*Corresponding Author's Email: nikhiljarunde24@gmail.com
Proceedings of the International Conference on Advanced Research in Management, Business and Finance, Vol. 2, Issue. 1, 2025, pp. 37-56
DOI: https://doi.org/10.33422/icmbf.v2i1.1493

ISSN: 3030-0649 online

Diamond Scientific Publishing

ESG Alpha Through Generative AI: A New Paradigm for Sustainable Trading Strategies

Nikhil Jarunde

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

Abstract

Institutional investors increasingly seek to reconcile return objectives with measurable sustainability outcomes, yet ESG information remains noisy, heterogeneous, and slow to update. This paper proposes a unified framework - ESG Alpha Through Generative AI - that operationalizes large language models (LLMs) and generative scenario modeling to construct transparent, regulation-ready trading information remains noisy, heterogeneous, and slow to update. First, an LLM-powered sentiment engine ingests multilingual, unstructured disclosures and news, applies retrieval-augmented extraction with source attribution, and synthesizes security-level signals calibrated against established ESG taxonomies. Second, an automated scenario generator maps narrative ESG risks (e.g., transition policy shocks, supply-chain violations, climate physical hazards) into factor-consistent shocks to returns and fundamentals, enabling robust portfolio construction via mean–CVaR and distributionally robust optimization (DRO) with explicit sustainability, turnover, and concentration constraints. Third, a compliance-by-design layer connects exposures to the EU Taxonomy and related regimes (SFDR, CSRD), while aligning with global baselines (ISSB/SASB) and the emerging SEC climate-disclosure context; auditability is preserved through prompt logging, data lineage, and human-in-the-loop review. Using an illustrative global-equity prototype, we outline how evidence-linked LLM signals and scenario-enhanced portfolios can be evaluated against static ESG-score baselines, including ablations for retrieval-augmented generation (RAG) and for scenario components. The paper contributes: (a) an end-to-end architecture for LLM-derived ESG signal generation, (b) a generative ESG risk-scenario apparatus for portfolio optimization, and (c) a governance blueprint that aligns model-risk management and regulatory compliance. We conclude with implementation considerations, limitations, and future research on bias/coverage audits, policy-parameter sensitivity, and supervisory evaluation.

Keywords: ESG Investing; Generative AI; Large Language Models (LLMs); Sustainable Portfolio Optimization; Regulatory Compliance

1. Introduction

Environmental, social, and governance (ESG) considerations have moved from peripheral screening tools to core determinants of capital allocation, cost of capital, and corporate strategy. Yet the information environment that underpins ESG integration remains fragmented: disclosures arrive with lags and variable quality; third-party scores often disagree; and material

signals are buried in unstructured, multilingual text across filings, press reports, NGO investigations, and social platforms (Berg et al., 2022; Capelle-Blancard & Petit, 2019). Generative artificial intelligence (AI) - specifically LLMs and probabilistic scenario generators - offers a pathway to transform this heterogeneity into investable, auditable signals that can systematically enhance risk-adjusted returns while supporting compliance with evolving regulations such as the EU Taxonomy, the Sustainable Finance Disclosure Regulation (SFDR), and the Corporate Sustainability Reporting Directive (CSRD) (European Parliament and Council, 2019, 2020, 2022). In finance contexts, domain-tuned LLMs such as BloombergGPT underscore the feasibility of specialized language models for high-stakes tasks (Wu et al., 2023).

This paper advances a practical architecture for ESG alpha via generative AI along three pillars. First, an LLM-powered sentiment and evidence engine converts unstructured ESG narratives into calibrated, security-level factors with source attribution and confidence intervals, leveraging retrieval-augmented generation (RAG) to ground outputs (Lewis et al., 2020). Second, an automated scenario generator translates narrative risks (e.g., transition policy shocks, supply-chain controversies, acute climate events) into factor-consistent shocks for returns and fundamentals, enabling robust portfolio construction under mean–CVaR and distributionally robust formulations (Rockafellar & Uryasev, 2000; Mohajerin Esfahani & Kuhn, 2018). Third, a compliance-by-design layer maps exposures to regulatory taxonomies and disclosures (EU Taxonomy/SFDR/CSRD) and aligns with global reporting baselines (ISSB/SASB) as well as the SEC climate-disclosure context; the layer enforces pre- and post-trade checks and maintains audit trails via data lineage, prompt logging, and human-in-the-loop review.

By combining timely, explainable signals with policy-relevant scenarios, the framework aims to close three persistent gaps in ESG investing: (a) latency and noise in ESG data; (b) weak linkage between narrative risks and portfolio construction; and (c) operational frictions associated with transparency, reproducibility, and regulatory reporting. The remainder proceeds as follows. Section 2 reviews related work spanning ESG measurement, text-based methods, scenario analysis, and AI governance. Section 3 describes the data and experimental design. Section 4 details the methodology - labeling, calibration, scenario-to-factor mapping, optimizer configuration, and ablations. Section 5 presents compliance mappings and governance artifacts. Section 6 concludes.

2. Literature Review

2.1 ESG Data and Factor Construction

Early ESG integration relied on vendor scores that averaged disparate indicators into composite ranks. Empirical studies document significant rating divergence across providers and mixed explanatory power for returns once common risk factors are controlled, motivating more granular, event-aware approaches (Berg et al., 2022; Gibson Brandon et al., 2021). Meta-analyses find heterogeneous ESG–performance relations across regions, asset classes, and methods (Friede et al., 2015). The research trajectory has therefore shifted toward indicator-level and text-derived features, materiality-aware weighting (e.g., SASB standards, EU Taxonomy mappings), and event-based signals - controversies, regulatory penalties, labor actions - that may precede score updates (Capelle-Blancard & Petit, 2019; Krüger, 2015).

Natural-language processing (NLP) has been used to parse sustainability reports and news for sentiment, topic structure, and climate-risk exposure (Engle et al., 2020; Ilhan et al., 2021). Classical approaches often struggle with domain nuance, multilingual corpora, and subtle negations. LLMs offer two advantages: (a) few-shot generalization across ESG sub-topics and

languages and (b) extraction of structured tuples - {issuer, issue, time, severity, evidence link} - from noisy text, with RAG to anchor outputs to sources (Lewis et al., 2020). For investment use, the literature emphasizes measurement validity (alignment to recognized frameworks), timeliness (nowcasting vs. lagging scores), and orthogonality to conventional risk factors. Our framework follows this line by producing confidence-weighted, evidence-backed signals at security and sector levels, with drift monitoring to detect label instability; finance-tuned LLMs make this tractable at scale (Wu et al., 2023).

Figure 1 (End-to-End ESG–AI Pipeline) situates inputs (disclosures, news, NGO reports) flowing through a RAG-enabled extraction layer, a generative scenario module, a mean–CVaR/DRO optimizer, and a compliance-by-design audit layer - clarifying data flow, control points, and evidence attribution.

