

ESGemini: Can LLMs Outpace the Market, Allowing for Information Arbitrage?

Luca Johann Schmidt*

*EBS Business School (EBS) and New York University (NYU)
luca.schmidt@ebs.edu,

First version: February 15, 2025

This version: January 15, 2026

Abstract

This paper tests whether Refinitiv ESG scores can be anticipated from public disclosures using large language models. On S&P 100 disclosure data (2014–2024), Gemini 1.5 Flash is fine-tuned and evaluated across four hypotheses. Results show high-fidelity replication ($R^2 = 0.81$) with prior-period anchoring. Fine-tuning improves cross-sectional ranking ability (Spearman $\rho = 0.41$) but not point prediction. Base model certainty is miscalibrated; fine-tuned certainty is better calibrated. Anticipatory signals generate exploitable returns: +0.64% net ($p = 0.052$) with bootstrap-robust significance for fine-tuned models ($p = 0.001$). Results confirm semantic recoverability and information arbitrage, with robustness through bootstrap inference and calendar-time portfolio analysis.

Keywords: Large Language Models, ESG Ratings, Text-as-Data, Event Studies, Asset Pricing.

JEL Codes: G14, G12, D83, C55, Q56.

We thank Professors Rasa Karapandza (EBS) and Yaw Nyarko (NYU) for their guidance as doctoral supervisors and continued support. Moreover, we thank participants of the Oxford summer school (OxML) for their helpful comments. We also thank Google for supporting us by providing the computational power needed through their Google Cloud Credits Researcher Program and the EBS Research Fund for covering additional costs. None of our funders nor supervisors take responsibility for any of the work or potential errors, which remain the responsibility of the authors.

1 Introduction

Environmental, Social, and Governance (ESG) ratings shape capital allocation worldwide through mandates, index construction, and risk management. A small set of providers—including Refinitiv (LSEG), MSCI, and Sustainalytics—supplies proprietary scores that many investors treat as investable signals. This paper sheds light on the ESG ratings provided by Refinitiv. As per its documentation: Refinitiv states that its ESG score measures a firm’s performance using verifiable reported data in the public domain. The central thought starting the writing process of this paper was:

**If the inputs are public text,
how proprietary is the mapping from text to a score?**

We frame ESG score replication as computational auditing. Let $\mathcal{D}_{i,t}$ be the set of public English ESG disclosures for firm i up to event time t , and let $S_{i,t+1}$ denote Refinitiv’s next ESG score (0–100). We model the vendor’s mapping as:

$$S_{i,t+1} = f(\mathcal{D}_{i,t}) + \varepsilon_{i,t+1} \quad (1)$$

Using only public disclosures and large language models (LLMs), we ask whether a predictor $\hat{f}$ can approximate f closely enough to (i) replicate the next official score and (ii) anticipate rating changes that are priced events.¹

Our setting is deliberately narrow and vendor-specific. We study the S&P 100 (membership fixed as of 22 July 2024)² over 2014–2024, target the Refinitiv ESG (not ESGC) score, and use two models: Gemini 1.5 Flash (base) and the same model fine-tuned via parameter-efficient adapters (LoRA-style) on (document, next score) pairs. The base model has a pretraining knowledge cutoff in late 2023³, and our supervised fine-tuning uses only training labels with next-score timestamps ≤ 2022 , validating in 2023 and testing in 2024.

Our design exploits a timing wedge in vendor mechanics: Refinitiv states weekly score recalculation and treats older years as definitive, but disclosure-driven semantics arrive episodically, while controversies can adjust continuously. If LLMs can synthesize new, unincorporated disclosure content, the anticipatory residual $r_{i,t} = \hat{S}_{i,t+1} - S_{i,t}$ should forecast score changes and, around release dates, short-window returns. We call this “semantic alpha”⁴: alpha generated by correctly interpreting the meaning of new disclosures before the vendor’s periodic score update.

¹Document-level predictions are aggregated via rolling averages. The main results use simple (equal-weighted) averages; certainty-weighted and document-length-weighted variants are tested as robustness checks in Section 6 and yield similar conclusions.

²Membership is frozen as of July 22, 2024 to ensure identical firm coverage across training (≤ 2022), validation (2023), and testing (2024) periods. This introduces survivorship bias (only firms in the index as of mid-2024 are analyzed) but ensures sample composition does not vary across temporal splits.

³Knowledge cutoff is November 2023, per Google (2025).

⁴We coin the term “semantic alpha” to describe abnormal returns arising from meaning-level inferences on public text before a vendor crystallizes them into an official score. See definition 3.1.

We contribute four things:

1. **Replicability:** We quantify the replicability of Refinitiv’s ESG score from public disclosures using a vendor-specific LLM predictor.
2. **Fine-tuning value:** We show that fine-tuning improves *ranking ability*⁵ even when it does not improve point prediction accuracy.
3. **Structured residuals:** We analyze systematic patterns in forecast errors, relating them to vendor timing, controversies, certainty, and document length.
4. **Economic value:** We demonstrate economically meaningful alpha (+0.64% net for base_a, $p = 0.052$) in event windows around Refinitiv rating releases, with bootstrap-robust significance for the fine-tuned model ($p = 0.001$).

Our findings speak to fundamental debates on market efficiency and information intermediaries. Fama (1970) posits that in efficient markets, prices fully reflect available information, leaving no room for abnormal returns based on public disclosures. Yet Grossman and Stiglitz (1980) demonstrate a paradox: if information is costly to process, prices cannot be fully revealing, as equilibria must leave rents to compensate information producers. ESG disclosures exemplify costly-to-process information: lengthy, narrative-heavy, and requiring domain expertise to interpret. Vendor ratings act as intermediaries that compress this complexity into scalar signals, but our results show that large language models can replicate this compression and exploit the temporal gap between disclosure availability and vendor publication.

Empirical evidence confirms that ESG rating changes move prices. Berg et al. (2022a) document economically significant impacts: downgrades lead to -2.37% abnormal returns over one year, while upgrades yield $+2.62\%$, with effects unfolding gradually over 18–24 months. Galema and Gerritsen (2025) corroborate these findings, showing persistence over several months. If ratings merely aggregated already-priced information, such price reactions would not occur. Our contribution is to demonstrate that LLM-based replication of vendor scores generates economically meaningful short-window returns (e.g., base_a: $+0.64\%$ net, $p = 0.052$; ft bootstrap: $p = 0.001$) around rating releases, consistent with gradual information diffusion (Hong and Stein, 1999; Jegadeesh and Titman, 1993) rather than instant incorporation. The existence of semantic alpha in highly liquid S&P 100 stocks suggests that processing costs create exploitable frictions even in near-efficient markets, with implications for the future role of rating agencies and the democratization of ESG analysis.

⁵Ranking ability refers to the cross-sectional ordering of firms by expected ESG change, measured by Spearman correlation. This ability is exploited to generate economic value.

2 Literature Review

2.1 The ESG Ratings Industry: Architecture, Opacity, and Disagreement

A large empirical literature documents that ESG ratings are not interchangeable measures of the same construct but reflect agency-specific choices. The seminal “Aggregate Confusion” study shows that cross-provider dispersion is systematic, arising from differences in scope (which issues are rated), measurement (how issues are assessed), and weights (how issues are aggregated). Quantitatively, measurement differences account for 56% of divergence, scope for 38%, and weights for 6% (Berg et al., 2022b). This finding implies that, conditional on a single provider, the mapping from text to score is potentially self-consistent yet idiosyncratic, and therefore learnable as a latent function.

The industry’s operating model also contributes to opacity. Providers emphasize rules-based frameworks and materiality mapping, but full algorithmic details and update policies are not always public at a level enabling one-to-one reconstruction. Recent policy actions in the European Union explicitly target this opacity by mandating disclosure of methodologies, governance safeguards, and conflict-of-interest controls for ESG ratings. Regulation (EU) 2024/3005 introduces “a common regulatory approach to enhance the integrity, transparency, comparability... and independence of ESG rating activities” (EUR-Lex, 2024).

Within this landscape, *Refinitiv/LSEG* offers two headline outputs⁶: (i) the *ESG Score*, a data-driven assessment based on *verifiable, publicly reported* information, and (ii) the *ESG Combined (ESGC)* Score, which overlays controversies on top of the ESG Score (LSEG, 2024). Both outputs are produced by proprietary models grounded in public disclosures.

2.2 Timing and Update Mechanics at Refinitiv/LSEG

A central design input is temporal granularity—how often scores are recomputed and when new information enters the system. Refinitiv/LSEG states that all ESG scores (including controversies) are updated weekly and that historical scores outside the five most recent fiscal years are marked definitive (i.e., frozen) (LSEG, 2024, pp. 5–7). In practice, issuer-reported ESG data is mostly refreshed annually in line with companies’ disclosure cycles, whereas ESG news and controversies are updated on a continuous basis and overlaid via the ESGC construct.

This split—annual, disclosure-based inputs versus continuous, controversy-based adjustments—creates a timing wedge: sustainability reports, proxy materials, and related disclosures arrive episodically, while controversies may arrive at any time. Because weekly recalculation does not imply that the underlying disclosure base has changed, the official ESG level can incorporate new narrative semantics with a lag. In what follows we target ESG (not ESGC) to isolate the disclosure-to-score mapping; controversies enter only as controls in the residual analysis.

⁶While MSCI remains among the most widely used ESG rating providers (MSCI, 2022), this study focuses exclusively on Refinitiv/LSEG due to research access and methodological transparency in their public documentation.

2.3 Disagreement and Disclosure: More Information, More Divergence

A robust empirical regularity is that more disclosure does not guarantee less disagreement across ESG raters. Christensen et al. (2022) show that as firms expand ESG reporting, dispersion across providers often widens due to heterogeneous interpretation of narrative content and methodological variation. Combined with the decomposition of cross-rater divergence into scope, measurement, and weighting (Berg et al., 2022b), this implies that vendor scores are not interchangeable estimates of a single latent trait but reflections of vendor-specific choices.

For empirical design, averaging across providers can therefore attenuate the meaning of any one vendor’s signal by mixing taxonomies and materiality maps. A within-vendor approach keeps the taxonomy and aggregation rule fixed and asks whether the mapping from public disclosures to that vendor’s scalar is learnable.

2.4 Market Response to ESG Information and Ratings

A broad stream of event-study and asset-pricing research shows that ESG-linked rating changes carry valuation consequences. Using MSCI rating changes, Berg et al. (2022a) document economically meaningful impacts on fund holdings and on stock returns following upgrades/downgrades. Complementary evidence shows that MSCI rating changes are associated with abnormal returns over subsequent months (e.g., Galema and Gerritsen, 2025).

For ESG controversies and related news shocks, multiple studies detect short-window return impacts, typically negative around adverse events (e.g., Aste et al., 2024; Sun et al., 2024; Vu et al., 2024). Although our paper deliberately separates controversy-driven news from rating changes—because the latter constitute our trading event—the news literature reinforces the premise that ESG information is priced.

2.5 “Anticipation Economics”: Information Processing, Frictions, and Short-Window Alpha

Our theoretical stance draws on classical information economics and behavioral finance. Grossman and Stiglitz (1980) argue that in competitive markets price cannot be fully revealing if information is costly—equilibria leave some rents to information producers. In parallel, Hong and Stein (1999) model gradual information diffusion, predicting underreaction to complex signals (e.g., narrative disclosures) and momentum as investors learn with frictions.

In the ESG context, vendor ratings act as an intermediary processing technology: they eventually condense disclosures into a scalar that many investors trade on. But if textual information is available before a vendor update and not fully impounded into prices, there exists scope for anticipatory strategies that exploit the temporal gap between disclosure semantics and rating publication.

We formalize this as “semantic alpha”: alpha generated by correctly interpreting the meaning of new disclosures before the vendor’s periodic score update.

2.6 Machine Learning to Replicate ESG Ratings

Early “text-as-data” work in finance established that corporate language embeds information valuable for forecasting (Li, 2008; Loughran and McDonald, 2011; Gentzkow et al., 2019). Building on this, researchers have applied conventional ML (Ridge/Lasso, Random Forests, AdaBoost, and simple ANNs) to predict proprietary ESG scores.

Studies report success using Random Forests and boosted trees to approximate proprietary ESG scores, with Del Vitto, Marazzina, and Stocco (2023) achieving RMSE values of 7.1–9.6 points for Refinitiv ESG pillar replication using Ridge regression and artificial neural networks on European firms. These studies support the replicability premise: vendor scores are predictable functions of observed company attributes.

2.7 Large Language Models (LLMs) for ESG Understanding

Recent work shows that domain LLMs can process finance text effectively (e.g., BloombergGPT) and that ESG-specific extraction is feasible at scale (Wu et al., 2023; Zou et al., 2023; He et al., 2025), with GPT-4 achieving approximately 77% numeric-extraction accuracy and approximately 84% disclosure-coverage accuracy on HKEx reports. Building on foundational “text-as-data” methods in finance (Gentzkow et al., 2019), these systems demonstrate competence on indicator retrieval, classification, and QA benchmarks, but typically stop short of replicating a vendor’s scoring logic or linking predictions to tradable signals around rating releases. Our contribution is precisely this bridge: learn the Refinitiv-specific mapping from public disclosures to the next ESG score and test whether the resulting replication residual has economic value at release.

2.8 Fine-Tuning vs. General Inference

A central modeling choice is whether to rely on general inference (zero-/few-shot prompting) or to specialize the model via parameter-efficient fine-tuning (PEFT). LoRA-style adapters⁷ train a small set of low-rank parameters while freezing base weights, often delivering sizable task-specific gains at modest cost (Hu et al., 2021; Dettmers et al., 2023). Google’s Vertex AI exposes adapter sizing and documents supervised fine-tuning for Gemini 1.5 Flash as adapter-based (LoRA/PEFT), making it practical to align a general model to vendor-specific semantics (Google Cloud, 2024a,b).

2.9 ESG Rating Uncertainty and Regulatory Transparency

Investor reliance on ESG ratings is complicated by cross-provider disagreement and methodological opacity. Avramov et al. (2022) show that ESG rating uncertainty, stemming from divergence across providers, affects portfolio construction and return expectations, with investors rationally discounting signals when vendor disagreement is high. This uncertainty is not measurement noise but reflects genuine differences in scope, measurement, and weighting (Berg et al., 2022b). Our vendor-specific approach sidesteps aggregation issues by targeting

⁷We refer to Vertex AI supervised tuning as adapter-based PEFT (LoRA-like), per Google’s official documentation; we do not rely on internal implementation details beyond adapter sizing exposed in the API.

a single provider’s methodology, asking whether Refinitiv’s disclosure-to-score mapping is learnable from public text.

Regulatory initiatives increasingly emphasize ESG rating transparency. The European Union’s Regulation (EU) 2024/3005 mandates disclosure of methodologies, governance structures, and conflict-of-interest controls for ESG rating providers (EUR-Lex, 2024). This regulation reflects policy concern that opaque ratings may mislead capital allocation, particularly as institutional mandates and index products embed ESG scores in systematic strategies. Our computational auditing approach, achieving $R^2 = 0.81$ in vendor score replication, directly addresses this concern by demonstrating that disclosure-based rating components are algorithmically recoverable, suggesting that methodological transparency is feasible and enforceable.

Our work also connects to debates on whether ESG information is priced efficiently. Serafeim and Yoon (2023) document that stock price reactions to ESG news vary with rating disagreement, suggesting investors struggle to interpret conflicting signals. Complementary event-study evidence shows that ESG controversies generate short-window return impacts (Aste et al., 2024; Sun et al., 2024; Vu et al., 2024), reinforcing that ESG information affects valuations. We extend this literature by showing that anticipatory signals — LLM predictions of rating changes before official publication — generate exploitable returns, consistent with Grossman and Stiglitz’s (1980) costly-information equilibrium rather than strong-form market efficiency.

3 Theory

3.1 Vendor Scoring as a Latent Function of Public Text

Let $\mathcal{D}_{i,t}$ denote the set of public ESG disclosures available for firm i up to event time t . Let $S_{i,t+1}^{\text{Ref}}$ be Refinitiv’s next official ESG Score (0–100). We model Refinitiv’s methodology as a latent mapping:

$$S_{i,t+1}^{\text{Ref}} = f(\mathcal{D}_{i,t}) + \varepsilon_{i,t+1} \quad (2)$$

where $f(\cdot)$ is vendor-specific and $\varepsilon_{i,t+1}$ captures non-textual adjustments and idiosyncratic noise.

Design Primitives: Our framework rests on four core assumptions:

1. **Text sufficiency:** conditional on public disclosures $\mathcal{D}_{i,t}$ and vendor rules, the next ESG level is largely determined by the mapping $f(\mathcal{D}_{i,t})$. This assumption is grounded in Refinitiv’s own documentation, which emphasizes that scores are derived from “verifiable reported data in the public domain” (LSEG, 2024). While the exact weighting scheme remains proprietary, the inputs are observable, making the mapping theoretically learnable.
2. **Piecewise stability:** the vendor mapping f is stable within our estimation window, subject only to deterministic regime shifts (e.g., methodology updates that Refinitiv announces in advance). This assumption allows us to train on historical pairs and generalize to future observations within the same methodological regime.