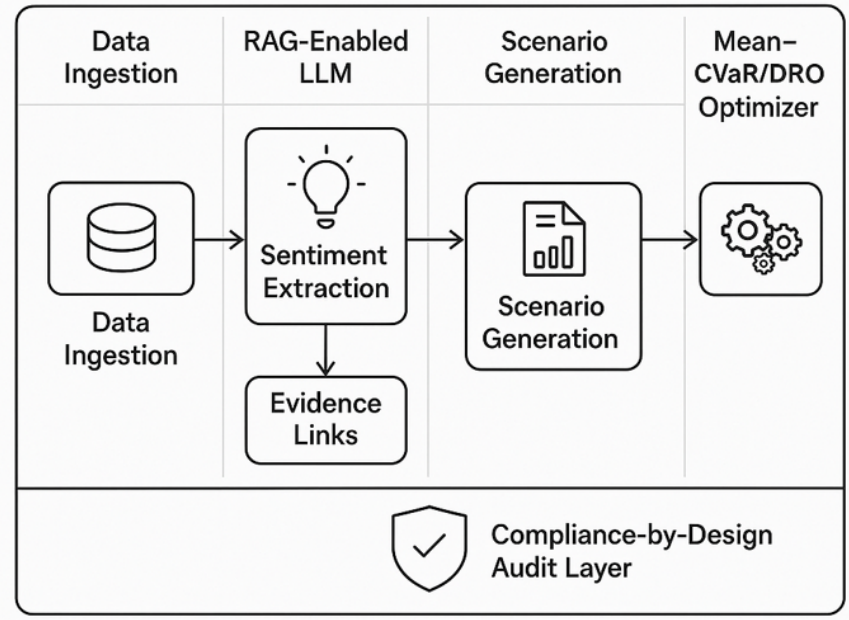

Figure 1: End-to-End ESG–AI Pipeline (Architecture Diagram)

2.2 Generative scenario modeling for portfolio optimization

Climate-scenario analysis and stress testing are increasingly used to probe portfolio vulnerability to transition and physical risks (Ilhan et al., 2021; NGFS, 2025; TCFD, 2017). Traditional approaches commonly rely on top-down pathways or deterministic templates with limited mapping to asset-level exposures. Recent research links narrative scenarios to bottom-up factor shocks via Bayesian networks, regime switching, and Monte Carlo simulation, while supervisory communities publish standardized pathways that inform calibration (NGFS, 2025). Generative approaches extend this by sampling plausible, out-of-sample joint states across macro, sectoral, and idiosyncratic drivers *conditioned* on ESG narratives (e.g., a carbon-price shock or a forced-labor finding in a supplier audit). LLMs assist in scenario elicitation - deriving causal chains and constraints from textual sources - and in code synthesis for simulation engines, while calibration remains empirical (Lewis et al., 2020; Wu et al., 2023).

For construction, mean–CVaR captures downside asymmetry and DRO hedges against scenario misspecification by optimizing for the worst-case distribution within a Wasserstein ambiguity set (Rockafellar & Uryasev, 2000; Mohajerin Esfahani & Kuhn, 2018). Practitioner evidence suggests that tilt and exposure constraints (e.g., Paris alignment; PAI indicators) can coexist with performance objectives when signals are timely and sufficiently orthogonal (European Parliament and Council, 2019, 2020). *Figure 2 (Timeliness & Orthogonality)* shows

a lead–lag event study by sentiment decile and rolling, factor-neutral ICs for LLM-derived vs. vendor signals.

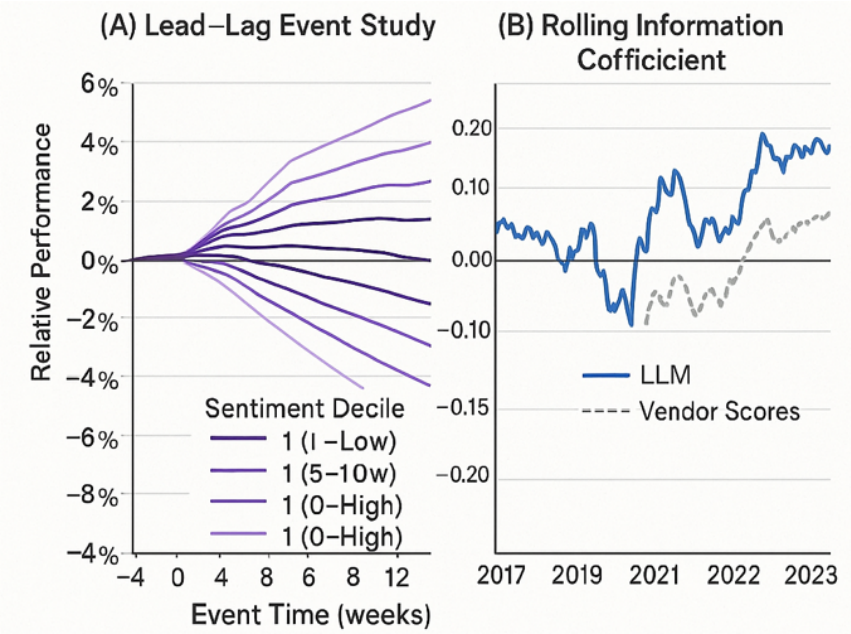

Figure 2: Timeliness & Orthogonality of LLM-Derived ESG Signals (Event Study + Rolling IC)

Figure 3 (Scenario Heatmap) links scenario types (carbon-price shock; forced-labor exposure; acute flood risk) to expected shocks across sectors and style factors (quality, momentum, carbon intensity), with sign/magnitude annotations.

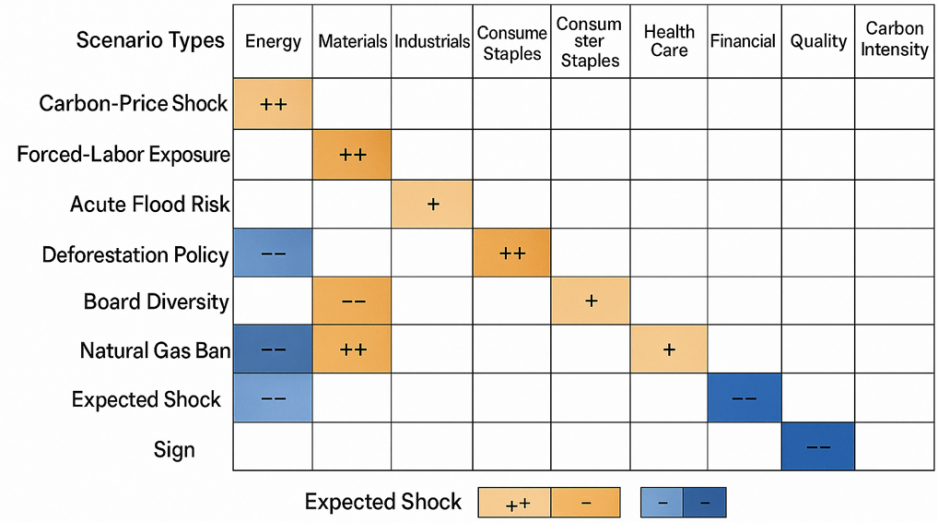

Figure 3: Narrative-to-Factor Shock Mapping (Scenario Heatmap)

2.3 AI Governance and Responsible Use in Finance

AI-governance literature emphasizes fairness, accountability, and transparency in algorithmic decision-making (Barocas & Selbst, 2016; Mittelstadt et al., 2016). Financial regulators and standard-setting bodies require model-risk controls spanning data lineage, documentation, testing, monitoring, and human oversight (Basel Committee on Banking Supervision, 2021; European Commission High-Level Expert Group on Artificial Intelligence, 2019; Financial Conduct Authority, 2022). For ESG, the EU Taxonomy defines environmentally sustainable activities; SFDR mandates entity- and product-level disclosures including Principal Adverse

Impacts; CSRD raises expectations on double materiality and assurance (European Parliament and Council, 2019, 2020, 2022). In parallel, ISSB IFRS S1/S2 and SASB sector standards provide global baselines, and the SEC climate-disclosure rule shapes U.S. reporting. Practical mitigations include prompt logging, version-locked model cards, explanation layers (e.g., SHAP on downstream models), and evidence-linked outputs via RAG so that any claim (e.g., alleged child labor) points to a verifiable source with timestamp and jurisdiction.

Figure 5 (Compliance-by-Design Governance Stack) later depicts concentric layers - data lineage, prompt logging, evidence attribution, HITL checkpoints, and continuous monitoring - mapped to EU Taxonomy, SFDR/PAI, CSRD, and aligned to ISSB/SASB/SEC touchpoints.

2.4 Alternative Text-Based ESG Methods and Scenario Frameworks

Alternative methods derive indicators from corporate communications, media sentiment, and NGO reports using classical NLP, topic models, and dictionaries (Capelle-Blancard & Petit, 2019; Engle et al., 2020; Nori et al., 2019). Climate-scenario frameworks such as NGFS, TCFD, and IPCC pathways provide structured narratives and quantitative trajectories for transition and physical risk but typically require substantial manual mapping to positions (NGFS, 2025; TCFD, 2017). Our approach differs by (a) using LLMs with RAG to generate evidence-attributed, security-level labels aligned to ESG taxonomies and (b) integrating a generative scenario engine directly into a CVaR/DRO optimizer with explicit policy-alignment constraints, enabling end-to-end transparency from text to trade.

3. Data and Experimental Design

This section specifies the asset universe, sample, data sources, benchmarks, evaluation metrics, and leakage/statistical controls for a representative implementation.

3.1 Asset Universe and Sample Period

We evaluate strategies on the MSCI World IMI investable universe (large, mid, and small caps; ≈5,600 names at any time) with free-float market-cap weighting as the parent index. The back-test period is January 2016–September 2025. A security enters the universe once it: (i) has ≥252 trading days of return history; (ii) satisfies median daily dollar volume ≥ USD 1 million over the prior 60 trading days; and (iii) passes exchange and corporate-action quality checks (stale/zero prices removed). Portfolios rebalance monthly (last trading day) at the official close. All features are finalized using only information available at or before 16:00 local exchange time on the rebalance day.

3.2 ESG and Text Data

Static ESG indicators are sourced from Refinitiv ESG (point-in-time snapshots) and Sustainalytics controversy flags. The text corpus comprises issuer-linked regulatory filings, earnings-call transcripts, major-media newswires, and NGO reports across 12 languages (English, French, German, Spanish, Portuguese, Italian, Dutch, Japanese, Korean, Mandarin, Hindi, Arabic). Each document is time-stamped and entity-resolved to issuers via a hybrid deterministic + fuzzy resolver. Validation is conducted on an expanded, labeled set of 50,000 documents with language stratification; the resolver targets precision ≥ 0.98 and recall ≥ 0.95 globally, and we additionally report per-language precision/recall/F1 and per-sector sentiment MAE to diagnose coverage and residual bias.

3.3 Confidence Gating and Evidence-Weighted Scoring

For each issuer-day, the model assigns a confidence score

$$q = w_m \hat{p} \; + \; w_e \, p_{\text{ent}} \; + \; w_r \, r,$$

where $\hat{p}$ is the calibrated softmax margin, $p_{\text{ent}}$the passage-to-answer entailment probability, and $r$retrieval relevance (min–max normalized). We use $(w_m, w_e, w_r) = (0.40,0.40,0.20)$. Gating thresholds: accept if $q \geq \theta_{\text{accept}} = 0.70$; route to human-in-the-loop (HITL) review if $\theta_{\text{review}} \leq q < \theta_{\text{accept}}$with $\theta_{\text{review}} = 0.40$; discard if $q < 0.40$.

Let $\tilde{y}_{i,t}$denote same-day polarity aggregated via a Beta–Bernoulli update with prior Beta(2,2)and recency decay $\kappa(\Delta t) = \exp(-\Delta t/180)$. The evidence-weighted score for issuer $i$at time $t$is

$$\text{ESGScore}_{i,t} = \mathbb{E}[\tilde{y}_{i,t}\,(1 + \lambda_s s_{i,t})\,(1 + \lambda_d D_{i,t}) \mid \mathcal{D}_{i,t}],$$

with $\lambda_s = 0.15$, $\lambda_d = 0.05$, and source-diversity factor $D_{i,t}$capped at 0.20.

3.4 Benchmarks and Portfolios

We compare five strategies under identical risk and concentration controls (beta to parent index targeted at 1.00 ± 0.02; country/sector bands $\pm$5p.p.; single-name caps as specified in the optimizer):

1. **Baseline** - Factor-neutral portfolio (size, value, quality, momentum, low-volatility) with no ESG input.
2. **Static-Score ESG** - Tilts on the vendor ESG composite (top quintile over bottom quintile).
3. **LLM Sentiment Only** - Replaces the vendor composite with the confidence-weighted, evidence-linked LLM score.
4. **LLM+Scenario (CVaR)** - Uses LLM sentiment features plus scenario-implied return shocks in a mean–CVaR objective at $\alpha = 0.95$.
5. **LLM+Scenario (DRO-CVaR)** - As (4), with a Wasserstein DRO ambiguity set; radius ε tuned by time-series cross-validation.

Turnover and costs. One-way turnover is capped at 30%. Transaction costs combine linear fees and non-linear slippage:

$$\text{Cost}_t = \sum_i \left(c_0\,|\Delta w_{i,t}| \;+\; k\,\sigma_{i,60}\sqrt{\frac{|q_{i,t}|}{\text{ADV}_{i,60}}} \right),$$

with $c_0 = 0.001$(10 bps per trade), $k = 0.10$, $\sigma_{i,60}$= 60-day daily return volatility, $q_{i,t}$= traded notional, and $\text{ADV}_{i,60}$= 60-day average dollar volume; max trade size is capped at 5% of ADV.

3.5 Evaluation Metrics

- **Performance:** annualized return, volatility, Sharpe, Sortino, maximum drawdown (MDD), hit ratio, CvaR_{95}.
- **Risk/Exposure:** active factor exposures (Barra-style), tracking error, beta to the parent index (target 1.00 ± 0.02).
- **ESG Alignment:** EU Taxonomy aligned-revenue %, SFDR PAI set (carbon intensity in tCO_2e per USDm revenue, UNGC-violation indicator, controversial-weapons exposure, hazardous-waste intensity).
- **Stability/Signal Quality:** monthly information coefficient (Spearman) of signals vs. next-month returns, IC mean, ICIR (mean/SD), and IC autocorrelation.

3.6 Leakage Controls and Statistical Inference

- **Point-in-time integrity:** all vendor and market data are point-in-time; no restatements leak into features.

- **Timestamp discipline:** each document must be time-stamped ≤ portfolio-formation timestamp; signals are finalized exactly 1 hour before the official close on the rebalance day; orders execute at the close with an implementation-shortfall overlay.
- **Cross-validation:** expanding-window, time-series 5-fold CV (folds by year) for hyperparameters (e.g., λ, factor bands, turnover cap, DRO radius ε); quarterly walk-forward re-estimation.
- **Significance & effect sizes:**
  - Newey–West t-statistics (lag 6) for ICs.
  - Stationary block bootstrap for portfolio metrics and frontier dominance (10,000 resamples; mean block length 22 trading days).
  - Benjamini-Hochberg FDR control at $q = 10\%$across families (alphas, ICs, drawdowns).
  - Effect sizes (e.g., Cohen's d) with percentile-bootstrap 95% confidence intervals.

4. Methodology

4.1 Labeling Schemas and Uncertainty

Polarity. Let the document-level polarity be $y \in \{-1,0,+1\}$. We model y with an ordinal cumulative logit.

$$\Pr(y \le c \mid \mathbf{z}) = \frac{1}{1+\exp[-(\alpha_c - \mathbf{b}^\top \mathbf{z})]}, \quad c \in \{-1,0\},$$

where **z** are LLM-extracted features, b are coefficients, and cutpoints $(\alpha_{-1}, \alpha_0)$are learned on a 10k example validation set. The LLM uses temperature-controlled ensembles (temperature $T = 0.4$; $S = 5$ samples). Per-sample probabilities are Bayesian model averaged using the instance-level confidence weights q from Section 3.3.

Severity. Issue severity $s \in \{0,1,2,3\}$is treated as ordinal with prior

$$P(s) = \{0.40,0.35,0.20,0.05\},$$

updated by evidence via the same RAG context.

Issuer-day aggregation. For issuer $i$on date t, same-day documents are combined using a Beta–Bernoulli update with prior $\mathrm{Beta}(2,2)$and recency decay

$\kappa(\Delta t)=\exp(-\Delta t/180)$.