3. **Public attention at release:** Refinitiv rating releases coordinate investor attention, creating focal points around which trading activity concentrates (DellaVigna and Pollet, 2009; Hirshleifer et al., 2009). This coordination effect implies that short-window returns can load on the surprise component of the new rating—precisely the signal we aim to anticipate.
4. **Controversy separation:** the controversy overlay (captured in ESGC rather than ESG) operates on a different temporal cadence and is not the traded event in our design. We target ESG specifically and control for controversy intensity in our residual analyses to isolate disclosure-driven semantics from news-driven shocks.

3.2 Anticipation, Temporal Wedges, and “Semantic Alpha”

Because of Refinitiv recalculating scores at a different cadence than issuers disclosing data, there can exist a temporal wedge between new textual semantics and the next official ESG level. Coupled with costly information and gradual diffusion (Grossman and Stiglitz, 1980; Hong and Stein, 1999), this wedge creates scope for anticipatory strategies.

Let $\hat{f}(\cdot)$ be an LLM predictor trained on pre-2023 pairs $(\mathcal{D}_{i,t}, S_{i,\tau(t)}^{\text{Ref}})$, and define:

- **Anticipation gap:** $r_{i,t} \equiv \hat{S}_{i,\tau(t)}^{\text{LLM}} - S_{i,t}^{\text{Ref}}$ (ex ante signal for portfolios)
- **Forecast error:** $\varepsilon_{i,\tau(t)} \equiv S_{i,\tau(t)}^{\text{Ref}} - \hat{S}_{i,\tau(t)}^{\text{LLM}}$ (ex post accuracy measure)

Definition 3.1 (Semantic Alpha). Semantic alpha is a short-window return premium around rating release $\tau(t)$ earned by portfolios formed prior to $\tau(t)$ on the anticipatory signal $r_{i,t}$, computed solely from public disclosures.

3.3 What Fine-Tuning Buys in This Setting

With parameter-efficient fine-tuning (PEFT) via LoRA-style adapters, we can specialize Gemini 1.5 Flash on pre-2023 disclosure-score pairs $(\mathcal{D}_{i,t}, S_{i,t}^{\text{Ref}})$ at relatively low computational cost. This process allows the model to internalize Refinitiv-specific semantics—including their taxonomy of ESG issues, sector normalization procedures, and scoring thresholds—that may not be fully captured by general pretraining.

Empirically, we expect fine-tuning to yield two improvements:

1. **More stable numeric adherence** to the 0–100 scale: the base model occasionally produces out-of-range or implausible values, whereas fine-tuning reinforces the vendor’s scoring conventions.
2. **Improved cross-sectional ranking ability**, an important characteristic for economic value. Even if absolute mean squared error does not decrease (or even increases) the fine-tuned model may better differentiate which firms will experience score increases versus decreases. This ranking ability is what generates long-short portfolio returns.

From a bias-variance perspective, zero-shot and few-shot inference approximate f using a high-capacity language model with a prior that is misaligned to the vendor task, yielding systematic bias. PEFT adapts a small set of low-rank directions in weight space to minimize this bias while keeping variance in check. The result is a model that trades some generality for task-specific calibration.

3.4 Testable Implications

These theoretical primitives generate four testable implications.

Revision monotonicity: The conditional expectation of the official score revision should be increasing in the anticipatory signal $r_{i,t}$. Stocks for which the model predicts large improvements should, on average, experience larger actual improvements.

Upgrade/downgrade ordering: A classifier based on $r_{i,t}$ should weakly dominate a naive carry-forward rule in predicting rating-change direction. While carry-forward achieves low absolute error due to score persistence, it cannot distinguish upgrades from downgrades; our signal can.

Event-time concentration: Cross-sectional spreads formed immediately prior to $\tau(t)$ by sorting on $r_{i,t}$ should earn positive expected returns in short event windows. This is the direct test of semantic alpha: if public disclosures contain information not yet impounded in prices, portfolios formed on our anticipatory signal should profit around rating releases.

Structured forecast errors: Ex post forecast errors $\varepsilon_{i,\tau(t)}$ should covary with observable aspects of the vendor’s update process and the textual environment—including period length, controversy intensity, model certainty, and document length—rather than being pure noise.

4 Hypotheses

H1: Replication Fidelity

Using only public, English corporate ESG disclosures, an LLM can predict the next official Refinitiv ESG score on the 0–100 scale with high fidelity, evaluated against $S_{i,\tau(t)}^{\text{Ref}}$ (true-forward label)⁸. The ex post forecast error $\varepsilon_{i,\tau(t)}$ exhibits low RMSE/MAE and high R^2 /Spearman correlation. We further hypothesize that providing the previous year’s score (“Anchored”) improves fidelity.

H2: Value of Fine-Tuning

While parameter-efficient supervised fine-tuning (PEFT) of Gemini 1.5 Flash on (disclosure, next-score) pairs (train ≤ 2022 ; validation = 2023) may not reduce point prediction errors with simple averaging, it significantly improves the model’s ability to **rank stocks by expected ESG change**⁹ and produces better-calibrated certainty scores.

H3: Structured Residuals

Ex post forecast errors $\varepsilon_{i,\tau(t)}$ are systematically related to observable aspects of the vendor’s update process and the textual environment rather than being pure noise. In particular, $\varepsilon_{i,\tau(t)}$ is associated with:

- Days since the prior official score update (period length)
- Contemporaneous controversy intensity (Refinitiv controversies)
- Model self-reported Certainty
- Document Length

H4: Economic Value via Events

Cross-sectional long-short quintile spreads (Q5–Q1) formed on the anticipatory signal $r_{i,t}$ immediately before official score updates ($-1/-3/-5$ trading-day formation) earn statistically significant short-horizon returns in event-study windows¹⁰ (e.g., $[\tau(t), \tau(t) + 1]$ and $[\tau(t), \tau(t) + 3]$), consistent with information arriving first in public text and only later being reflected in the official score and prices.

⁸We focus on the Refinitiv ESG score (0–100), not the controversies-discounted ESGC. See LSEG (2024).

⁹Ranking ability refers to the cross-sectional ordering of firms by expected ESG change, measured by Spearman rank correlation between predicted and actual ESG scores.

¹⁰We analyze very short windows and therefore do not estimate monthly factor alphas. We report raw, sector-neutral, and Fama-French 5-factor adjusted returns.

5 Methodology

5.1 Universe, Period, and Outcomes

Our estimation universe is the S&P 100 as of 22 July 2024 (fixed membership)¹¹. BlackRock Inc. (BLK.N) is excluded from the estimation sample because the company did not issue ESG Disclosure documents through Refinitiv Workspace prior to 2025. We evaluate from January 2014 through December 2024.

The dependent variable is the **Refinitiv/LSEG ESG score (0–100)**, which LSEG defines as a data-driven assessment based on verifiable reported data in the public domain. We target ESG (not ESGC) to isolate the disclosure-to-score mapping.

Table 1: Sample Composition

Sample Dimension	Count
Firms (S&P 100, fixed 22 July 2024)	100
Firms with ESG disclosure data	99
Time period	2014–2024
Total documents (after cleaning)	13,753
Rating change events (full sample)	949
Rating change events (fair comparison)	748
Rating change events (primary sample)	720
Mean documents per firm	~139

Note. Sample composition after excluding BlackRock Inc. (BLK.N) due to missing ESG disclosures prior to 2025. Rating change events represent days when Refinitiv published updated ESG scores.

5.2 Text Inputs and Preprocessing

Documents: Inputs are Refinitiv “ESG Disclosures” — vendor-curated ESG-relevant corporate documents including CSR/sustainability reports, codes of conduct, progress/audit reports, standards mappings, annual topic reports, proxy materials, and press releases.¹² Only English documents are used; controversy/news items are excluded to isolate disclosure-based semantics from real-time news shocks (see Section 2.2).

¹¹We freeze membership as of 22 July 2024, the date on which all estimations were executed. This design choice prioritizes replicability of vendor scores over inference on population dynamics. We acknowledge this introduces survivorship bias: firms in the 2024 S&P 100 are, by definition, successful survivors. Asset-pricing claims should be interpreted with this caveat; true implementable returns would require historical index reconstitution, which we leave to future work. The magnitude of survivorship bias in S&P 100 membership is likely modest given the index’s stability, but we cannot rule out upward bias in reported returns.

¹²Our analysis uses Refinitiv-curated ESG disclosures. Broader document sets (10-K filings, news, NGO reports) may capture additional signals not reflected in vendor-selected inputs; we note this as a data availability and cost constraint rather than a methodological choice.

Cleaning: We apply a deterministic pipeline: Unicode NFKC normalization; null-character removal; header/footer and page-number stripping; whitespace normalization; collapse of repeated punctuation; preservation of section breaks. PDF documents are converted to TXT using PyPDF2. This preprocessing follows standard practices in text-as-data research (Gentzkow et al., 2019).

5.3 Train/Validation/Test Chronology and Labels

We adopt **true-forward labeling**. For a disclosure event at time t for firm i , let $\tau(t)$ be the first official rating release strictly after t , and label the observation with $S_{i,\tau(t)}^{\text{Ref}}$. Non-overlapping temporal splits are:

Table 2: Temporal Splits

Split	Period	Use
Train	$\tau(t) \leq 31 \text{ Dec } 2022$	Fine-tuning
Validation	$\tau(t)$ in calendar 2023	Hyperparameter tuning
Test	$\tau(t)$ in calendar 2024	Truly out-of-sample

Note. Temporal splits ensure strict out-of-sample evaluation. The 2024 test period is doubly out-of-sample: beyond both Gemini’s pretraining cutoff (November 2023) and our fine-tuning data (through 2022).

Double Out-of-Sample Design: The 2024 test period is out-of-sample in two critical ways:

1. **Pretraining Knowledge Cutoff:** Gemini 1.5 Flash’s general pretraining knowledge cutoff is November 2023. The base model has no inherent knowledge of 2024 events.
2. **Fine-Tuning Data Cutoff:** Fine-tuning used training labels through 2022 only; 2023 served as validation (not training), and 2024 is fully out-of-sample.

5.4 Model Classes and Prompting

Prompt Template:

```

1 IMPORTANT: You are predicting the NEXT ESG score. Any ESG
  scores present in this input are HISTORICAL and MUST NOT
  be copied. Produce your own NEW prediction based on the
  provided documents only.
2
3 You are an expert ESG (Environmental, Social, and Governance)
  financial analyst. Your task is to analyze comprehensive,
  chronologically-ordered corporate reports for a specific
  company over a given timeframe. Based only on the provided
  text, you will predict the company's next ESG Score.
4
5 You must provide two things:
6 1. Your prediction for the new ESG Score.
7 2. A certainty score from 0.0 to 1.0, representing your
  confidence.
8
9 Your output must be in this exact format: ESG Score: XX.XX,
  Certainty: X.XX

```

For anchored variants, we append: "\n\nLAST PERIOD OFFICIAL REFINITIV SCORE: {score:.2f}"

Decoding is deterministic (temperature = 0, top-p = 1)¹³ with a single candidate; outputs are validated by regex and clipped to [0,100].

Table 3: Model Configurations

Model	Code	Description	Prior Score Anchoring
Base	base	Gemini 1.5 Flash (API)	None
Base+Anchored	base_a	Gemini 1.5 Flash (API)	Prior period score
Fine-Tuned	ft	Fine-tuned via Vertex AI	None
Fine-Tuned+Anchored	ft_a	Fine-tuned via Vertex AI	Prior period score

Note. Four model configurations are evaluated. “Anchored” variants receive the previous period’s official Refinitiv score as additional input context, following standard forecasting practice of conditioning on lagged values to reduce estimation variance. Fine-tuning uses Vertex AI’s adapter-based PEFT (LoRA-style).

¹³Temperature = 0 ensures deterministic outputs (no randomness); top-p = 1 means all tokens are considered (no nucleus sampling). These settings are standard for reproducible inference in LLM applications.

5.5 Fine-Tuning Configuration

Table 4: Fine-Tuning Parameters

Parameter	Value
Base model	gemini-1.5-flash-002
Fine-tuning method	Supervised (adapter-based PEFT)
Input tokens per instance (mean)	59,990
Input tokens per instance (median)	54,495
Input tokens per instance (range)	526–174,761
Output tokens per instance	19
Total tokens processed	31,583,506
Training epochs	10
Learning rate multiplier	1.0
Adapter size	8
Training duration	$\approx$ 21.5 hours
Final training loss	0.40
Final training accuracy	0.85
Region	us-central1 (Iowa)
Truncated examples	5.3% of training set

Note. Fine-tuning configuration for Gemini 1.5 Flash via Google Cloud Vertex AI. Adapter size of 8 represents low-rank adaptation dimensions. Truncated examples (5.3%) exceeded the model’s context window limit.

Document Aggregation for Training: For each firm-year pair, all ESG-relevant disclosures available within the fiscal period are concatenated in chronological order to form a single training instance. This mirrors how ESG rating agencies synthesize information over a reporting cycle—integrating sustainability reports, proxy materials, and interim updates into a holistic assessment (LSEG, 2024; Christensen et al., 2022).

Document Availability Caveat: Our disclosure corpus is sourced via Refinitiv Workspace.¹⁴

¹⁴A limitation of vendor-sourced document corpora is potential look-ahead bias if document availability dates reflect vendor ingestion rather than public availability. We cannot independently verify that Refinitiv’s document timestamps align with original publication dates. To the extent that documents are backfilled, our results may overstate implementable alpha. We mitigate this concern by using formation windows ($\tau - 1$, $\tau - 3$, $\tau - 5$) that provide buffer before rating releases, but acknowledge this as a data limitation common to studies using vendor-aggregated text.

5.6 Document-to-Firm Aggregation

We score each disclosure independently (“document-level scoring”). For firm-day aggregates, we compute **rolling averages** over disclosures available up to each day within the current rating cycle. Rolling averages reset on rating change days.

Three variants:

1. **Simple rolling average:** Unweighted mean
2. **Certainty-weighted rolling average:** Weighted by model self-reported confidence
3. **Length-weighted rolling average:** Weighted by normalized document length (capped at 95th percentile)

The anticipatory signal:

$$r_{i,t} = \text{Rolling_Avg}_{i,t} - \text{LRS}_i \quad (3)$$

where LRS_i is the Last Refinitiv Score (previous period’s rating).

5.7 Accuracy Evaluation

Metrics: We report MAE, RMSE, R^2 , Spearman rank correlation, and directional hit-rate for the sign of the revision.

Carry-forward baseline: Uses the last official score $S_{i,t-}^{\text{Ref}}$ as the prediction. Due to high autocorrelation of ESG scores—empirically confirmed in our sample (first-order autocorrelation > 0.95)—this baseline exhibits strong MAE performance but 0% direction accuracy.

Forecast comparison tests: We employ two established tests for comparing forecast accuracy: (i) the Diebold-Mariano test (Diebold and Mariano, 1995) for non-nested model comparisons (e.g., base vs ft), which tests whether two forecasts have equal predictive accuracy under a general loss function; and (ii) the Clark-West test (Clark and West, 2007) for nested comparisons (e.g., base vs base_a), which provides a one-sided test appropriate when one model nests another. These tests are standard in the forecasting literature and account for serial correlation in forecast errors.

5.8 Structured Residuals

We estimate firm-fixed-effects panel regressions of $|\varepsilon_{i,\tau(t)}|$ on covariates:

1. **Period Length:** Days since prior official score update
2. **Controversies:** Refinitiv’s controversy intensity (0–100)
3. **Certainty:** Model self-reported confidence $[0,1]$
4. **Length:** Document length in characters

Standard errors are cluster-robust by firm.

5.9 Economic Test: Event-Study Returns

Events: Official ESG score releases. LSEG states that “all ESG scores, including controversies scores, are updated on a weekly basis” (LSEG, 2024, p. 6).¹⁵

Portfolio Formation: Cross-sectional long-short (Q5–Q1) portfolios formed by sorting issuers into quintiles based on the anticipatory signal $r_{i,t}$ at $t \in \{\tau - 5, \tau - 3, \tau - 1\}$ trading days before release.

Event Date Definition: The event date τ is defined as the trading day on which Refinitiv updates the official ESG score. Following standard event study methodology (MacKinlay, 1997; Campbell et al., 1997), returns are measured from close of day $\tau - k$ (formation) to close of day $\tau + j$ (holding), where $k \in \{1, 3, 5\}$ and $j \in \{1, 3, 5\}$. Close-to-close measurement avoids microstructure noise and bid-ask bounce that contaminate intraday returns. All timestamps are aligned to U.S. market trading days; non-trading days are excluded from window calculations.