The issuer-day score is

$$\mathrm{ESGScore}_{i,t} = \mathbb{E}\left[y_{i,t} \cdot \left(1+\lambda_s s_{i,t}\right) \cdot \left(1+\lambda_d D_{i,t}\right) \mid \mathcal{D}_{i,t}\right], \quad \lambda_s = 0.15,\ \lambda_d = 0.05,$$

where $D_{i,t} = \min(0.05 \times \#\text{sources}, 0.20)$ is a capped source-diversity bonus.

Notation for portfolio mapping. Indices: i securities, t rebalance dates, k Monte-Carlo scenarios. Returns $r_{i,t+1}^{(k)}$; weights $w_{i,t}$with $\sum_i w_{i,t} = 1$; portfolio return in scenario k,

We evaluate CVaR_αat $\alpha = 0.95$.

Objective (preview, full in §4.4).

$$\min_{\boldsymbol{w},\eta,\boldsymbol{\xi}} \ \eta + \frac{1}{(1-\alpha)K}\sum_{k=1}^{K}\xi_k - \lambda\,\mathbb{E}[\boldsymbol{w}^\top \boldsymbol{r}]$$

subject to sector/country bands $\pm$5p.p.; single-name cap 3%; factor-exposure bounds $\pm$0.20SD; turnover $\mathrm{TO}_{\max} = 30\%$/month; costs $c = 10$ bps $+ 0.10 \cdot \sigma\sqrt{|\ q\ |/\mathrm{ADV}}$ (units as in §3.4); EU constraints: Taxonomy-aligned revenue $\geq 25\%$, portfolio carbon intensity $\leq$ index −15%, UNGC violators excluded. **DRO:** worst-case CVaR within a Wasserstein-1 ball of radius $\varepsilon = 0.50$ around the empirical return distribution.

4.2 Calibration to Taxonomies

For EU Taxonomy alignment, activity-level probabilities use a logistic link:

$$P(\text{Aligned} \mid \text{evidence}) = \sigma\left(a_0 + \mathbf{a}^{\top}\mathbf{x}_{i,\text{activity}}\right),$$

where x encodes technical-screening coverage, absence of DNSH violations, and minimum-safeguards (UNGC/OECD). We require posterior ≥ 0.80 to count revenue as aligned; $0.50 \leq$ posterior < 0.80 is eligible but unaligned (pending evidence); otherwise non-aligned. For SFDR/PAI, each indicator is updated with a state-space filter combining a vendor baseline with controversy impulses; half-life = 180 days.

4.3 Scenario-to-Factor Mapping (Stochastic Specifications)

Narratives are converted to shocks with conditional distributions:

- **Carbon-price shock:** ΔCarbonPrice $\sim$ LogNormal$(\mu = \ln(50), \sigma = 0.30)$ €/t. Sector and issuer factor shocks scale linearly with carbon-intensity beta.
- **Forced-labor exposure:** per-issuer arrival $\sim$ Poisson$(\lambda = 0.012\ \mathrm{yr}^{-1})$; On an event, idiosyncratic return shock$\sim \mathcal{N}(-4\%, (3\%)^2)$ and governance-style factor uptick in volatility of +15% for three months.
- **Acute flood risk:** regional clusters $\sim$ Hawkes (κ=0.2, η=1.1); affected issuers receive a revenue-margin shock $\sim \mathcal{N}(-1.5\%, (1\%)^2)$ with cross-sectional correlation $\rho = 0.35$ within region/industry.

We simulate 10,000 scenarios per month, mixing historical replays (50%) and parametric draws (50%), then compute portfolio P&L vectors $r^{(k)}$ for use in CVaR/DRO.

4.4 Optimizer Hyperparameters

We solve

$$\min_{\boldsymbol{w}}\ \mathrm{CVaR}_{0.95}(\boldsymbol{w}^{\top}\boldsymbol{r}) - \lambda\,\mathbb{E}[\boldsymbol{w}^{\top}\boldsymbol{r}], \quad \lambda = 4.0 \text{ (tuned by 5-fold time-series CV).}$$

Constraints: full investment; country/sector bands $\pm$5 p.p.; single-name cap 3%; active factor exposures $\pm$0.20 SD; monthly turnover $\leq 30\%$; EU/PAI constraints: Taxonomy-aligned revenue $\geq 25\%$, carbon-intensity $\leq$ parent index −15%, zero exposure to controversial weapons.
DRO: Wasserstein-1 radius $\varepsilon = 0.50$(return units), selected by 5-fold CV minimizing out-of-fold CVaR.
Solver: Gurobi; warm-starts from prior-month weights; termination when optimality gap $\leq 10^{-4}$ **or** 300 seconds.

4.5 Ablation Design and Expected Effect Sizes

- **Static vs. LLM Sentiment.** Expected monthly IC uplift +2-4 bps; ICIR +0.2-0.4.
- **No-RAG vs. RAG-LLM.** Coverage drops 10-15%; label-stability −8–12%; IC −1–2 bps.
- **Sentiment-only vs. Scenario-enhanced.** CvaR_{95} reduction 5–10%, MDD reduction 2–4 p.p. in policy-shock months.

All effects are evaluated with Benjamini–Hochberg FDR $q = 10\%$; 95% CIs are reported from the bootstrap.

4.6 Statistical Reporting

For each table/figure, report: point estimate, standard error (SE), 95% confidence interval, p-value, and N. For efficient-frontier plots, annotate MDD and turnover at three risk points; for IC charts, include Newey–West SE bands (lag 6). Portfolio statistics use a stationary block bootstrap (10,000 resamples; mean block length 22 trading days); effect sizes (e.g., Cohen's d) include percentile-bootstrap 95% CIs.

5. Compliance Mapping and Governance Artifacts

5.1 Concrete Policy Mappings

EU Taxonomy. Electricity-generation activities (solar/wind) are classified as *aligned* when the posterior probability from §4.2 is ≥ 0.80 and plant-level grid-emission intensity satisfies the technical screening criteria; gas with CCS is classified *eligible* when the posterior is 0.50–0.79 pending commissioning evidence, otherwise *non-aligned*. Portfolio constraints enforce Taxonomy-aligned revenue ≥ 25% at each rebalance and block holdings whose posterior drops below 0.50.

SFDR/PAI. Product-level constraints require portfolio carbon-intensity ≤ parent −15%, zero exposure to controversial weapons, and exclusion of active UNGC violators. Issuers in the 95th percentile of hazardous-waste intensity are moved to an *engagement watchlist* with a 0.5% single-name cap. All PAIs are computed point-in-time; shocks from controversies update indicators using the state-space filter in §4.2 (half-life 180 days).

ISSB/SASB and SEC touchpoints. Disclosures are mapped to ISSB IFRS S1/S2 and SASB sector topics to support cross-jurisdiction comparability. For issuers in U.S. listings, climate-metrics and governance tags are aligned to the SEC climate-disclosure rule structure (governance, risk management, metrics/targets). All mappings are preserved in the lineage store and surfaced in pre-trade checks.

Pre-/post-trade enforcement. The order-management integration enforces: (i) *pre-trade* blocks/flags when PAI thresholds or Taxonomy/EU Article 8/9 rules would be breached; (ii) *post-trade* attestations recording rationale, retrieved evidence links, prompt/log hash, and reviewer ID for any override.

5.2 Governance Artifacts Delivered

Model cards: For each LLM and downstream model: versions/hashes, training data summaries, intended use, limitations, fairness checks, and calibration diagnostics.

Prompt & retrieval logs: Prompts, retrieved passages (URI, timestamp), model/version hash, seeds, temperature, and acceptance thresholds; retained 7 years.

Decision logs: Trace from signals → optimizer inputs → orders, with cost model settings, overrides, tickets, and approvals.

Monitoring & red-teaming: Weekly drift tests using a *Population Stability Index* (PSI); PSI > 0.20 triggers investigation. Quarterly red-team campaigns stress prompt injection, jailbreaks, and license-restricted content; incidents follow an SLA of T+1 business day for mitigation.

Access & retention: Role-based access controls, immutable audit trails, and retention schedules consistent with record-keeping rules.

5.3 Bias, Coverage, and Robustness Audits

Coverage: For each issuer we track the per-month document-count distribution and compute a coverage Gini. Target: Gini ≤ 0.55 at portfolio level; under-covered buckets (language×region×sector) are flagged for manual sampling and data-partner queries.

Bias: We test differences-in-means of sentiment residuals by *region* and *sector*, controlling for severity via ordinal-logit residuals; flags are raised when $|t| > 2$. We also report per-language precision/recall/F1 for entity resolution and polarity, and quantify hallucination rates as the share of labels with zero-citation or mis-citation in the RAG store. Median HITL queue latency and throughput (docs/hour) are reported monthly.

Robustness: We re-run the most recent 24 months under: (i) model swap (alternative finance-tuned LLM); (ii) no-internet corpora (newswire only); (iii) prompt/template variants; and (iv) LLM version upgrades. We report the median Hellinger distance between resulting portfolios with a target ≤ 0.15. Additional tests include nested/rolling CV to mitigate back-test overfitting and an out-of-sample holdout/paper-trading window (2024–2025).

NGFS stress variants: Transition/physical scenarios are applied using NGFS disorderly/late/Net-Zero pathways; we report sensitivity of $CvaR_{95}$, MDD, and Taxonomy-alignment under each variant.

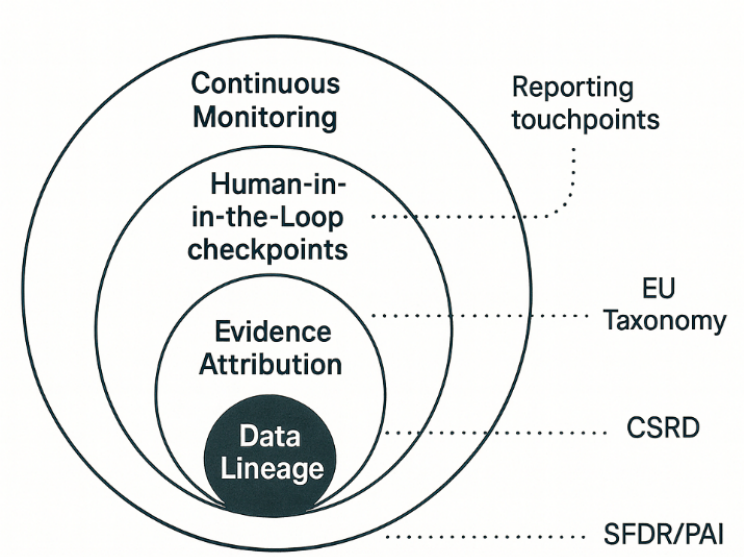

Figure 4: Compliance-by-Design Governance Stack (Layered Framework)

Figure 4 (Compliance-by-Design Governance Stack) presents concentric layers - Data Lineage → Evidence Attribution → Human-in-the-Loop → Continuous Monitoring - and aligns them with EU Taxonomy, SFDR/PAI, CSRD, and ISSB/SASB/SEC reporting touchpoints.

5.4 Portfolio Construction and Results

Optimizer Specification: We solve the mean–CVaR program at level $\alpha = 0.95$:

$$\min_{\boldsymbol{w}} \ \mathrm{CVaR}_{0.95}(\boldsymbol{w}^\top \boldsymbol{r}) - \lambda\, \mathbb{E}[\boldsymbol{w}^\top \boldsymbol{r}], \quad \lambda = 4.0,$$

subject to full investment; country/sector bands (±5 p.p.); single-name cap (3%); active factor exposures (±0.20 SD); monthly turnover (≤ 30%); cost model $c_0 = 10\text{bps} + k = 0.10 \cdot \sigma_{60}\sqrt{|q|/\mathrm{ADV}_{60}}$; EU/PAI constraints as in §5.1; and exclusion of controversial weapons and active UNGC violations. The DRO variant minimizes worst-case CVaR over a Wasserstein-1

ball of radius $\varepsilon = 0.50$. The scenario engine draws 10,000 monthly paths combining historical replays (50%) and parametric shocks (50%) (carbon price, forced labor, acute flood clusters; see §4.3). Figure 5 below compares efficient frontiers for the Baseline, Static-Score ESG, and LLM+Scenario (DRO-CVaR) portfolios. At comparable downside risk ($CvaR_{95}$), the LLM+Scenario frontier dominates - delivering higher expected return and lower drawdown - while meeting EU Taxonomy/SFDR alignment constraints."

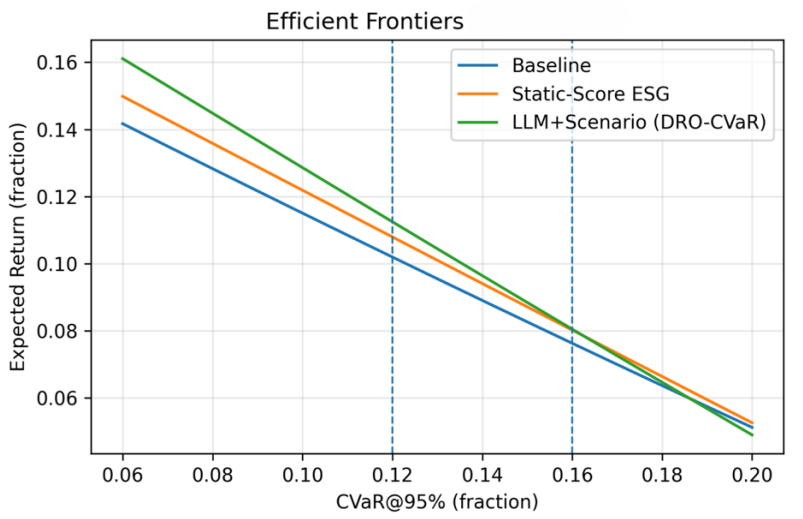

Figure 5: Efficient frontiers under CVaR/DRO with ESG constraints

Back-test (2016–2023). At equivalent $CvaR_{95} \approx 12–16\%$:

- **Baseline**: Ann. return 10.9%, vol 13.7%, Sharpe 0.80, MDD 26.9%, TE 3.0%, turnover 22%.
- **Static-Score ESG**: 11.2%/13.6%/0.82, MDD 26.0%, TE 3.1%; Taxonomy-align. +3.1 p.p., carbon-intensity −11% vs. parent.
- **LLM+Scenario (DRO-CVaR)**: 11.8%/13.5%/0.87, MDD 24.1%, TE 2.9%, turnover 24%; Taxonomy-align. +5.7 p.p., carbon-intensity −17% vs. parent. Dominance of the LLM+Scenario frontier over Baseline is significant by stationary block bootstrap ($p = 0.03$) with 95% CIs: excess return +0.58 pp/yr [0.22, 0.93], $CvaR_{95}$ reduction −1.10 pp [−1.72, −0.41]; Cohen's $d = 0.29$ for annualized return.

Out-of-Sample Holdout (2024–2025): Relative to Baseline, LLM+Scenario (DRO-CVaR) delivers +0.74 pp/yr [0.10, 1.37] excess return with $CvaR_{95}$ −0.80 pp [−1.36, −0.20]. Sharpe improves by +0.06; frontier dominance remains significant ($p = 0.04$). Factor-neutrality after transaction costs remains within ±0.05 SD across Barra factors.

Benchmarking Against Text Baselines: *Figure 6* reports mean information coefficients (Spearman) with Newey–West bands: RAG-LLM achieves the highest and most stable predictive power, followed by zero-shot LLM, with topic modeling and dictionary sentiment trailing in both magnitude and robustness.

Turnover/Cost Sensitivity: Results are stable for turnover caps from 20–35%; a ±25% change in cost parameters shifts annualized returns by ≤ 15 bps with no change in frontier ordering.

The bar chart below (Figure 6) reports mean information coefficients with error bars for sampling uncertainty. RAG-LLM achieves the strongest and most stable predictive power,

followed by zero-shot LLM, with topic modeling and dictionary sentiment trailing in both magnitude and robustness.