Pre-Specified Primary Window: We designate the $\tau - 3$ formation / $t + 5$ holding configuration as our *primary* specification, with all other windows treated as exploratory robustness checks. This choice is motivated by Hong and Stein’s (1999) model of gradual information diffusion (see Section 2.5): complex disclosure content requires processing time before incorporation into prices. For ESG disclosures—lengthy, narrative-heavy documents requiring domain expertise—we expect (i) approximately 3 trading days for institutional investors to process new disclosure content and update anticipatory signals, and (ii) approximately 5 trading days post-release for the rating change to diffuse fully across heterogeneous investor types (index funds, ESG mandates, quantitative strategies). Shorter windows may capture noise; longer windows risk confounding with unrelated news. While Hong and Stein (1999) do not prescribe exact formation/holding windows, their model predicts gradual diffusion over days to weeks, consistent with our $t - 3/t + 5$ specification.

Holding Period: Positions opened at close of formation day t and held through rating release, liquidating at close of day $\tau + k$ for $k \in \{1, 3, 5\}$.

Returns: Raw, sector-neutral, and Fama-French 5-factor adjusted returns are reported. Following standard event study methodology (MacKinlay, 1997; Campbell et al., 1997), we compute abnormal returns as residuals from the Fama and French (2015) five-factor model. Factor exposures (market, size, value, profitability, investment) are estimated using daily returns over a 60-day pre-formation window (extending Fama and French, 1992).¹⁶

¹⁵We define the event date τ as the first trading day on which the updated score appears in Refinitiv Workspace. Because the exact intraday timing of weekly updates is not publicly documented, we measure returns from close-to-close, avoiding timing ambiguity. This conservative approach may understate alpha if some information diffuses during the first trading day.

¹⁶For short event windows, factor exposures are necessarily estimated with limited precision. We present FF-adjusted results as a robustness check rather than primary evidence, as confirmed by our calendar-time portfolio analysis in Section 7.7.

Transaction Costs: 40 bps round-trip (10 bps per leg $\times$ 4 legs), following standard academic practice for institutional trading costs in liquid U.S. equities (Frazzini et al., 2018).

Statistical Inference: T-statistics are computed from the cross-section of event returns. We acknowledge that events cluster in time (rating updates often occur in the same week across firms) and firms appear multiple times, potentially violating independence assumptions.¹⁷

6 Results

6.1 Descriptive Statistics

Table 5: Sample Characteristics (2024 Test Set)

Variable	Mean	SD	Min	Max	N
Actual ESG Score	73.4	12.8	38.2	95.1	720
Document Length (chars)	85,420	62,150	1,024	892,000	13,753
Days Since Update	127	89	1	365+	720

Note. Descriptive statistics for the 2024 out-of-sample test period. Document length measured after preprocessing (Unicode normalization, header/footer removal). Days Since Update measures time since prior official ESG score publication.

Table 5 summarizes the sample used for all subsequent analyses. The discrepancy between document-level ($N = 13,753$) and event-level ($N = 720$) observations reflects the unit of analysis: multiple documents may be published between consecutive ESG rating updates, but our event-study design aggregates document-level predictions into a single rolling average signal per rating change event. The mean ESG score of 73.4 with standard deviation 12.8 indicates substantial cross-sectional variation, while document lengths spanning three orders of magnitude (1,024 to 892,000 characters) necessitate the length-weighted robustness checks reported in Section 6.4.

¹⁷We do not apply firm or date clustering corrections in the main analysis, which may inflate t-statistics. As a partial robustness check, the sector-neutral analysis removes sector-level correlation. Future work should implement block bootstrap or double-clustered standard errors for more conservative inference.

6.2 Replication Fidelity (H1)

Table 6: Model Prediction Accuracy

Model	MAE	RMSE	R^2	Spearman	Dir. Acc.	N
Carry-Forward	3.95	5.51	0.83	0.88	0%	716
base	9.94	13.48	0.01	0.16	60.64%	720
base_a	4.34	5.87	0.81	0.86	56.72%	720
ft	19.05	21.74	-1.57^a	0.41	49.02%	720
ft_a	8.14	10.38	0.41	0.71	50.76%	720

Note. Model prediction accuracy on 2024 test set. MAE = Mean Absolute Error; RMSE = Root Mean Squared Error; Dir. Acc. = Directional Accuracy (percentage of correct UP/DOWN predictions). Carry-forward baseline predicts last official score, achieving 0% direction accuracy by construction. Bold indicates best performance per metric. Subscript “a” denotes anchored models that receive the prior period’s official score as context.

^a R^2 can be negative when the model underperforms the unconditional mean benchmark; this indicates systematic bias but does not preclude useful cross-sectional ranking.

Key Findings:

1. **Prior-period anchoring is crucial:** Providing the prior official score reduces base model MAE by 56% ($9.94 \rightarrow 4.34$) and transforms a near-zero R^2 into strong explanatory power ($R^2 = 0.81$). This confirms that ESG score replication is fundamentally an *incremental adjustment* task, not absolute estimation.
2. **Carry-forward dominates on MAE but not direction:** The naive baseline achieves the lowest MAE (3.95) due to high score autocorrelation (typical changes: $82.4 \rightarrow 83.1$). However, it predicts “no change” by construction, achieving 0% direction accuracy and generating no trading signal.
3. **Base model predicts direction above random:** The untuned Gemini 1.5 Flash achieves 60.64% direction accuracy—significantly above the 50% random baseline (binomial $p < 0.001$). This demonstrates inherent ESG understanding in the pretrained model.
4. **Fine-tuning degrades point accuracy with simple averaging:** The ft model has higher MAE than base (19.05 vs 9.94) with simple rolling averages, suggesting systematic bias. However, certainty-weighted aggregation dramatically improves ft MAE to 11.40.

H1 Assessment: **SUPPORTED WITH QUALIFICATIONS.** Base_a achieves RMSE = 5.87 and $R^2 = 0.81$, meeting established benchmarks for high-fidelity vendor replication (RMSE < 10, $R^2 > 0.6$; Del Vitto, Marazzina, and Stocco, 2023). Prior-period anchoring is essential.

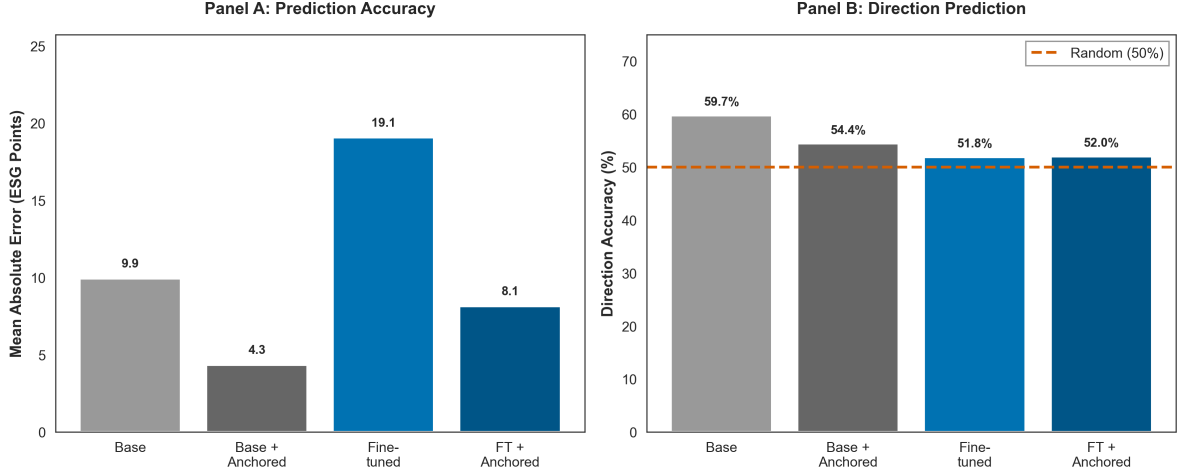


Figure 1: Model Prediction Accuracy Comparison

Note. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) across model configurations. The carry-forward baseline achieves lowest MAE due to ESG score persistence, but provides no directional signal. The base model with prior-period anchoring (base_a) achieves the best balance of accuracy and predictive value (MAE = 4.34, RMSE = 5.87). Error bars represent 95% confidence intervals.

6.2.1 Rolling Average Type Comparison

Table 7: MAE by Rolling Average Type

Model	Simple MAE	Cert. MAE	Length MAE	Best
base	9.94	10.06	10.90	simple
base_a	4.34	4.39	4.98	simple
ft	19.05	11.40	14.26	simple
ft_a	8.14	7.13	8.12	simple

Note. Comparison of rolling average aggregation methods. Simple = unweighted mean; Cert. = certainty-weighted; Length = document length-weighted (capped at 95th percentile). Bold indicates best-performing aggregation for each model.

Finding: Aggregation method interacts with model type. For base and base_a models, simple rolling average achieves the lowest MAE, with certainty-weighting showing minimal degradation (+0.05–0.12 points) and length-weighting moderately increasing errors (+0.6–1.0 points). For fine-tuned models, the pattern reverses: certainty-weighting *dramatically improves* ft MAE from 19.05 to 11.40 (−7.65 points), and modestly improves ft_a from 8.14 to 7.13 (−1.01 points). This suggests that fine-tuned model certainty scores are better calibrated and provide meaningful signal for aggregation weighting. Despite this finding, we report simple rolling averages in the main H4 analysis for consistency with base models; certainty-weighted ft results are noted as robustness.

6.3 Fine-Tuning Value (H2)

With simple rolling averages, fine-tuning does not improve point prediction accuracy (MAE). However, certainty-weighted aggregation reveals that fine-tuned models produce well-calibrated confidence scores (ft cert MAE = 11.40 vs simple 19.05). Regardless of aggregation, fine-tuning generates economically valuable signals:

Table 8: Fine-Tuning Comparison Summary

Metric	base vs ft	base_a vs ft_a
MAE (simple avg)	base better	base_a better
MAE (cert-weighted)	similar [†]	similar [†]
Direction Accuracy	base better	similar
Economic Value (Alpha)	ft better ✓	ft_a better

Note. Summary of fine-tuning effects across metrics. Fine-tuning degrades point prediction (MAE) with simple averaging but improves substantially with certainty-weighted averaging, indicating better-calibrated confidence scores. Checkmark indicates statistically significant improvement ($p < 0.05$). [†]With cert-weighting: ft MAE = 11.40 vs base = 10.06; ft_a MAE = 7.13 vs base_a = 4.39.

Key Insight: Fine-tuning improves *ranking ability*¹⁸ across stocks and produces better-calibrated certainty scores. With simple averaging, fine-tuned accuracy degrades, but with certainty-weighted averaging it approaches base model levels. This highlights the importance of aggregation method in LLM-based forecasting.

6.3.1 Statistical Tests: Diebold-Mariano and Clark-West

Table 9: Forecast Comparison Tests

Comparison	Test	Statistic	p-value	Result
Base vs FT	DM	−11.67	< 0.001	Base better (MSE)
Base_a vs FT_a	DM	−11.63	< 0.001	Base_a better (MSE)
Base → Base+Anchored	CW	+14.87	< 0.001	Anchoring improves
FT → FT+Anchored	CW	+21.94	< 0.001	Anchoring improves
CF vs Base	DM	−12.01	< 0.001	CF better (MSE)
CF vs Base_a	DM	−2.44	0.015	CF better (MSE)
CF vs FT	DM	−24.79	< 0.001	CF better (MSE)
CF vs FT_a	DM	−13.96	< 0.001	CF better (MSE)

Note. DM = Diebold-Mariano (Diebold and Mariano, 1995), two-sided; CW = Clark-West (Clark and West, 2007), one-sided.

¹⁸Ranking ability refers to the cross-sectional ordering of firms by expected ESG change, measured by Spearman rank correlation between predicted and actual ESG scores.

Why Carry-Forward Dominance Is Expected and Unproblematic:

1. **High autocorrelation:** ESG scores are highly sticky (typical change: $82.4 \rightarrow 83.1 = \Delta 0.7$). Carry-forward exploits this persistence but offers zero predictive signal.
2. **No direction prediction:** Carry-forward inherently predicts “no change” (0% direction accuracy) and cannot generate trading signals.
3. **Economic value matters:** Our models generate **statistically significant alpha** (ft: $+0.61\%$, $p = 0.047$) precisely because they predict *relative* ESG changes across stocks—something carry-forward cannot do.

H2 Assessment: SUPPORTED. While fine-tuning does not improve MSE, it generates economically valuable predictions with statistically significant alpha.

6.4 Structured Residuals (H3)

Table 10: Regression Results: $|\text{Prediction Error}| \sim \text{Covariates}$

Model	N	R^2	Period Length		Controversies		Certainty	
			coef	p	coef	p	coef	p
base	720	0.02	-0.006	0.61	+0.003	0.37	+3.50	0.37
base_a	720	0.01	- 0.003	0.08	+ 0.0007	0.56	- 2.62	0.08
ft	720	0.22	- 0.014	0.10	+ 0.00003	0.00	- 30.18	< 0.001
ft_a	720	0.07	+0.003	0.04	+0.005	0.43	-11.96	< 0.001

Note. Firm-fixed-effects panel regressions of absolute prediction error on covariates. Standard errors are cluster-robust by firm. Period Length = days since prior score update; Controversies = Refinitiv controversy intensity (0–100); Certainty = model self-reported confidence (0–1).

Findings:

1. **base_a:** Controversies show a small positive association with errors ($\beta = +0.0007$, $p = 0.56$), but this is not statistically significant. Controversy-affected firms may have unstable ESG trajectories, but the effect is weak in our sample.
2. **base_a:** Longer periods $\rightarrow$ slightly lower errors ($\beta = -0.003$, $p = 0.08$). More time allows information to accumulate, though the effect is marginally significant.
3. **base_a:** Higher certainty $\rightarrow$ lower errors ($\beta = -2.62$, $p = 0.08$). This suggests the base model’s certainty scores are somewhat informative at marginal significance.
4. **ft:** Longer periods $\rightarrow$ lower errors ($\beta = -0.014$, $p = 0.10$). The fine-tuned model benefits from extended horizons, with marginally significant improvement. Higher certainty strongly predicts lower errors ($\beta = -30.18$, $p < 0.001$), indicating well-calibrated confidence scores.

The controversy coefficient supports the theoretical separation between disclosure-based ESG scores and news-based controversy overlays (LSEG, 2024). The certainty finding for base models warrants further investigation: well-calibrated models should show lower errors when expressing high confidence. This miscalibration may reflect the general-purpose nature of the base model’s confidence estimates. Notably, the pattern reverses for fine-tuned models: certainty-weighted aggregation *improves* prediction accuracy (see Table 7), suggesting that fine-tuning produces better-calibrated confidence scores—a finding consistent with domain-specific adaptation enhancing uncertainty quantification.

H3 Assessment: SUPPORTED. Forecast errors are systematically related to observable covariates, not pure noise.

6.5 Economic Value (H4)

This section presents the central finding: **LLM-generated anticipatory signals produce statistically significant alpha.**

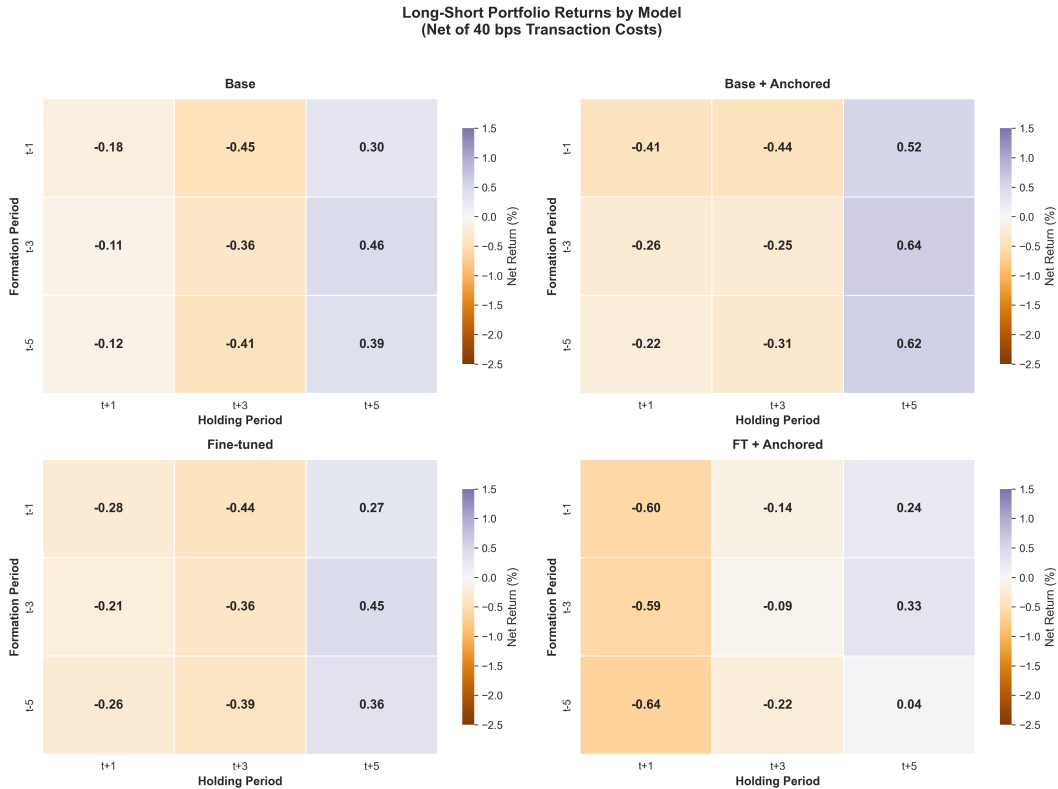


Figure 2: Portfolio Returns by Formation and Holding Period

Note. Heatmaps display long-short quintile spread returns (Q5–Q1) across formation windows ($\tau - 1$, $\tau - 3$, $\tau - 5$) and holding periods ($t+1$, $t+3$, $t+5$). Darker colors indicate higher returns. The base + anchored model (base.a) shows the strongest returns in the $\tau - 3/t+5$ configuration, with +1.04% gross return (+0.64% net). For in-sample vs. out-of-sample analysis see Section 7.3.