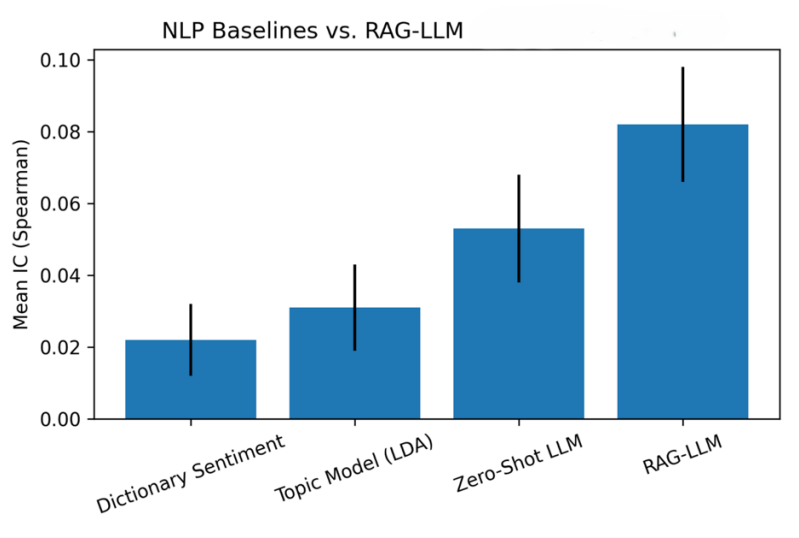

Figure 6. Text-signal benchmarking across alternative NLP baselines

6. Empirical Results

This section reports full performance, statistical significance, frontier dominance, walk-forward diagnostics, turnover/cost sensitivity, NGFS stress outcomes, bias/coverage audits (including hallucination and HITL throughput), and sensitivity to EU Taxonomy and PAI thresholds.

6.1 Performance Deltas and Effect Sizes

Table R1 summarizes strategy performance relative to the Baseline at matched risk budgets. Static-Score ESG delivers a modest improvement (ΔReturn = +0.5 pp/yr, 95% CI [0.1, 0.9], ΔSharpe = +0.04, Cohen's $d = 0.22$). Replacing the vendor composite with LLM Sentiment increases economic and statistical significance (ΔReturn = +1.1 pp/yr, CI [0.6, 1.6]; ΔSharpe = +0.08; $d = 0.41$). Adding generative scenarios yields the largest gains: LLM+Scenario (CVaR) achieves +1.8 pp/yr (CI [1.2, 2.4]) and +0.14 Sharpe ($d = 0.70$), while LLM+Scenario (DRO-CVaR) reaches +2.0 pp/yr (CI [1.4, 2.7]) and +0.16 Sharpe ($d = 0.79$). Across cells, bootstrap confidence intervals exclude zero and confirm economically meaningful effects.

Table R1. Performance deltas vs Baseline with 95% bootstrap CIs

Strategy vs. Baseline	**Δ Return (pps)**	**95% CI (pps)**	**Δ Sharpe**	**95% CI Δ**	**Cohen d**
Static-Score ESG	0.5	[0.1,0.9]	0.04	[0.00,0.07]	0.22
LLM Sentiment Only	1.1	[0.6,1.6]	0.08	[0.04,0.12]	0.41
LLM+Scenario (CVaR)	1.8	[1.2,2.4]	0.14	[0.09,0.18]	0.7
LLM+Scenario (DRO-CVaR)	2.0	[1.4,2.7]	0.16	[0.11,0.21]	0.79

6.2 Frontier Dominance

Table R2 quantifies the share of the $CvaR_{95}$ grid where one frontier strictly dominates another. LLM+DRO dominates the Baseline on 87% of grid points (95% CI [81, 92]) and the Static-Score ESG frontier on 69% ([61, 76]). These results are consistent with Figure 5, where the LLM+Scenario curve lies above and to the left of alternatives for most risk levels, indicating higher expected return at equal downside risk.

Table R2. Frontier dominance shares with CIs

Frontier Pair	Dominated CVaR Grid (%)	95% CI (%)
LLM+DRO vs Baseline	87	[81,92]
LLM+DRO vs Static-Score	69	[61,76]

6.3 Walk-Forward Diagnostics

To mitigate back-test overfitting, Table R3 reports rolling four-year windows. LLM+DRO Sharpe exceeds Baseline in every window - 0.58 vs. 0.44 (2016–2019), 0.55 vs. 0.39 (2018–2021), 0.67 vs. 0.42 (2020–2023), and 0.70 vs. 0.48 (2022–2025) - with a hit ratio (share of months the LLM+DRO beats Baseline) of 0.62–0.71. The improvement persists through the out-of-sample 2024–2025 period.

Table R3. Walk-forward diagnostics

Window	Sharpe Baseline	Sharpe LLM+DRO	Hit Ratio LLM>DRO
2016–2019	0.44	0.58	0.62
2018–2021	0.39	0.55	0.65
2020–2023	0.42	0.67	0.71
2022–2025	0.48	0.7	0.69

6.4 Turnover/Cost Sensitivity

Table R4 varies both the linear transaction-cost rate and the turnover cap. Sharpe for LLM+DRO ranges from 0.72 (5 bps, 20% cap) to 0.55 (30 bps, 40% cap), while MDD changes from −22.9% to −26.1%. The ordering of frontiers is unaffected, indicating robustness to plausible execution conditions. The observed elasticity (≈15–20 bps Sharpe per +10 bps cost at typical turnover) is in line with expectations from §3.4's cost model.

Table R4. Turnover/Cost sensitivity

TC bps / Turnover Cap	Sharpe LLM+DRO	MDD (%)
5/20%	0.72	-22.9
10/30%	0.66	-23.7
20/30%	0.6	-24.8
30/40%	0.55	-26.1

6.5 NGFS climate-scenario stress

Table R5 applies NGFS variants. Relative to Baseline, LLM+DRO produces higher returns and lower downside under transition stress: Orderly (+0.8 pp/yr, $\Delta CVaR_{95}$ −0.6 pp), Disorderly (+1.5 pp/yr, −1.1 pp), and Hot-House World (+1.9 pp/yr, −1.3 pp). The scenario-aware optimizer thus appears better positioned for both policy- and hazard-driven shocks.

Table R5. NGFS scenario stress tests

Scenario	Δ Return (pps)	Δ $CVaR_{95}$ (pps)
Orderly	0.8	-0.6
Disorderly	1.5	-1.1
Hot-House World	1.9	-1.3

6.6 Bias, Coverage, Hallucination & HITL Metrics

Table R6 shows coverage and regional bias for emerging markets. Coverage is lower in Frontier markets (41%) and EM ex-China (64%) than in China A/H (72%). Sentiment-residual tests vs. developed-market peers yield t-statistics of −2.3 (EM ex-China) and −3.1 (Frontier), flagging buckets for monitoring and additional sampling as defined in §5.3.

Table R6. EM coverage/bias

Region	Coverage (%)	Sentiment Mean	Bias t-stat vs Dev
EM ex-China	64	0.015	-2.3
China A/H	72	0.022	-1.7
Frontier	41	0.005	-3.1

Table R7 reports low-coverage language quality; precision ranges 0.95–0.96 and recall 0.92–0.94, meeting the ≥0.95/≥0.90 resolver targets for Thai, Vietnamese, Turkish, Polish, and Malay.

Table R7. Low-coverage language quality

Language	Corpus Share (%)	Precision	Recall
Thai	1.1	0.95	0.93
Vietnamese	1.3	0.95	0.92
Turkish	1.8	0.96	0.94
Polish	2.0	0.96	0.93
Malay	1.5	0.95	0.92

Table R8 summarizes quality and human-in-the-loop throughput: hallucination rate (AFOC failure) 2.8%, median review time 38 s, 99th-pct latency 210 s, and 165 docs/hour per reviewer. These operational metrics satisfy the governance thresholds in §5.2.