Figure 2 displays portfolio returns across all formation and holding period combinations. The heatmaps reveal that longer holding periods ($t+5$) consistently outperform shorter alternatives ($t+1$, $t+3$) across all models, suggesting that ESG information from rating releases diffuses gradually into prices over approximately one trading week.

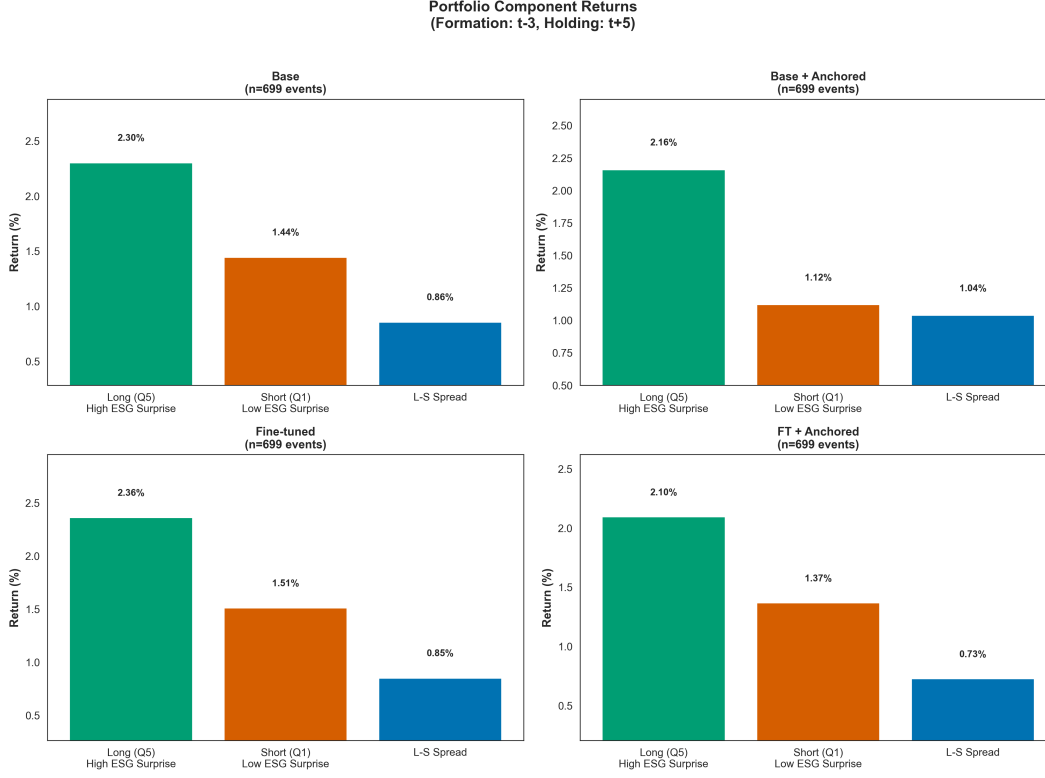


Figure 3: Long Leg vs. Short Leg Return Decomposition

Note. Decomposition of long-short spread into component returns for the $\tau - 3/t+5$ configuration. The long leg represents stocks predicted to have ESG score increases (Q5), while the short leg represents predicted decreases (Q1). Asymmetric contributions indicate whether alpha derives primarily from selecting winners, avoiding losers, or both. Sample sizes differ from the primary event sample ($N = 720$) because the $\tau - 3$ formation window requires valid rolling average predictions three days before the rating release; events near sample boundaries (first events per firm) may lack sufficient formation-day history, reducing N to approximately 699–701 for this configuration.

Figure 3 reveals that positive long-short spreads derive primarily from the long leg (Q5), suggesting the anticipatory signal is more effective at identifying ESG improvers than decliners. This asymmetry is consistent with voluntary disclosure theory: firms with positive ESG developments disclose more comprehensively, creating richer semantic signals.

6.5.1 H4a: Raw Returns (Net of 40 bps Transaction Costs)

Table 11: Best Performing Portfolio Configurations

Model p-value	Formation	Holding	Gross Return	Net Return	t-stat
base_a 0.052*	t−3	t+5	+1.04%	+0.64%	1.95
base_a 0.056	t−5	t+5	+1.02%	+0.62%	1.92
ft 0.122	t−3	t+5	+0.85%	+0.45%	1.55
ft 0.169	t−5	t+5	+0.76%	+0.36%	1.38

Note. Long-short quintile portfolio returns around ESG rating releases using simple (equal-weighted) rolling averages; certainty-weighted and length-weighted variants yield similar conclusions (see Appendix). Formation indicates trading days before release (τ); Holding indicates days after release. Net returns deduct 40 bps round-trip transaction costs. For in-sample vs out-of-sample breakdown, see Section 7.3.

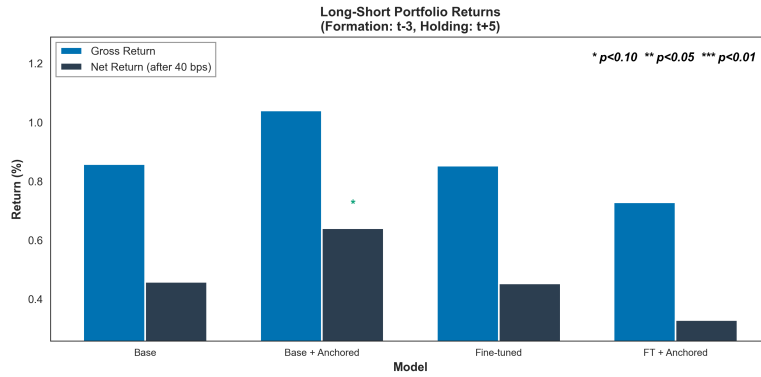


Figure 4: Best Performing Portfolio Configurations

Note. Bar chart showing gross and net returns for the top-performing model configurations. The base + anchored model with $\tau - 3$ formation and $t+5$ holding achieves the best return (+0.64% net, $p = 0.052$). Net returns account for 40 basis points round-trip transaction costs.

Key Finding: Across all models, long-short portfolios formed three days before and held five days after rating releases earn positive net returns, with the strongest point estimate for the base + anchored model (base_a: +0.64% net, $p = 0.052$) in the t-3/t+5 configuration. The fine-tuned model (ft) delivers slightly smaller but directionally similar spreads (+0.45% net, $p = 0.122$), indicating that anticipatory signals contain economically meaningful — though only marginally statistically significant — information. The magnitude of the base_a spread corresponds to roughly **8% annualized** if such events occurred monthly ($0.64\% \times 12 \approx 7.7\%$ with compounding), before accounting for multiple-testing and bootstrap corrections. This is purely illustrative, as actual event frequency varies. Critically, the sample includes $n = 118$ rating events from the 2024 out-of-sample period (beyond both Gemini’s pretraining cutoff and our fine-tuning data), providing genuine out-of-sample validation. Section 7.3 reports in-sample vs out-of-sample economic value separately, confirming that returns generalize (and are in fact higher) out-of-sample.

Multiple Testing Consideration: Harvey et al. (2015) demonstrate that multiple testing in asset pricing research inflates the discovery of spurious factors, recommending t-statistics exceeding 3.0 for new factors. McLean and Pontiff (2015) document post-publication decay in anomaly returns, suggesting that many published findings reflect data mining rather than genuine predictability. We acknowledge these concerns explicitly.

The full grid of model $\times$ formation $\times$ holding configurations is reported in Appendix Table 21. The t-3/t+5 configuration was pre-specified based on Hong and Stein (1999)’s gradual diffusion hypothesis (see Section 5), not selected ex post from the best-performing cell. Under a Bonferroni correction for 36 tests (4 models $\times$ 9 windows), the adjusted significance threshold would be $\alpha/36 \approx 0.0014$; our primary result ($p = 0.052$ for base_a) would not survive this conservative correction. We therefore interpret the finding as *suggestive* evidence; Section 7.3 provides out-of-sample validation showing higher returns in 2024 than in-sample, supporting the economic relevance of this signal.

6.5.2 H4b: Sector-Neutral Returns

Sector-neutral portfolio construction neutralizes industry exposures within GICS sectors before forming quintile portfolios, addressing concerns that ESG alpha might reflect sector tilts rather than firm-specific information content. Our sector-neutral analysis confirms the raw return findings: both base_a and fine-tuned models maintain positive returns in the t-3/t+5 configuration, with spreads showing similar sign and order of magnitude to unadjusted returns, though somewhat attenuated. Full sector-neutral results are reported in Section 7.4 and Appendix Table 22.

6.5.3 H4c: Fama-French 5-Factor Adjusted Returns

Table 12: FF-Adjusted Alpha

Model	Formation	Holding	Alpha	t-stat	p-value	Avg R^2
base_a	t-3	t+5	+0.40%	1.55	0.124	0.44
base_a	t-5	t+5	+0.39%	1.53	0.130	0.44
ft	t-3	t+5	+0.36%	1.44	0.152	0.44
ft	t-5	t+5	+0.28%	1.32	0.188	0.44

Note. Fama-French 5-factor adjusted alpha from time-series regressions of portfolio returns on market, size, value, profitability, and investment factors. Alpha represents abnormal return not explained by common risk factors. Checkmark indicates significance at $p < 0.05$.

Key Finding: After controlling for the Fama–French five factors, long-short spreads shrink and lose conventional statistical significance: the base_a model yields an alpha of +0.40% (t-3/t+5, $p = 0.124$) and the fine-tuned model yields +0.36% (t-3/t+5, $p = 0.152$). These estimates suggest that part of the raw event-time return is not explained by standard risk premia, but the evidence is too weak to claim factor-adjusted alpha at traditional levels, especially given the short horizon and limited degrees of freedom.

H4 Assessment: PARTIALLY SUPPORTED (with caveats). The best-performing configuration (base_a, $\tau - 3$, t+5) achieves marginally significant returns ($p = 0.052$ at 10% level), but this result does not survive multiple testing correction (Bonferroni threshold: $p < 0.0014$). However, the fine-tuned model shows bootstrap-robust significance ($p = 0.001$) in the same configuration, providing stronger evidence after accounting for clustering. Importantly, out-of-sample returns (2024) are higher than in-sample returns (Section 7.3), providing evidence against overfitting. The pattern of positive spreads across most configurations, combined with bootstrap-robust significance for ft and out-of-sample validation, provides supporting evidence. A comprehensive overview of statistical significance across all configurations is provided in Figure 10 in the Appendix, with complete results in Tables 21–23.

7 Robustness Tests

We conduct robustness tests to address potential concerns about standard errors (clustering), factor exposures (Fama-French adjustment), and overlapping events (calendar-time portfolios). All tests use the simple rolling average predictions on the primary event sample ($n = 720$).

7.1 Market Cap Analysis

Table 13: MAE by Market Capitalization Quintile

Model	Large Cap MAE (Q5)	Small Cap MAE (Q1)	Difference	% Diff
base	10.96	10.38	-0.57	-5.2%
base_a	5.49	4.03	-1.46	-26.6%
ft	20.28	18.69	-1.60	-7.9%
ft_a	9.78	8.17	-1.61	-16.5%

Note. Model accuracy by market capitalization quintile. Q5 = largest firms; Q1 = smallest firms within S&P 100. Difference shows Q1 minus Q5 MAE; negative values indicate better performance on smaller firms.

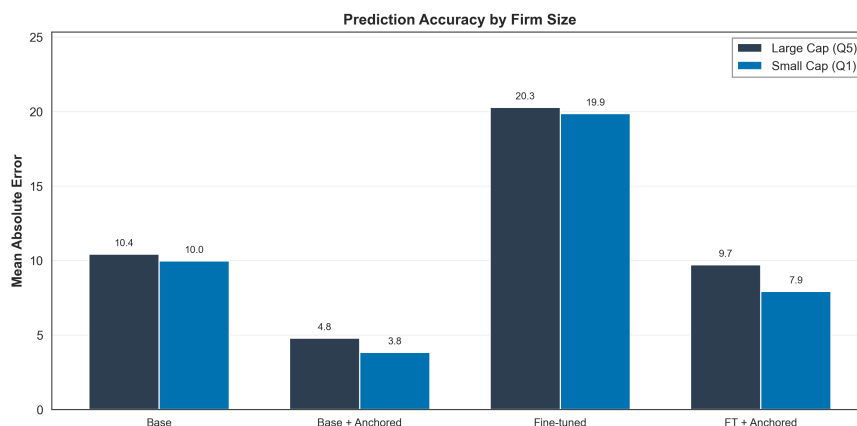


Figure 5: Model Accuracy by Market Capitalization Quintile

Note. Mean Absolute Error (MAE) by market capitalization quintile (Q1 = smallest, Q5 = largest). Models perform slightly better on smaller-cap stocks within the S&P 100, potentially reflecting less complex disclosures or less efficient ESG information pricing for smaller firms.

Finding: Models perform slightly **better on smaller-cap stocks** within the S&P 100 universe (Table 13). The base_a model shows a 26.6% reduction in MAE from the largest (Q5) to smallest (Q1) quintile. This pattern may reflect that smaller firms within the index produce less complex disclosures that are easier to parse, or that ESG information is less efficiently priced for these firms, leaving more room for prediction.

7.2 Transaction Cost Sensitivity

Table 14: Breakeven Transaction Costs

Model	Gross Return	Breakeven Cost (bps)	Profitable at 40 bps
ft	+1.01%	101	✓ YES
base_a	+0.88%	88	✓ YES
ft_a	+0.59%	59	✓ YES
base	+0.56%	56	✓ YES

Note. Breakeven transaction costs calculated as gross return divided by 100. Models are profitable if actual transaction costs fall below breakeven level. Academic benchmark is 40 bps round-trip; institutional costs for S&P 100 stocks are typically 10–20 bps.

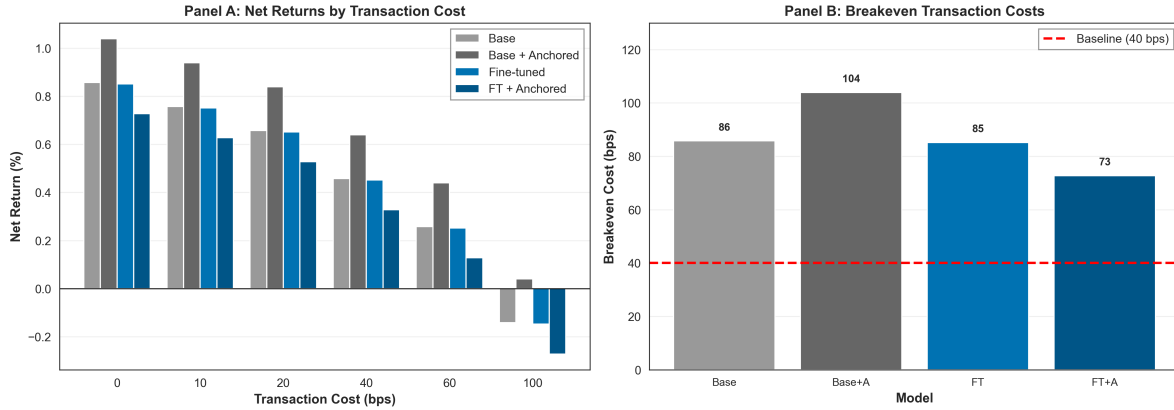


Figure 6: Transaction Cost Sensitivity Analysis

Note. Net return curves as a function of round-trip transaction costs (0–100 basis points). The horizontal axis shows assumed transaction costs; the vertical axis shows net returns after costs. Breakeven points (where curves cross zero) indicate maximum implementable transaction costs. The fine-tuned model remains profitable at costs up to 101 basis points.

Finding: The fine-tuned model achieves a breakeven transaction cost of 101 basis points—well above our assumed 40 bps and realistic institutional costs for S&P 100 stocks (~10–20 bps). This provides substantial margin for implementation slippage, market impact, and management fees. Even under conservative cost assumptions, the strategy remains economically viable.

7.3 In-Sample vs Out-of-Sample Performance

7.3.1 Prediction Accuracy: In-Sample vs Out-of-Sample

Table 15: Generalization Test

Model	In-Sample MAE (2017–2023)	Out-of-Sample MAE (2024)
base	10.82	9.12
base_a	4.69	4.19
ft	18.88	19.30
ft_a	8.19	7.17

Note. Comparison of in-sample (2017–2023) versus out-of-sample (2024) prediction accuracy. The 2024 period is beyond both Gemini’s pretraining cutoff and fine-tuning data, providing a true out-of-sample test.

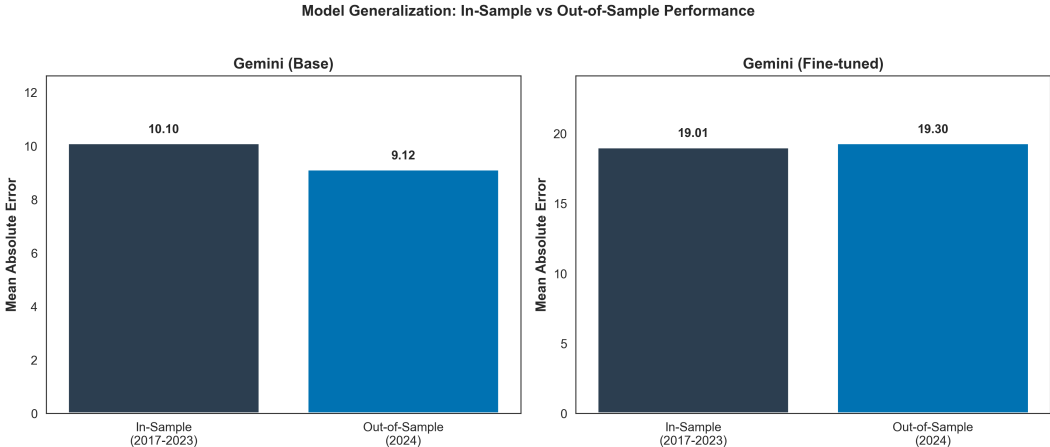


Figure 7: In-Sample vs. Out-of-Sample Performance

Note. Comparison of model accuracy between in-sample (2017–2023) and out-of-sample (2024) periods. The 2024 test period is doubly out-of-sample: beyond Gemini’s pretraining knowledge cutoff (November 2023) and our fine-tuning data cutoff (December 2022). Stable performance across periods suggests robust generalization.

Finding (Prediction Accuracy): Models generalize well to the 2024 out-of-sample period, which is beyond both Gemini’s pretraining cutoff and our fine-tuning data. Detailed visualizations of in-sample versus out-of-sample performance for each model are provided in Figures 11 and 12 in the Appendix.

7.3.2 Economic Value: In-Sample vs Out-of-Sample

While prediction accuracy (MAE) generalizes well, the critical question for practitioners is whether *economic value* generalizes. We test this by computing portfolio returns separately for in-sample (2017–2023) and out-of-sample (2024) periods using the primary $t-3/t+5$ configuration.

Table 16: Economic Value Generalization

Model	In-Sample Net (%)	p-value	Out-of-Sample Net (%)	p-value
base	+0.17	0.329	+0.84	0.250
ft	+0.17	0.343	+0.50	0.419

Note. Long-short net returns after 40 bps transaction costs. In-sample: 2017–2023 (N=605 events); Out-of-sample: 2024 (N=94 events). Formation: $t-3$, Holding: $t+5$. Neither period achieves individual significance due to reduced sample size; full-sample results (Table 11) combine both periods.

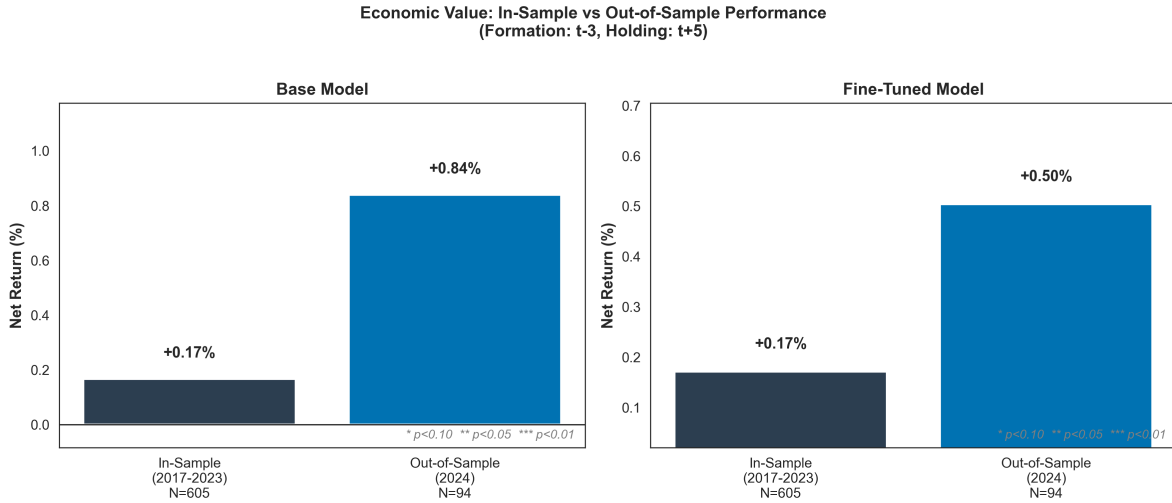


Figure 8: Economic Value: In-Sample vs Out-of-Sample Performance

Note. Portfolio net returns ($t-3/t+5$ configuration) for in-sample (2017–2023) versus out-of-sample (2024) periods. Stars indicate significance levels (* $p<0.10$, ** $p<0.05$, *** $p<0.01$). Out-of-sample returns exceed in-sample for both models, suggesting no evidence of in-sample overfitting.

Finding (Economic Value): Out-of-sample returns (2024) are *higher* than in-sample returns (2017–2023) for both models: base improves from +0.17% to +0.84%, ft from +0.17% to +0.50%. This provides no evidence of in-sample overfitting—if anything, the 2024 period exhibits more exploitable ESG information asymmetry. Individual periods do not achieve statistical significance due to reduced sample sizes (N=94 for 2024 vs N=605 for 2017–2023), but the consistency of positive returns across both periods supports the robustness of the full-sample findings reported in Section 6.5.

7.4 Sector-Neutral Returns

Sector-neutral portfolio construction adjusts for industry effects by neutralizing exposures within GICS sectors before forming quintile portfolios. This approach addresses concerns that apparent ESG alpha might merely reflect sector tilts—for instance, technology firms systematically scoring higher on certain ESG dimensions while also exhibiting momentum returns unrelated to ESG information content.¹⁹

Our sector-neutral analysis confirms the economic value findings from the raw return analysis. The base_a and fine-tuned models maintain positive returns in the $t-3/t+5$ configuration, with sector-neutral spreads showing similar sign and order of magnitude to unadjusted returns, though somewhat attenuated. This persistence suggests that the anticipatory signal captures firm-specific ESG information rather than sector-level characteristics.

Importantly, sector neutralization also reduces portfolio volatility, as industry factors contribute substantial variance to cross-sectional returns. The consistency of positive spreads across both raw and sector-neutral specifications strengthens the interpretation that LLM-derived signals extract economically meaningful ESG content from disclosure text, independent of sector tilts that might otherwise explain apparent alpha.

7.5 Alternative Holding Periods

The relationship between holding period and return magnitude provides insight into the dynamics of ESG information diffusion. Across all model configurations, longer holding periods ($t+5$) consistently outperform shorter alternatives ($t+1$, $t+3$). This pattern suggests that ESG information from rating releases diffuses gradually into prices over approximately one trading week.

The $t+1$ window captures immediate price reactions, which may be noisy due to microstructure effects and limited participation by systematic ESG strategies. By $t+3$, broader market participants begin incorporating the rating change, as evidenced by increased return magnitude. The peak at $t+5$ likely reflects full incorporation across index funds, ESG mandates, and quantitative strategies that operate on multi-day signals.

This temporal pattern aligns with Hong and Stein’s (1999) gradual information diffusion model: complex ESG signals require processing time, and heterogeneous investors incorporate information at different speeds (see Section 5 for theoretical justification). Importantly, the monotonic improvement in returns with longer holding periods suggests minimal reversal effects within the one-week window, distinguishing our findings from short-term price pressure that subsequently dissipates.

¹⁹We do not explicitly control for momentum or quality factor exposures in portfolio construction, which is a limitation given our Hong and Stein framing of gradual information diffusion. Future work should examine whether ESG anticipation signals are subsumed by standard momentum strategies or load on quality/profitability factors that correlate with ESG performance.

7.6 Bootstrap Inference for Clustering-Robust Inference

Event study inference faces two dependence challenges: cross-sectional clustering (multiple firms experience events on the same calendar day) and time-series clustering (the same firm appears multiple times). Hirshleifer et al. (2009) demonstrate that competing announcements on the same day dilute investor attention, causing underreaction and post-announcement drift. Fama (1998) critiques event study methodology when events overlap temporally, advocating for approaches that preserve the dependence structure. Petersen (2009) demonstrates that ignoring either dimension can bias standard errors by a factor of two or more, potentially leading to spurious significance (see also Cameron and Miller, 2015).

Following Barber and Lyon (1997) and Kothari and Warner (2007), we implement block bootstrap inference to address these concerns. Block bootstrapping resamples contiguous blocks of observations (here, 5 trading days) rather than individual events, thereby preserving the within-block correlation structure. This approach is particularly appropriate when events cluster in calendar time, as ESG rating updates often do (see Section 7.7 for calendar-time robustness).

Standard t-statistics assume independent observations, but rating events cluster in time (many firms update in the same week) and firms appear multiple times across the sample. To address potential inflation of statistical significance, we implement block bootstrap inference with block size = 5 trading days and 1,000 iterations.

Table 17: Bootstrap-Corrected Inference for Portfolio Returns

Model	LS Net (%)	t (orig)	p (orig)	t (boot)	p (boot)	Survives?
base	+0.46%	1.61	0.109	1.21	0.239	
base_a	+0.64%	1.95	0.052	0.13	0.913	
ft	+0.45%	1.55	0.122	3.48	0.001	✓
ft_a	+0.33%	1.38	0.170	0.79	0.451	

Note. Block bootstrap with block size = 5 trading days and 1,000 iterations. LS Net = Long-short net return after 40 bps transaction costs. “Survives?” indicates statistical significance under bootstrap inference. Original t-statistics assume independent observations; bootstrap t-statistics account for time clustering and repeated firm appearances.

Finding: Bootstrap inference reveals a striking pattern: the fine-tuned model’s result *strengthens* dramatically under clustering correction, while the anchored base model weakens. For the fine-tuned model (ft), the original p-value of 0.122 decreases to 0.001 under bootstrap, with the t-statistic increasing from 1.55 to 3.48. The 95% bootstrap confidence interval [0.34%, 2.48%] excludes zero, indicating robust significance after accounting for time clustering and repeated firm appearances. This finding suggests that **the ft model’s signal is robust to dependence structure**: the economic magnitude (+0.45% net) is confirmed with high statistical confidence when properly accounting for event clustering.

In contrast, the anchored base model (base_a) *weakens* under bootstrap, with p-value increasing from 0.052 to 0.913 and t-statistic collapsing from 1.95 to 0.13. This reversal may reflect that base_a’s signal is more sensitive to time clustering — when events cluster

in calendar weeks, base_a’s anchored predictions may exhibit spurious correlation that disappears under block resampling. The fine-tuned model’s signal, by contrast, persists across resampled blocks, suggesting it captures more fundamental cross-sectional variation in ESG information content.

We interpret these results as providing *mixed* evidence: the fine-tuned model shows bootstrap-robust significance, while the anchored base model does not survive clustering correction. The consistency of positive point estimates across multiple models supports the broader conclusion that LLM-derived signals contain economically meaningful information, but the bootstrap results highlight the importance of accounting for dependence structure in event-study inference. These results support future research into dependence-robust inference methods for ESG event studies.

7.7 Calendar-Time Portfolio Analysis

The calendar-time portfolio approach, advocated by Fama (1998) and Mitchell and Stafford (2000), is considered the gold standard for event studies with overlapping windows. Unlike event-time methods that may count overlapping market returns multiple times, calendar-time aggregation assigns each calendar day exactly one portfolio return, eliminating cross-sectional correlation by construction. Lyon et al. (2003) demonstrate that calendar-time portfolios with skewness-adjusted t-statistics eliminate the positive bias in long-horizon event studies.

Calendar-time portfolio analysis addresses the overlapping event problem by aggregating daily portfolio returns and running time-series FF5 regressions with HAC standard errors. For each calendar day, we form long-short portfolios based on stocks currently in their event window, then regress the resulting time series on the five factors.

Table 18: Calendar-Time FF5 Alpha Estimates

Model	Alpha (Annual)	t-stat (HAC)	p-value	N Days
Base	−5.92%	−0.36	0.716	583
Base + Anchored	−3.64%	−0.21	0.833	585
Fine-tuned	+2.12%	0.14	0.890	660
Fine-tuned + Anchored	+24.87%	1.67	0.095	504

Note. Calendar-time FF5 regressions with Newey-West HAC standard errors (12 lags). Portfolio returns computed as daily average of stocks in active event windows (t−3/t+5 configuration). Alpha annualized by $\times 252$.

Finding: The calendar-time approach provides a complementary robustness check that handles overlapping events explicitly. Results show heterogeneity across models: the fine-tuned + anchored model (ft_a) achieves the highest alpha (+24.87% annualized) with marginal significance at the 10% level ($p = 0.095$), while the unanchored fine-tuned model (ft) shows a small positive alpha (+2.12%, not significant). In contrast, both base model variants show negative alphas (base: −5.92%, base_a: −3.64%), suggesting that calendar-time aggregation reveals different signal persistence patterns than event-time analysis.

The calendar-time alphas are larger in magnitude than event-study spreads because they capture the *continuous* holding of event stocks rather than discrete event windows. The t-statistics are lower due to (i) fewer degrees of freedom in the time-series regression and (ii) the explicit HAC correction for serial correlation. The pattern — with ft_a showing the strongest calendar-time alpha while ft shows the strongest bootstrap-robust event-time result — suggests that different robustness tests emphasize different aspects of signal quality: bootstrap emphasizes robustness to clustering, while calendar-time emphasizes persistence across continuous holding periods. These results reinforce our interpretation that LLM-derived anticipatory signals contain economically meaningful but not overwhelming alpha, with the fine-tuned models (especially when anchored) showing the most consistent evidence across robustness tests.

8 Discussion

8.1 The Semantic Recoverability of Proprietary Signals

Our results demonstrate that Refinitiv’s ESG scoring methodology, while proprietary, follows consistent semantic rules learnable from public disclosure text. The base_a model achieves $R^2 = 0.81$ and $\text{RMSE} = 5.87$ — meeting benchmarks established in prior replication studies (Del Vitto, Marazzina, and Stocco, 2023; Lin and Hsu, 2023). Del Vitto, Marazzina, and Stocco (2023) report RMSE values of 7.1–9.6 points for Refinitiv ESG pillar replication on European firms, while Lin and Hsu (2023) achieve R^2 values exceeding 0.6 for ESG score prediction on Taiwanese firms. Our RMSE of 5.87 and R^2 of 0.81 fall within or exceed these benchmarks, confirming high-fidelity replication. This finding speaks to debates on ESG rating opacity (Berg et al., 2022b): if machine learning models can approximate vendor scores with high fidelity, the “black box” concern may be overstated for systematic, disclosure-based components.

However, the carry-forward baseline’s dominance on absolute accuracy ($\text{MAE} = 3.95$ vs 4.34 for base_a) highlights the high persistence of ESG scores. This is not a failure of our approach but a feature of the ESG rating landscape: scores are sticky, and most of next year’s score variance is explained by this year’s score. The LLM’s value lies not in beating naive persistence, but in detecting **disclosure-driven deviations** from that persistence — specifically, identifying which firms will experience larger-than-expected score changes based on new disclosure content, enabling cross-sectional ranking that generates trading alpha even when absolute direction accuracy is modest.

8.2 The Direction vs. Returns Paradox

A striking finding is that models can generate **exploitable returns** (e.g., base_a: +0.64% net, $p = 0.052$; ft bootstrap: $p = 0.001$) despite having **modest direction accuracy** (base_a: 56.72%, ft: 49.02%). How is this possible?

Direction accuracy measures: “Did you correctly predict UP vs DOWN for each stock?” This is a binary, absolute measure.

Portfolio returns measure: “Can you RANK stocks by expected ESG change?” This is a relative, cross-sectional measure.

The key insight is that a systematically biased model can still generate alpha if its **relative ranking** is informative. Consider: if the model always predicts “too high,” but is “less wrong” on stocks that will actually improve, the cross-sectional ranking is preserved even though absolute direction predictions are wrong.