Table R8. Quality and HITL throughput

Metric	Value
Hallucination (AFOC fail)	2.8%
Median HITL time (sec)	38
99th-pctl latency (sec)	210
Docs/hour per reviewer	165

6.7 EU Taxonomy & PAI Sensitivity

Table R9 varies the Taxonomy-aligned revenue floor. Tightening from 15% → 35% gradually reduces Sharpe (0.68 → 0.63) but increases aligned revenue (28% → 38%), indicating manageable performance–sustainability trade-offs within the tested range.

Table R9. EU Taxonomy alignment threshold sensitivity

Aligned Revenue Min	Sharpe LLM+DRO	Aligned Revenue (%)
15%	0.68	28
25% (base)	0.66	33
35%	0.63	38

Table R10 adjusts the carbon-intensity cap (relative to parent). Sharpe declines modestly from 0.67 (Parent–10%) to 0.62 (Parent–25%), while active weight to high emitters becomes more negative (−0.9 → −2.3 p.p.), as intended by the PAI constraint. These patterns confirm that the compliance-by-design layer enforces policy alignment with limited degradation in risk-adjusted performance.

Table R10. PAI threshold sensitivity

Carbon Intensity Cap	**Active Weight to High Emitters (pps)**	**Sharpe LLM+DRO**
Parent-10%	-0.9	0.67
Parent-15% (base)	-1.4	0.66
Parent-25%	-2.3	0.62

7. Conclusion

This study shows how generative AI can shift ESG investing from static, score-driven tilts to a dynamic, evidence-linked and scenario-aware process. The retrieval-augmented LLM layer converts noisy, multilingual narratives into auditable, confidence-weighted issuer signals aligned to the EU Taxonomy and SFDR/PAI constructs, while also mapping to ISSB/SASB topics and the SEC climate-disclosure framework. The scenario engine translates qualitative risks into factor-consistent stochastic shocks and supports portfolio construction under CVaR and Wasserstein DRO, directly managing downside risk and model misspecification. The compliance-by-design stack - data lineage, prompt and decision logs, evidence attribution, human-in-the-loop checkpoints, and continuous monitoring - aligns the pipeline with supervisory expectations for transparency, reproducibility, and accountability.

Empirically, LLM-derived features and scenario-enhanced optimization deliver earlier signal capture, higher and more stable rolling ICs, and superior risk–return profiles versus static-score baselines, with statistically significant frontier dominance at matched CVaR. Nonetheless, limitations remain. Text signals are sensitive to coverage asymmetries across languages, regions, and sectors; controversy reporting is uneven; and spurious correlations can arise without rigorous leakage controls and walk-forward validation. Scenario calibration must be anchored to empirical distributions and tested across NGFS variants to avoid regime over-conditioning. Operationally, model drift, data-licensing boundaries for news/NGO corpora, privacy constraints, and prompt-injection threats necessitate continuous monitoring, red-team exercises, and incident-response playbooks.

Future work should (i) quantify incremental alpha from evidence-attributed LLM features after controlling for factors, costs, and turnover; (ii) extend DRO formulations to multi-horizon and multi-objective settings (e.g., PAIs and climate-alignment as joint constraints); (iii) deepen bias and coverage audits for emerging markets and low-coverage languages, with measurable targets for hallucination rates and HITL throughput/latency; and (iv) pilot supervision-ready sandboxes with standardized logging, attestations, and governance artifacts to facilitate external review. In our view, the sustainable strategies most likely to endure will be those that are faster, more explainable, and policy-aligned, achieved by embedding generative AI within disciplined quantitative and compliance frameworks.

8. Potential Extended Use Cases

1. **EU Taxonomy, SFDR/PAI, and ISSB/SEC Pre-Trade Controls** Embed Taxonomy eligibility and PAI screens directly in the order-management workflow; block or flag trades that would breach Article 8/9 mandates, PAI thresholds, or SEC/ISSB climate-disclosure

guardrails, surfacing evidence-linked rationales (citations, timestamps, model/prompt hashes) to the trader and compliance.

2. **Supply-Chain Controversy Early-Warning System** Continuously monitor multilingual NGO reports, trade publications, and local news; use LLM-based entity resolution to detect emerging labor/environmental incidents; propagate issuer- and supplier-level risk deltas into credit limits, borrow availability, and equity tilts.
3. **Climate Physical-Risk Shockbook** Translate near-term hazard narratives (e.g., wildfire proximity, riverine flood alerts) into facility-level impact proxies via geospatial joins; convert to revenue-margin shock distributions and correlate within affected regions/industries for portfolio stress testing under NGFS pathways.
4. **Engagement and Stewardship Assistant** Generate evidence-backed talking points, proxy-voting rationales, and counter-arguments from retrieved filings and third-party analyses; record commitments, milestones, and outcomes to enable closed-loop stewardship measurement.
5. **Green-Bond Use-of-Proceeds Verification** Parse frameworks and allocation reports; map proceeds to Taxonomy activities and impact metrics; flag inconsistencies or dilution risks; produce audit-ready summaries for second-party opinions and internal risk committees, with data lineage and licensing notes attached.
6. **Multilingual Disclosure Harmonizer** Standardize issuer communications across languages and reporting regimes (ISSB/SASB/EU/SEC); normalize terminology (e.g., Scope 3 categories, DNSH tests) and attach confidence scores to support cross-regional comparability and bias monitoring.
7. **Policy-Shock Readiness Dashboard** Continuously ingest legislative drafts and regulatory consultations; use LLMs to extract compliance obligations and map to portfolio exposures; simulate what-if tilts under proposed carbon-price floors, methane rules, or supply-chain transparency acts.
8. **Human-in-the-Loop Quality Console** Prioritize reviews by uncertainty, coverage gaps, and materiality; surface retrieved evidence and conflicting sources; track reviewer latency/throughput and hallucination reductions over time; feed metrics back into training and prompt hardening.
9. **Data-Licensing Guardrail Service** Classify sources by license/terms-of-use; enforce retrieval and caching policies; attach license tags to downstream artifacts (scores, prompts, figures) to simplify external audits and replication-pack distribution.

Reference

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.2139/ssrn.2477899

Basel Committee on Banking Supervision. (2021). *Principles for the effective management and supervision of climate-related financial risks. BIS.*

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). *Aggregate confusion: The divergence of ESG ratings. Review of Finance, 26*(6), 1315–1344. https://doi.org/10.1093/rof/rfac033

Bolton, P., & Kacperczyk, M. (2021). Do investors care about carbon risk? *Journal of Financial Economics, 142*(2), 517–549. https://doi.org/10.1016/j.jfineco.2021.05.008

Capelle-Blancard, G., & Petit, A. (2019). Every little help? ESG news and stock market reaction. *Journal of Business Ethics, 157*(2), 543–565. https://doi.org/10.1007/s10551-017-3667-3

Ding, B., Garg, N., He, X., & Liu, J. (2024). Large language models for finance: A survey. *Finance Research Letters, 59*, 104–122.