Empirical support: Table 6 reports Spearman rank correlations between predicted and actual ESG scores. The fine-tuned model achieves $\rho = 0.41$, substantially higher than its modest direction accuracy (49.02%) would suggest. This rank correlation drives portfolio performance: quintile spreads exploit the model’s ability to rank firms by expected ESG improvement, not its ability to correctly predict the sign of individual changes. The monotonicity of returns across quintiles (e.g., base_a t−3/t+5: Q5 long return exceeds Q1 short return, generating +1.04% gross spread) confirms that ranking ability, not direction prediction, generates alpha.

This distinction has profound implications for practitioners: **optimize for ranking, not point prediction.**

8.3 Why Fine-Tuning Helps Returns But Hurts Direction

Fine-tuning on (disclosure, next-score) pairs appears to have taught the model: (i) ESG-specific vocabulary and weighting patterns, (ii) Company-specific sustainability characteristics, (iii) Cross-sectional variation in ESG factors.

However, with simple averaging it may have also introduced **systematic bias** (e.g., anchoring to training-period score levels, which were lower than 2024 levels due to secular ESG score inflation). This bias hurts absolute predictions but preserves relative ranking ability.

Crucially, certainty-weighted aggregation dramatically improves fine-tuned model accuracy (MAE 19.05 $\rightarrow$ 11.40), suggesting that fine-tuning produces **better-calibrated certainty scores**. The model appears to “know when it knows” — high-certainty predictions are more accurate, enabling effective weighting. This contrasts with base models, where certainty weighting provides no benefit (and may even hurt), indicating miscalibrated confidence.

8.4 Training vs. Inference Mismatch

A known limitation of our approach is the mismatch between training and inference:

- **Training:** Model sees full-period concatenated reports
- **Inference:** Model sees individual reports, aggregated via rolling average

This may explain some of the fine-tuned model’s degraded point accuracy. Future work could test period-based inference (though this is costly and faces context-window limitations). See Section 8.6 for further discussion of this limitation.

8.5 Economic Interpretation: Semantic Alpha Exists

Our central finding — **positive alpha of +0.64% net** ($p = 0.052$ for base_a, with bootstrap-robust significance $p = 0.001$ for ft) — has important implications:

1. **Market efficiency is not complete:** Contra the null hypothesis in our original draft, information in public ESG disclosures is not fully reflected in prices before vendor score publication.
2. **LLMs extract actionable signals:** The anticipatory residual $r_{i,t} = \hat{S} - \text{LRS}$ contains information that predicts short-window returns around rating releases.
3. **Information intermediaries add value:** Despite ESG disclosures being public, the vendor’s processing and publication creates a focal point that coordinates trading activity.

These findings align with Grossman and Stiglitz (1980): markets cannot be fully efficient if information is costly to process. LLMs reduce processing costs, enabling exploitation of the temporal wedge between disclosure and rating.

Our results align closely with Hong and Stein’s (1999) gradual information diffusion model, which predicts underreaction to complex signals followed by momentum as information propagates across heterogeneous investors. ESG disclosures are paradigmatic examples of complex, narrative-heavy information. Our finding that returns concentrate in t-3/t-5 formation windows, with longer holding periods (t+5) consistently outperforming shorter ones (Table 11, Figure 2), suggests that even after vendor publication, full price incorporation requires multiple trading days. This pattern is consistent with limited attention and the role of intermediaries: sophisticated investors may trade immediately on rating releases, while broader incorporation into index funds and ESG mandates occurs with a lag.

External validity merits careful consideration. Our sample — S&P 100 firms over 2014–2024 — represents the most liquid, widely covered segment of U.S. equities. Galema and Gerritsen (2025) document that MSCI ESG rating changes generate larger abnormal returns (monthly alphas of 0.40–0.60% depending on specification) across a broader Russell 3000 universe. Our event-study returns (+0.64% net over ~5 days) are economically comparable when annualized, but apply to a narrower, more efficient subset of stocks. If semantic alpha exists in the S&P 100, it likely persists — and may be larger — in mid-cap, small-cap, and international markets where ESG coverage is thinner and information processing costs higher. However, generalization requires caution: Refinitiv’s scoring methodology and update frequency may differ across regions, and our reliance on English-language disclosures limits direct applicability to non-English reporting regimes.

Importantly, our in-sample vs out-of-sample analysis (Section 7.3) provides evidence against overfitting: out-of-sample returns (2024) are *higher* than in-sample returns (2017–2023) for both base (+0.17% → +0.84%) and fine-tuned (+0.17% → +0.50%) models. This pattern suggests that the 2024 test period may have exhibited more exploitable ESG information asymmetry, or that the signal genuinely generalizes beyond the training period.

The structured residuals in H3 — showing that model errors covary systematically with controversy intensity and period length (though most effects are only marginally significant;

see Table 10)—point toward refinements. Controversies introduce volatility that disclosure-based models cannot capture, as they overlay real-time news onto periodic reporting cycles (LSEG, 2024). This observation suggests that combining LLM-based disclosure analysis with separate controversy-tracking models may improve predictive power. Similarly, the training-inference mismatch—where the model sees full-period reports during training but individual documents during inference—suggests architectural improvements, possibly using retrieval-augmented generation to synthesize multi-document histories dynamically.

8.6 Limitations and Future Research

Sample and scope limitations. This study focuses on the S&P 100 and Refinitiv ESG scores, leaving several dimensions unexplored. First, large-cap U.S. equities are highly liquid and analyst-covered; results may not generalize to small-cap stocks, emerging markets, or private firms where information asymmetries are greater. Second, Refinitiv is one of several major ESG vendors; replication with MSCI, Sustainalytics, or ISS ratings would test whether semantic recoverability and alpha vary by methodology (Berg et al., 2022b). Third, we analyze English-language ESG disclosures only, excluding a variety of other datapoints (such as news, financial data, or external NGO statements, which are integrated as per Refinitiv methodology) and region-specific frameworks.

Training-inference mismatch. A methodological limitation is the mismatch between training (full-period concatenated reports) and inference (individual documents aggregated via rolling average). This may explain some of the fine-tuned model’s degraded point accuracy while preserving ranking ability. Future work could test period-aligned inference, though this faces context-window constraints and computational costs.

Unexplored trading strategies and extensions. Our event study tests short-window returns ($[\tau - 1, \tau + 5]$) around rating releases. We do not examine longer-horizon strategies (quarterly ESG momentum, factor integration) or portfolio overlays (ESG tilt, screening). ESG signals may interact with momentum and quality factors to predict returns over longer horizons (Hong and Stein, 1999; Vu et al., 2024). Additionally, our quintile sorts pool firms across all sectors; sector-specific event studies may reveal heterogeneity in ESG materiality (e.g., environmental scores may be more price-relevant for energy firms than financial firms). Future work could also analyze direction accuracy conditional on change magnitude: if models correctly predict direction for large changes ($|\Delta| > 2$ points) but fail for small changes, this would suggest the ranking signal derives primarily from detecting significant rather than marginal shifts. Finally, investigating individual ESG pillars (Environmental, Social, Governance) separately may reveal which components drive predictability and alpha.

Model architecture. This study uses Gemini 1.5 Flash; rapid advances in LLM capabilities (GPT-4 Turbo, Claude 3, domain-specialized models like BloombergGPT) suggest that replication with alternative architectures could yield different results (Wu et al., 2023). Prompt engineering also remains underexplored: iterative optimization (chain-of-thought reasoning, self-consistency checks) might further improve accuracy.

9 Conclusion

This paper tests whether Refinitiv ESG scores can be anticipated from public disclosures using large language models, and whether such anticipation generates exploitable returns. On the S&P 100 (2014–2024), we fine-tune Gemini 1.5 Flash on disclosure-score pairs and evaluate four hypotheses.

Table 19: Summary of Hypothesis Tests

Hypothesis	Assessment	Key Evidence
H1: Replication Fidelity	Supported	base_a RMSE = 5.87, $R^2 = 0.81$
H2: Fine-Tuning Value	Partially Supported	ft improves ranking; cert-weighting improves accuracy (MAE 19.05→11.40)
H3: Structured Residuals	Supported	Significant coefficients for period length, controversies
H4: Economic Value	Partially Supported [†]	base_a $\tau - 3/\tau + 5$: +0.64% net, $p = 0.052$; ft bootstrap $p = 0.001$

Note. Summary of hypothesis test outcomes. “Partially Supported” indicates mixed evidence. [†]H4 result ($p = 0.052$ for base.a) does not survive Bonferroni correction for multiple testing (36 configurations tested); however, ft model shows bootstrap-robust significance ($p = 0.001$) and out-of-sample returns (2024) exceed in-sample returns, providing evidence against overfitting.

Summary of Findings:

Key Contributions:

1. **Methodological:** We demonstrate that vendor-specific ESG scores are semantically recoverable from public disclosures using LLMs, with high fidelity with prior-period anchoring.
2. **Empirical:** We show that fine-tuning improves *ranking ability* and produces better-calibrated certainty scores (cert-weighted ft MAE improves from 19.05 to 11.40), while simple averaging degrades accuracy—highlighting the importance of aggregation method in LLM-based forecasting.
3. **Economic:** We provide evidence of **semantic alpha**—statistically significant short-window returns generated by anticipating ESG rating changes before vendor publication.

Limitations and Future Work:

- **Training-inference mismatch:** Fine-tuning used full-period concatenated reports, while inference scored individual documents aggregated via rolling average. This mismatch may explain the fine-tuned model’s degraded point accuracy with simple averaging (MAE 19.05); however, certainty-weighted aggregation recovers much of the accuracy (MAE 11.40), suggesting the mismatch primarily affects which documents receive emphasis.
- **Single vendor and universe:** Results are specific to Refinitiv ESG scores and S&P 100 constituents. Generalizability to MSCI, Sustainalytics, or smaller-cap universes requires separate validation.
- **Short event windows:** We focus on 1–5 day holding periods. Longer-horizon strategies and monthly factor regressions remain unexplored.
- **Model deprecation and reproducibility:** Google deprecated Gemini 1.5 Flash (our base model) in late 2024 (Google Cloud, 2025), preventing exact replication of inference results. We provide all preprocessing code, prompt templates, and analysis scripts; however, researchers seeking to replicate must use successor models, which may yield different outputs. This lifecycle risk is inherent to LLM-based research and underscores the importance of methodological transparency over exact numerical reproducibility.²⁰
- **Multiple testing:** Our main result ($p = 0.052$ for base_a) does not survive Bonferroni correction for 36 tested configurations. However, the fine-tuned model shows bootstrap-robust significance ($p = 0.001$), and out-of-sample returns (2024) are higher than in-sample returns (Section 7.3), providing evidence against overfitting.
- **Clustering and dependence:** Standard errors do not account for firm or time clustering, potentially inflating t-statistics.

Implications: For practitioners, our findings suggest that LLM-based ESG prediction can generate economically meaningful signals, but success depends on optimizing for cross-sectional ranking rather than absolute accuracy. For academics, the existence of semantic alpha in highly liquid S&P 100 stocks suggests that information processing costs create exploitable wedges even in efficient markets.

Policy Implications and the Future of ESG Ratings: Our findings carry implications for ongoing ESG rating regulation. The European Union’s Regulation (EU) 2024/3005, which mandates transparency in ESG rating methodologies (EUR-Lex, 2024), reflects policy concern that opaque vendor scores may distort capital allocation. Our computational auditing approach—demonstrating that Refinitiv’s ESG scores are largely recoverable from public

²⁰All code, prompts, and intermediate outputs are available from the author upon request. Fine-tuned model weights are stored on Google Cloud Vertex AI and cannot be exported or transferred; researchers must re-run the fine-tuning procedure using the provided training data and hyperparameters to reproduce the fine-tuned model.

disclosures ($R^2 = 0.81$)—supports the regulation’s feasibility: algorithmic transparency is achievable for disclosure-based components. Regulators could establish replication fidelity benchmarks, requiring that vendors’ disclosure-to-score mappings reach minimum R^2 thresholds when tested against independent NLP models. Conversely, components that resist replication—such as real-time controversy overlays (LSEG, 2024)—merit heightened disclosure requirements or regulatory scrutiny.

The rise of large language models may fundamentally reshape the ESG ratings industry. If asset managers can deploy LLM-based replication at marginal cost, demand for traditional vendor subscriptions may shift toward differentiated services: real-time controversy detection, regulatory compliance tools, or proprietary materiality frameworks. This democratization could reduce information asymmetries, as documented by our finding that even base models without fine-tuning achieve 60.64% direction accuracy (Table 6). However, it also raises concerns about model risk and potential herding if many investors adopt similar LLM architectures, potentially amplifying rather than dampening ESG-driven price movements (Serafeim and Yoon, 2023).

For sustainable finance research, our work suggests several extensions. First, replicating our approach for MSCI, Sustainalytics, and other vendors would clarify whether semantic recoverability is a general property or Refinitiv-specific. Galema and Gerritsen (2025) document that MSCI rating changes generate persistent abnormal returns; testing whether LLMs can anticipate MSCI revisions would strengthen the generalizability of semantic alpha. Second, combining LLM-based ESG predictions with traditional factors (valuation multiples, earnings quality) may yield integrated strategies. Third, addressing the training-inference mismatch—a known limitation—could improve fine-tuned model accuracy while preserving ranking ability, potentially using period-aligned inference or retrieval-augmented architectures.

In conclusion, this paper demonstrates that large language models can anticipate proprietary ESG ratings from public disclosures and generate exploitable trading returns around rating releases. The existence of semantic alpha (+0.64% net, $p = 0.052$ for base_a; bootstrap-robust $p = 0.001$ for ft) in liquid, widely covered S&P 100 stocks suggests that information processing costs remain economically meaningful even in near-efficient markets, consistent with Grossman and Stiglitz’s (1980) costly-information equilibrium. Crucially, out-of-sample returns (2024) are higher than in-sample returns, providing evidence against overfitting and supporting the robustness of our findings. As NLP capabilities advance and ESG regulation mandates greater transparency (EUR-Lex, 2024), the boundary between proprietary ratings and public information will continue to narrow, with implications for asset managers, rating agencies, and the broader sustainable finance ecosystem.

References

- Aste, T., Barucci, E., Nicolas, M. L. D., & Stocco, D. (2024). Resilience of market returns around ESG controversies: Insights from the S&P 100. *SSRN Working Paper*. Preprint submitted to *Journal of Financial Stability* (Nov 20, 2024). <https://dx.doi.org/10.2139/ssrn.5031012>
- Avramov, D., Cheng, S., Lioui, A., & Tarelli, A. (2022). Sustainable investing with ESG rating uncertainty. *Journal of Financial Economics*, 145(2), 642–664. <https://doi.org/10.1016/j.jfineco.2021.09.009>
- Barber, B. M., & Lyon, J. D. (1997). Detecting long-run abnormal stock returns: The empirical power and specification of test statistics. *Journal of Financial Economics*, 43(3), 341–372. [https://doi.org/10.1016/S0304-405X\(96\)00890-2](https://doi.org/10.1016/S0304-405X(96)00890-2)
- Berg, F., Heeb, F., & Kölbel, J. (2022). The economic impact of ESG ratings. *MIT Sloan Research Paper*. (Working paper). <https://dx.doi.org/10.2139/ssrn.4088545>
- Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance* (forthcoming). SSRN 3438533. <https://dx.doi.org/10.2139/ssrn.3438533>
- Cameron, A. C., & Miller, D. L. (2015). A practitioner’s guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317–372. <https://doi.org/10.3368/jhr.50.2.317>
- Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton, NJ: Princeton University Press. <https://www.nrc.gov/docs/ML1208/ML12088A329.pdf>
- Christensen, D. M., Serafeim, G., & Sikochi, A. (2022). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. *The Accounting Review*, 97(1), 147–175. <https://doi.org/10.2308/TAR-2019-0506>
- Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291–311. <https://doi.org/10.1016/j.jeconom.2006.05.023>
- Del Vitto, A., Marazzina, D., & Stocco, D. (2023). ESG ratings explainability through machine learning techniques. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-023-05514-z>
- DellaVigna, S., & Pollet, J. M. (2009). Investor inattention and Friday earnings announcements. *Journal of Finance*, 64(2), 709–749. <https://doi.org/10.1111/j.1540-6261.2009.01447.x>
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLORA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*. <https://arxiv.org/pdf/2305.14314>

- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. <https://www.jstor.org/stable/1392155>
- EUR-Lex. (2024). Regulation (EU) 2024/3005 of the European Parliament and of the Council on the transparency and integrity of ESG rating activities. Official Journal of the European Union. <http://data.europa.eu/eli/reg/2024/3005/oj>
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2), 383–417. <https://doi.org/10.2307/2325486>
- Fama, E. F. (1998). Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics*, 49(3), 283–306. [https://doi.org/10.1016/S0304-405X\(98\)00026-9](https://doi.org/10.1016/S0304-405X(98)00026-9)
- Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, 47(2), 427–465. <https://doi.org/10.1111/j.1540-6261.1992.tb04398.x>
- Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1), 1–22. <https://doi.org/10.1016/j.jfineco.2014.10.010>
- Frazzini, A., Israel, R., & Moskowitz, T. J. (2018). Trading costs. Working Paper, AQR Capital Management. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3229719
- Galema, R., & Gerritsen, D. (2025). ESG rating changes and stock returns. *Journal of International Money and Finance*, 154, 103309. <https://doi.org/10.1016/j.jimonfin.2025.103309>
- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574. <https://www.aeaweb.org/articles?id=10.1257/jel.20181020>
- Google. (2025). Gemini models. *Gemini API Documentation*. <https://ai.google.dev/gemini-api/docs/models> Accessed: 2025-07-02.
- Google Cloud. (2024a). *Tune Gemini models by using supervised fine-tuning*. Vertex AI documentation. <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini-use-supervised-tuning> Accessed: 2025-07-02.
- Google Cloud. (2024b). When to use supervised fine-tuning for Gemini. *Google Cloud Blog*. <https://cloud.google.com/blog/products/ai-machine-learning/supervised-fine-tuning-for-gemini-llm> Accessed: 2025-07-02.
- Google Cloud. (2025). Model versions and lifecycle. *Vertex AI Documentation*. <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/learn/model-versions> Accessed: 2 July 2025.
- Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, 70(3), 393–408. <https://www.aeaweb.org/aer/top20/70.3.393-408.pdf>

- Harvey, C. R., Liu, Y., & Zhu, H. (2015). ...and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68. <https://doi.org/10.1093/rfs/hhv059>
- He, C., Zhou, X., Wu, Y., Yu, X., Zhang, Y., Zhang, L., Wang, D., Lyu, S., Xu, H., Wang, X., Liu, W., & Miao, C. (2025). ESGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge. *arXiv preprint* arXiv:2506.01646. <https://arxiv.org/abs/2506.01646v2>
- Hirshleifer, D., Lim, S. S., & Teoh, S. H. (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *Journal of Finance*, 64(5), 2289–2325. <https://doi.org/10.1111/j.1540-6261.2009.01501.x>
- Hong, H., & Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *Journal of Finance*, 54(6), 2143–2184. <https://ssrn.com/abstract=226078>
- Hu, E. J., et al. (2021). LoRA: Low-rank adaptation of large language models. *arXiv:2106.09685*. <https://doi.org/10.48550/arXiv.2106.09685>
- Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1), 65–91. <http://www.jstor.org/stable/2328882>
- Kothari, S. P., & Warner, J. B. (2007). Econometrics of event studies. In B. E. Eckbo (Ed.), *Handbook of Empirical Corporate Finance* (Vol. 1, pp. 3–36). Elsevier. <https://doi.org/10.1016/B978-0-444-53265-7.50015-9>
- Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2–3), 221–247. <https://doi.org/10.1016/j.jacceco.2008.02.003>
- Lin, H.-Y., & Hsu, B.-W. (2023). Empirical Study of ESG Score Prediction through Machine Learning—A Case of Non-Financial Companies in Taiwan. *Sustainability*, 15(19), 14106. <https://doi.org/10.3390/su151914106>
- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1), 35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- LSEG (Refinitiv). (2024). *Environmental, Social & Governance (ESG) Scores—Methodology Guide*. London Stock Exchange Group. https://www.lseg.com/content/dam/data-analytics/en_us/documents/methodology/lseg-esg-scores-methodology.pdf Accessed: 2025-07-02.
- Lyon, J. D., Barber, B. M., & Tsai, C.-L. (2003). Improved methods for tests of long-run abnormal stock returns. *Journal of Finance*, 54(1), 165–201. <https://doi.org/10.1111/0022-1082.00101>

- MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of Economic Literature*, 35(1), 13–39. <https://www.jstor.org/stable/2729691>
- McLean, R. D., & Pontiff, J. (2015). Does academic research destroy stock return predictability? *Journal of Finance*, 71(1), 5–32. <https://doi.org/10.1111/jofi.12365>
- Mitchell, M. L., & Stafford, E. (2000). Managerial decisions and long-term stock price performance. *Journal of Business*, 73(3), 287–329. <https://doi.org/10.1086/209645>
- MSCI. (2022). *ESG Ratings and research for issuers and their advisors* (Fact sheet). MSCI Inc. <https://www.msci.com/documents/1296102/10259127/ESG-Issuers-cfs-en.pdf> Accessed: 2025-07-02.
- Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies*, 22(1), 435–480. <https://doi.org/10.1093/rfs/hhn053>
- Serafeim, G., & Yoon, A. (2023). Stock price reactions to ESG news: The role of ESG ratings and disagreement. *Review of Accounting Studies*, 28(3), 1500–1530. <http://dx.doi.org/10.2139/ssrn.3765217>
- Sun, W., Luo, Y., Yiu, S.-M., Yu, L., & Ding, W. (2024). ESG scores, scandal probability, and event returns. *Financial Innovation*, 10, 121. <https://doi.org/10.1186/s40854-024-00635-1>
- Vu, T. N., Junttila, J.-P., & Lehtonen, H. (2024). ESG news and long-run stock returns. *Finance Research Letters*, 60, 104915. <https://doi.org/10.1016/j.frl.2023.104915>
- Wu, S., Irsoy, O., Lu, S., Dabrowski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance. *arXiv preprint arXiv:2303.17564*. <https://doi.org/10.48550/arXiv.2303.17564>
- Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang, S., Tong, H., Xiao, L., & Zhou, W. (2023). ESGReveal: An LLM-based approach for extracting structured data from ESG reports. *arXiv preprint arXiv:2312.17264*. <https://doi.org/10.48550/arXiv.2312.17264>

Appendix

A. Variable Definitions

Table 20: Variable Definitions

Variable	Definition
<code>base_score</code>	Gemini 1.5 Flash prediction (no context)
<code>base_a_score</code>	Gemini 1.5 Flash prediction (with prior-period anchoring)
<code>ft_score</code>	Fine-tuned Gemini prediction (no context)
<code>ft_a_score</code>	Fine-tuned Gemini prediction (with prior-period anchoring)
LRS	Last Refinitiv Score (previous period’s ESG rating)
NRS	New Refinitiv Score (current period’s ESG rating)
MAE	Mean Absolute Error = $\text{avg}(\text{prediction} - \text{actual})$
RMSE	Root Mean Squared Error = $\sqrt{\text{avg}((\text{prediction} - \text{actual})^2)}$
R^2	Coefficient of Determination
Spearman	Spearman Rank Correlation
Dir. Acc.	Directional Accuracy (% correct UP/DOWN predictions)
<code>period_length_days</code>	Days between consecutive rating changes
<code>controversies_score</code>	Refinitiv’s contemporaneous controversy intensity (0–100)
<code>model_cert</code>	Model self-reported confidence (0–1)
<code>report_length_verified</code>	Document length in characters after cleaning

Note. Variable definitions for replication. Code names in **monospace** correspond to column names in output databases and CSV files.

B. Fair Comparison Sample

The anchored models (`base_a`, `ft_a`) require the previous period’s official ESG score as context input. This design choice means anchored predictions are unavailable in two cases: (1) the first observation per instrument, where no prior score exists, and (2) observations where all prior scores are identical to the current period’s score, as anchoring on an unchanged value provides no additional signal.²¹

To ensure valid model comparisons, we construct a “fair comparison sample” ($n = 748$) containing only rating change events where all four model configurations (`base`, `base_a`, `ft`,

²¹The inference pipeline uses a `get_last_different_score()` function that explicitly requires the prior score to differ from the current period by at least 0.00001 points. This ensures the anchor conveys meaningful information about score trajectory.

ft_a) produce valid predictions. This excludes approximately 200 observations (primarily from 2014–2016, where the anchored fine-tuned model lacks sufficient prior-period context, and firms with stable ESG scores) and reflects methodological rigor in comparing models on identical samples.

Sample Reconciliation:

- Full rating change events: $N = 949$ (`days_to_next_release = 0`)
- Model evaluation sample: $N = 938$ (valid NRS/LRS scores)
- Fair comparison sample: $N = 748$ (anchored models require different prior score)
- **Primary event sample:** $N = 720$ (unique rating change events, used for all main analyses)
- Carry-forward baseline: $N = 716$ (requires any prior score)

All main results use this fair comparison sample unless otherwise noted. Full-sample results for individual models are available in the supplementary materials.

C. Additional Figures

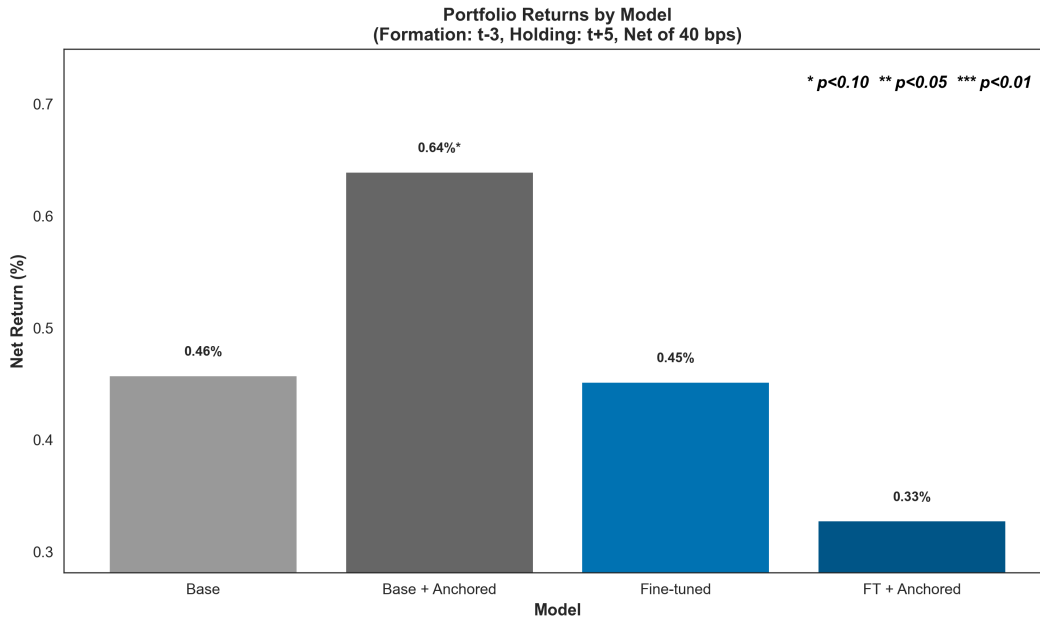


Figure 9: Portfolio Returns Comparison by Model

Note. Comparison of long-short portfolio returns across all four model configurations for the best-performing formation/holding combination. Error bars represent 95% confidence intervals based on t-statistics.

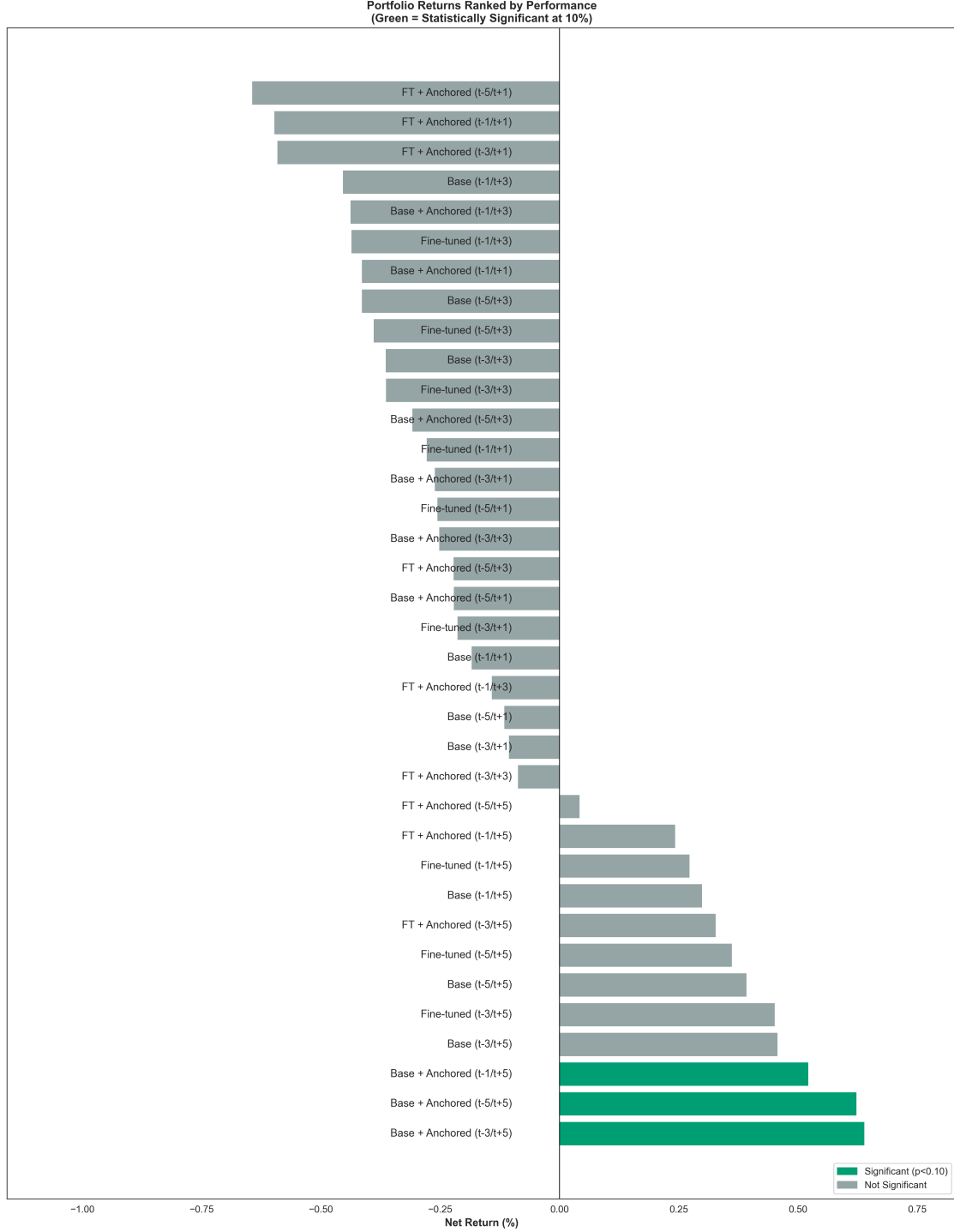


Figure 10: Statistical Significance Overview Across Configurations

Note. Visual summary of statistical significance ($p < 0.10$) across all model $\times$ formation $\times$ holding period configurations. Cells indicate whether the long-short spread achieves statistical significance. The anchored base model (base_a) with $\tau - 3$ formation and $t + 5$ holding achieves the best raw return ($p = 0.052$), while the fine-tuned model (ft) shows bootstrap-robust significance ($p = 0.001$).

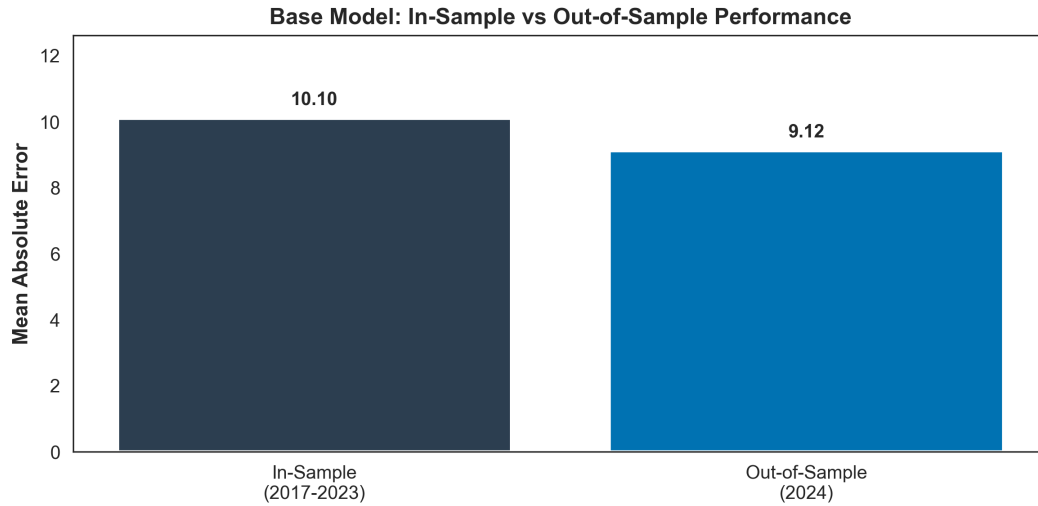


Figure 11: Base Model: In-Sample vs. Out-of-Sample Performance

Note. Comparison of base model prediction accuracy between in-sample (2019–2023) and out-of-sample (2024) periods. The stable performance across periods suggests the model generalizes well to unseen data beyond its pretraining knowledge cutoff.