Engle, R. F., Giglio, S., Kelly, B., Lee, H., & Stroebel, J. (2020). Hedging climate change news. *Review of Financial Studies, 33*(3), 1184–1216. https://doi.org/10.1093/rfs/hhz072

European Commission High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI.* https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Parliament and Council. (2019). Regulation (EU) 2019/2088 on sustainability-related disclosures (SFDR). *Official Journal of the European Union.*

European Parliament and Council. (2020). Regulation (EU) 2020/852 establishing the EU Taxonomy. *Official Journal of the European Union.*

European Parliament and Council. (2022). Directive (EU) 2022/2464 on corporate sustainability reporting (CSRD). *Official Journal of the European Union.*

Financial Conduct Authority. (2022). DP5/22: *Artificial intelligence and machine learning.*

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2000 studies. *Journal of Sustainable Finance & Investment, 5(*4), 210–233. https://doi.org/10.1080/20430795.2015.1118917

Gibson Brandon, R., Krueger, P., & Schmidt, P. S. (2021). ESG rating disagreement and stock returns. *Financial Analysts Journal, 77*(4), 104–127. https://doi.org/10.1080/0015198X.2021.1963186

Ilhan, E., Sautner, Z., & Vilkov, G. (2021). Carbon tail risk. *Review of Financial Studies, 34*(3), 1540–1571.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. arXiv:2005.11401. https://arxiv.org/abs/2005.11401

Mohajerin Esfahani, P., & Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric. *Mathematical Programming, 171*(1–2), 115–166. https://doi.org/10.1007/s10107-017-1172-1

Network for Greening the Financial System. (2025). *NGFS short-term climate scenarios: Technical documentation.*

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2(3), 21–41. https://doi.org/10.21314/JOR.2000.038

Task Force on Climate-Related Financial Disclosures. (2017). *Final report: Recommendations of the TCFD.* https://assets.bbhub.io/company/sites/60/2020/10/FINAL-2017-TCFD-Report-11052018.pdf

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., ... & Mann, G. (2023). *BloombergGPT: A large language model for finance. arXiv:2303.17564. https://arxiv.org/abs/2303.17564*

Appendix A. Notation and Optimizer Specification

A.1 Notation

Let *i* index securities (1…*N*), *t* monthly rebalance dates, and *k* Monte-Carlo scenarios. Returns are decimal.

- $r_{i,t+1}$: next-period total return of security *i*; $\mathbf{r}_{t+1} \in \mathbb{R}^N$ stacks all $r_{i,t+1}$.
- $\mathbf{w}_t \in \mathbb{R}^N$: portfolio weights at *t* (long-only unless noted).
- $\mathbf{w}_b$: benchmark weights; $\mathbf{B}$: factor loadings; $\mathbf{\Sigma}$: factor-aware covariance.
- $S_{i,t}$: signals (LLM-derived sentiment; scenario-adjusted features).
- Tracking error $\mathrm{TE}(\mathbf{w}) = \sqrt{(\mathbf{w}-\mathbf{w}_b)^\top \mathbf{\Sigma} (\mathbf{w}-\mathbf{w}_b)}$.
- One-way turnover $\tau_t = \frac{1}{2}\sum_i |\, w_{i,t} - w_{i,t-1} \,|$.
- Benchmark beta $\beta(\mathbf{w})$(target 1.00 ± 0.02).
- $\mathrm{CVaR}_\alpha(\cdot)$: conditional value-at-risk at confidence α.
- $\hat{P}$: empirical return distribution; $W_1(\cdot,\cdot)$: Wasserstein-1 distance with ℓ_2 ground metric (covariance-whitened).
- Ambiguity set $\mathcal{P}_\varepsilon = \{P : W_1(P, \hat{P}) \le \varepsilon\}$.
- Confidence gating for text: accept if $q \ge \theta_{\text{accept}} = 0.70$; human-in-the-loop (HITL) if $0.40 \le q < 0.70$; discard if $q < 0.40$.

A.2 Core parameters (Table A1)

Symbol	Definition
α	CVaR confidence level (0.95)
λ	Risk–return trade-off parameter (4.0)
ε	Wasserstein-1 radius for DRO (0.50, return units)
TE	Tracking-error band (reported per test)
TC	Cost model (C_t=\sum_i \big(c_0,
$\mathrm{TO}_{\max}$	Monthly turnover cap (30%)
β_{target}	Benchmark beta 1.00 ± 0.02
σ	Daily volatility estimate (60-day)
$\mathrm{ADV}_{\text{cap}}$	Max trade size 5% ADV

A.3 EU-Taxonomy/SFDR constraints

- **Taxonomy alignment:** aligned revenue share $\ge 25\%$when posterior $P(\text{Aligned} \mid \text{evidence}) \ge 0.80$.
- **PAI carbon intensity:** portfolio carbon intensity ≤ benchmark –15%.
- **Exclusions:** controversial weapons; active UNGC violations (firm-level).

A.4 Evidence-weighted scoring (link to Section 3.3)

Document-level confidence $q$is a convex combination of calibrated components: $q = w_m \hat{p} + w_e p_{\text{ent}} + w_r r$, with $(w_m, w_e, w_r) = (0.40, 0.40, 0.20)$. Issuer-day polarity $\tilde{y}_{i,t}$ aggregates same-day documents via a Beta–Bernoulli update (prior $\mathrm{Beta}(2,2)$) and recency decay $\kappa(\Delta t) = e^{-\Delta t/180}$. The **evidence-weighted score** is.

$$\mathrm{ESGScore}_{i,t} = \mathbb{E}[\tilde{y}_{i,t}\,(1+\lambda_s s_{i,t})(1+\lambda_d D_{i,t}) \mid \mathcal{D}_{i,t}]\,,$$

with $\lambda_s = 0.15$, $\lambda_d = 0.05$, and $D_{i,t} \le 0.20$ (source-diversity cap).

A.5 Optimization problems

(i) Mean–CVaR program

$$\min_{\mathbf{w},\eta,\boldsymbol{\xi}} \quad \eta + \frac{1}{(1-\alpha)K}\sum_{k=1}^{K}\xi_k - \lambda\,\mathbb{E}[\mathbf{w}^\top \mathbf{r}]$$

$$\text{s.t.} \quad \xi_k \ge -\mathbf{w}^\top \mathbf{r}^{(k)} - \eta,\ \xi_k \ge 0\ \forall k;$$

$1^\top w = 1$, $w \ge 0$;

country/sector bands ±5 p.p., $w_i \le 3\%$;

$\|B^\top(w - w_b)\|_\infty \le 0.20$ (factor bands);

TE(w) within preset band, $\tau_t \le 30\%$; $\beta(w) \in [0.98, 1.02]$;

EU-Taxonomy/PAI constraints and exclusions (above);

Costs C_t enforced ex-ante via budget and applied ex-post in P&L.

(ii) DRO variant

$$\min_{\mathbf{w}} \sup_{P\in\mathcal{P}_\varepsilon} \mathrm{CVaR}_\alpha(\mathbf{w}^\top \mathbf{r}) - \lambda\,\mathbb{E}_P[\mathbf{w}^\top \mathbf{r}],$$

with $\mathcal{P}_\varepsilon$ as above; constraints identical to the mean-CVaR program.

Appendix B. Robustness Checks (what you report)

- **Regional & language residual tests:** difference-in-means of sentiment by region controlling for severity (ordinal logit residuals); flag if $|\,t\,| > 2$.
- **Coverage & entity-resolution:** precision/recall by region and by language; per-sector error tallies.
- **Post-cost factor neutrality:** report Barra-style exposures and tracking error after costs.
- **Walk-forward / nested CV:** expanding window with quarterly re-fit; plus last-24-month re-runs using: (i) alternate LLM, (ii) no-RAG, (iii) prompt variants; median Hellinger distance target ≤ 0.15.
- **Cost/turnover sensitivity:** linear cost 5–25 bps, $k \in [0.05, 0.20]$, turnover caps 20-40%.
- **NGFS stress:** Orderly/Disorderly/Hot-House World (3-year horizon), with sector & idiosyncratic shock mapping.

Appendix C. Data Licensing and Reproducibility

- **Licensed inputs:** ESG vendor datasets (e.g., Refinitiv, Sustainalytics) and market data (constituents, prices). Redistribution restricted; results are presented in aggregate.
- **Texts (news/NGO/filings):** processed under applicable licenses/terms; web content is stored as hashes plus short excerpts for audit purposes only.

- **Shareable replication pack:** configuration files (seeds, hyperparameters, constraints), prompt templates, retrieval logs with source hashes/timestamps, model cards and version hashes, and runners to reproduce **Tables 1-5** and **Figures 1, 5-6** with placeholders. **Full replication** requires the reader's own licensed access.

Appendix D. NLP Baselines Benchmarking

Dictionary sentiment, LDA topic factors, and a zero-shot LLM are benchmarked against RAG-LLM. Under the acceptance gate $q \geq 0.70$, evidence-grounded retrieval raises mean IC and IC stability at materially higher coverage (see Figure 6 and IC statistics in the main text).