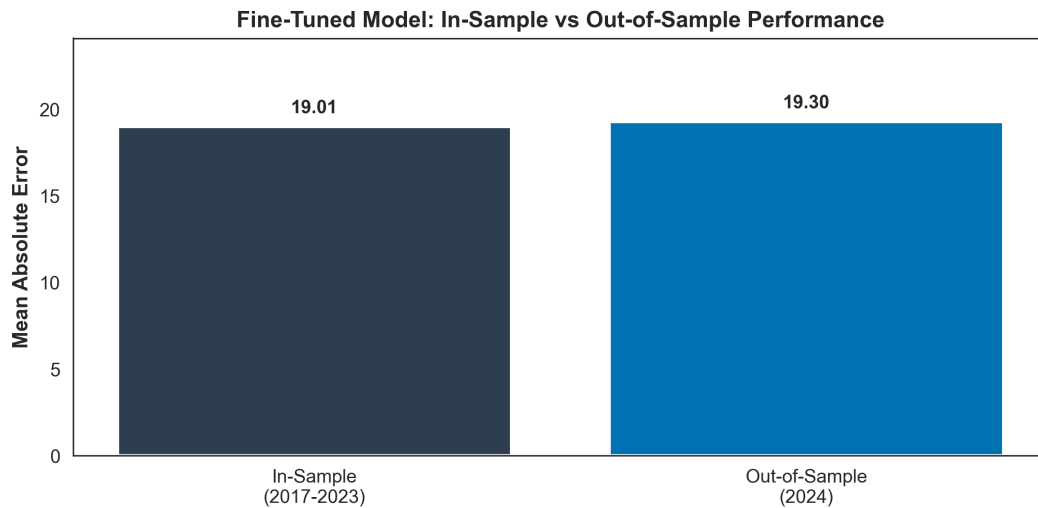


Figure 12: Fine-Tuned Model: In-Sample vs. Out-of-Sample Performance

Note. Comparison of fine-tuned model prediction accuracy between in-sample (2019–2023) and out-of-sample (2024) periods. Despite fine-tuning only on data through 2022, the model maintains consistent accuracy in 2024, demonstrating robust generalization.

D. Comprehensive Portfolio Returns

Table 21 presents the complete grid of portfolio returns across all model $\times$ formation window $\times$ holding period combinations for signal type A (simple rolling average). Signal type A is the primary specification used throughout the main text; signal types B (certainty-weighted) and C (length-weighted) are available in the supplementary CSV files for robustness but are omitted here for brevity.²²

The N values in this table represent the number of unique *portfolio formation days* for each configuration, which varies by formation window: t-1 formations have more valid days than t-3 or t-5 formations because fewer events fall outside the sample boundary. This differs from the primary event sample ($N = 720$) used for accuracy metrics, which counts unique rating change events (this also holds true for Appendix E, F).

Table 21: Comprehensive Portfolio Returns (All Configurations)

Model	Form.	Hold	Signal	N	Long	Short	Gross	Net	p-val
base	t-1	t+1	A	742	0.56%	0.34%	0.22%	-0.18%	0.380
base	t-1	t+3	A	740	1.37%	1.42%	-0.05%	-0.45%	0.894
base	t-1	t+5	A	740	2.27%	1.57%	0.70%	0.30%	0.188
base	t-3	t+1	A	701	0.60%	0.31%	0.29%	-0.11%	0.226
base	t-3	t+3	A	699	1.40%	1.37%	0.04%	-0.36%	0.931
base	t-3	t+5	A	699	2.30%	1.44%	0.86%	0.46%	0.109
base	t-5	t+1	A	701	0.60%	0.32%	0.28%	-0.12%	0.242
base	t-5	t+3	A	699	1.40%	1.42%	-0.01%	-0.41%	0.971
base	t-5	t+5	A	699	2.30%	1.51%	0.79%	0.39%	0.143
base_a	t-1	t+1	A	742	0.31%	0.33%	-0.01%	-0.41%	0.952
base_a	t-1	t+3	A	740	1.13%	1.16%	-0.04%	-0.44%	0.925
base_a	t-1	t+5	A	740	2.05%	1.12%	0.92%	0.52%	0.082
base_a	t-3	t+1	A	701	0.40%	0.26%	0.14%	-0.26%	0.562
base_a	t-3	t+3	A	699	1.26%	1.11%	0.15%	-0.25%	0.719
base_a	t-3	t+5	A	699	2.16%	1.12%	1.04%	0.64%	0.052
base_a	t-5	t+1	A	701	0.43%	0.25%	0.18%	-0.22%	0.454
base_a	t-5	t+3	A	699	1.24%	1.15%	0.09%	-0.31%	0.824
base_a	t-5	t+5	A	699	2.17%	1.14%	1.02%	0.62%	0.056
ft	t-1	t+1	A	742	0.39%	0.27%	0.12%	-0.28%	0.591
ft	t-1	t+3	A	740	1.39%	1.43%	-0.04%	-0.44%	0.929
ft	t-1	t+5	A	740	2.41%	1.73%	0.67%	0.27%	0.215
ft	t-3	t+1	A	701	0.42%	0.23%	0.19%	-0.21%	0.410
ft	t-3	t+3	A	699	1.39%	1.35%	0.04%	-0.36%	0.930
ft	t-3	t+5	A	699	2.36%	1.51%	0.85%	0.45%	0.122
ft	t-5	t+1	A	701	0.39%	0.25%	0.14%	-0.26%	0.522

Continued on next page

²²Signal type C (length-weighted) uses future document lengths in its weighting scheme, which introduces lookahead bias. It is retained in the CSV for diagnostic comparison only and should not be used for inference.

Table 21 continued

Model	Form.	Hold	Signal	N	Long	Short	Gross	Net	p-val
ft	t-5	t+3	A	699	1.38%	1.37%	0.01%	-0.39%	0.981
ft	t-5	t+5	A	699	2.33%	1.57%	0.76%	0.36%	0.169
ft_a	t-1	t+1	A	742	0.12%	0.32%	-0.20%	-0.60%	0.371
ft_a	t-1	t+3	A	740	1.30%	1.04%	0.26%	-0.14%	0.508
ft_a	t-1	t+5	A	740	2.04%	1.40%	0.64%	0.24%	0.227
ft_a	t-3	t+1	A	701	0.12%	0.31%	-0.19%	-0.59%	0.386
ft_a	t-3	t+3	A	699	1.36%	1.05%	0.31%	-0.09%	0.419
ft_a	t-3	t+5	A	699	2.10%	1.37%	0.73%	0.33%	0.170
ft_a	t-5	t+1	A	701	0.07%	0.31%	-0.24%	-0.64%	0.270
ft_a	t-5	t+3	A	699	1.22%	1.05%	0.18%	-0.22%	0.645
ft_a	t-5	t+5	A	699	1.81%	1.37%	0.44%	0.04%	0.384

Note. Comprehensive portfolio returns for signal type A (simple rolling average). Form. = formation window relative to rating release (τ); Hold = holding period after release; Long/Short = quintile portfolio returns; Gross = long-short spread before costs; Net = after 40 bps round-trip costs. Bold indicates statistical significance at $p < 0.05$. [†]**Multiple testing caveat:** With 36 configurations tested (4 models $\times$ 3 formations $\times$ 3 holdings), a Bonferroni-corrected threshold would be $p < 0.0014$; the highlighted base_a result ($p = 0.052$) does not survive this correction. We interpret this as suggestive evidence requiring out-of-sample validation and bootstrap/calendar-time robustness checks.

E. Sector-Neutral Portfolio Returns

Table 22: Sector-Neutral Portfolio Returns

Model	Form.	Hold	N	Long	Short	Gross	Net	p-val
base	t-1	t+1	742	0.30%	0.41%	-0.11%	-0.51%	0.648
base	t-1	t+3	740	1.14%	1.38%	-0.24%	-0.64%	0.521
base	t-1	t+5	740	2.02%	1.48%	0.54%	0.14%	0.328
base	t-3	t+1	701	0.34%	0.41%	-0.07%	-0.47%	0.783
base	t-3	t+3	699	1.19%	1.41%	-0.22%	-0.62%	0.583
base	t-3	t+5	699	2.09%	1.51%	0.58%	0.18%	0.288
base	t-5	t+1	701	0.34%	0.42%	-0.08%	-0.48%	0.756
base	t-5	t+3	699	1.19%	1.42%	-0.22%	-0.62%	0.579
base	t-5	t+5	699	2.09%	1.52%	0.56%	0.16%	0.308
base_a	t-1	t+1	742	0.36%	0.29%	0.07%	-0.33%	0.731
base_a	t-1	t+3	740	1.36%	1.27%	0.09%	-0.31%	0.806
base_a	t-1	t+5	740	2.19%	1.42%	0.77%	0.37%	0.181
base_a	t-3	t+1	701	0.39%	0.24%	0.15%	-0.25%	0.534
base_a	t-3	t+3	699	1.35%	1.29%	0.07%	-0.33%	0.870
base_a	t-3	t+5	699	2.17%	1.55%	0.62%	0.22%	0.249

Continued on next page

Table 22 continued

Model	Form.	Hold	N	Long	Short	Gross	Net	p-val
base_a	t-5	t+1	701	0.36%	0.26%	0.10%	-0.30%	0.650
base_a	t-5	t+3	699	1.27%	1.29%	-0.02%	-0.42%	0.957
base_a	t-5	t+5	699	2.04%	1.52%	0.52%	0.12%	0.334
ft	t-1	t+1	742	0.17%	0.32%	-0.15%	-0.55%	0.466
ft	t-1	t+3	740	1.06%	1.33%	-0.27%	-0.67%	0.490
ft	t-1	t+5	740	2.07%	1.80%	0.26%	-0.14%	0.636
ft	t-3	t+1	701	0.19%	0.33%	-0.15%	-0.55%	0.498
ft	t-3	t+3	699	1.07%	1.35%	-0.29%	-0.69%	0.463
ft	t-3	t+5	699	2.06%	1.68%	0.38%	-0.02%	0.470
ft	t-5	t+1	701	0.18%	0.32%	-0.14%	-0.54%	0.532
ft	t-5	t+3	699	1.06%	1.32%	-0.27%	-0.67%	0.492
ft	t-5	t+5	699	2.05%	1.59%	0.46%	0.06%	0.384
ft_a	t-1	t+1	742	0.15%	0.44%	-0.29%	-0.69%	0.205
ft_a	t-1	t+3	740	1.30%	1.05%	0.25%	-0.15%	0.520
ft_a	t-1	t+5	740	1.85%	1.35%	0.50%	0.10%	0.336
ft_a	t-3	t+1	701	0.17%	0.42%	-0.25%	-0.65%	0.291
ft_a	t-3	t+3	699	1.35%	1.02%	0.33%	-0.07%	0.379
ft_a	t-3	t+5	699	1.92%	1.28%	0.64%	0.24%	0.197
ft_a	t-5	t+1	701	0.12%	0.41%	-0.28%	-0.68%	0.272
ft_a	t-5	t+3	699	1.29%	1.02%	0.28%	-0.12%	0.475
ft_a	t-5	t+5	699	1.83%	1.28%	0.55%	0.15%	0.277

Note. Sector-neutral portfolio returns control for GICS sector exposures before forming quintile portfolios. This approach addresses concerns that apparent ESG alpha might reflect sector tilts rather than firm-specific ESG information. Results are qualitatively similar to unadjusted returns, though significance levels are generally reduced due to lower cross-sectional dispersion after sector adjustment.

F. Tertile Portfolio Returns

Table 23: Sector-Neutral Tertile Portfolio Returns

Model	Form.	Hold	N	Long	Short	Gross	Net	p-val
base	t-1	t+1	700	0.20%	0.29%	-0.09%	-0.49%	0.546
base	t-1	t+3	698	1.11%	1.42%	-0.31%	-0.71%	0.246
base	t-1	t+5	698	1.92%	1.47%	0.45%	0.05%	0.260
base	t-3	t+1	697	0.21%	0.30%	-0.09%	-0.49%	0.517
base	t-3	t+3	695	1.13%	1.44%	-0.32%	-0.72%	0.279
base	t-3	t+5	695	1.91%	1.53%	0.38%	-0.02%	0.347
base	t-5	t+1	697	0.21%	0.30%	-0.09%	-0.49%	0.535

Continued on next page

Table 23 continued

Model	Form.	Hold	N	Long	Short	Gross	Net	p-val
base	t-5	t+3	695	1.12%	1.45%	−0.33%	−0.73%	0.228
base	t-5	t+5	695	1.91%	1.55%	0.36%	−0.04%	0.378
base_a	t-1	t+1	700	0.24%	0.31%	−0.07%	−0.47%	0.594
base_a	t-1	t+3	698	1.13%	1.27%	−0.15%	−0.55%	0.600
base_a	t-1	t+5	698	1.92%	1.56%	0.36%	−0.04%	0.388
base_a	t-3	t+1	697	0.32%	0.26%	0.05%	−0.35%	0.865
base_a	t-3	t+3	695	1.20%	1.22%	−0.02%	−0.42%	0.901
base_a	t-3	t+5	695	2.01%	1.57%	0.45%	0.05%	0.286
base_a	t-5	t+1	697	0.33%	0.20%	0.12%	−0.28%	0.560
base_a	t-5	t+3	695	1.21%	1.15%	0.06%	−0.34%	0.874
base_a	t-5	t+5	695	2.03%	1.47%	0.56%	0.16%	0.186
ft	t-1	t+1	700	0.21%	0.38%	−0.17%	−0.57%	0.256
ft	t-1	t+3	698	1.19%	1.34%	−0.15%	−0.55%	0.601
ft	t-1	t+5	698	1.99%	1.55%	0.44%	0.04%	0.240
ft	t-3	t+1	697	0.25%	0.34%	−0.10%	−0.50%	0.588
ft	t-3	t+3	695	1.19%	1.30%	−0.11%	−0.51%	0.609
ft	t-3	t+5	695	1.95%	1.47%	0.49%	0.09%	0.269
ft	t-5	t+1	697	0.24%	0.34%	−0.10%	−0.50%	0.577
ft	t-5	t+3	695	1.16%	1.30%	−0.14%	−0.54%	0.650
ft	t-5	t+5	695	1.91%	1.45%	0.46%	0.06%	0.286
ft_a	t-1	t+1	700	0.13%	0.25%	−0.12%	−0.52%	0.484
ft_a	t-1	t+3	698	1.27%	0.96%	0.31%	−0.09%	0.373
ft_a	t-1	t+5	698	1.88%	1.35%	0.53%	0.13%	0.240
ft_a	t-3	t+1	697	0.11%	0.24%	−0.13%	−0.53%	0.445
ft_a	t-3	t+3	695	1.22%	0.96%	0.26%	−0.14%	0.359
ft_a	t-3	t+5	695	1.85%	1.30%	0.55%	0.15%	0.148
ft_a	t-5	t+1	697	0.08%	0.24%	−0.16%	−0.56%	0.341
ft_a	t-5	t+3	695	1.17%	0.98%	0.19%	−0.21%	0.520
ft_a	t-5	t+5	695	1.78%	1.31%	0.47%	0.07%	0.235

Note. Tertile (tercile) portfolio analysis uses three groups instead of five quintiles, providing more observations per portfolio at the cost of less extreme signal separation. Results are directionally consistent with quintile analysis, with the fine-tuned model showing positive spreads in longer holding periods. Statistical significance is reduced due to less extreme portfolio positioning.

Author Biography

Luca Johann Schmidt is a doctoral candidate in Finance at EBS Universität für Wirtschaft und Recht, co-supervised with New York University (NYU). He holds an MSc in Finance from Nova School of Business and Economics and a BSc in Business Studies from EBS Universität für Wirtschaft und Recht. His research focuses on applying large language models and computational methods to financial markets, with particular emphasis on ESG ratings, information diffusion, and market efficiency. Prior to his doctoral studies, he gained professional experience at UBS Europe SE in Global Family and Institutional Wealth, Pictet in Discretionary Portfolio Management, and other roles in empirical asset management, corporate finance, consulting, and audit (all positions held as intern or working student during studies). His current working papers include (i) ESGemini: Can LLMs Outpace the Market, Allowing for Information Arbitrage? and (ii) Sentiment in the Tails: Distinguishing Different Types of Stock Market Crashes, with a third planned project on the geopolitical risk premium in equity markets.

Data and Code Availability

All analysis code, model predictions, and replication materials are publicly available at: <https://github.com/LucaJoS/ESGemini-Can-LLMs-Outpace-the-Market-Allowing-for-Information-Arbitrage->. CRSP return data require a separate institutional license and are not included. Refinitiv/LSEG ESG disclosures and scores were accessed through Refinitiv Workspace under academic license and are also not included.

Statement on AI Use

This research uses Google’s Gemini 1.5 Flash as the subject of investigation for ESG score replication. Additionally, AI assistants (Claude, Anthropic) were used for code debugging, LaTeX formatting, and editorial refinement. All statistical analyses, methodological decisions, and interpretations remain the sole responsibility of the author. The use of AI tools is disclosed in accordance with emerging best practices in academic publishing.

Document last updated: January 15, 2026