Future Technology
Open Access Journal
ISSN 2832-0379

November 2026| Volume 05 | Issue 04 | Pages 181-198

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.4.15

Future Publishing LLC

Article

A transformer-based explainable machine learning framework for heterogeneous ESG risk and return prediction using text intelligence

Feifei Fan, Asri Marsidi*

Faculty of Economics and Business, Universiti Malaysia Sarawak (UNIMAS), 94300 Kota Samarahan, Malaysia

ARTICLE INFO

Article history:
Received 24 March 2026
Received in revised form
15 August 2026
Accepted 03 September 2026

Keywords:
ESG disclosure, Long-document encoding, Sparse expert routing, Explainable AI, Downside risk, Text mining

*Corresponding author
Email address:
maasri@unimas.my

DOI: 10.55670/fpll.futech.5.4.15

ABSTRACT

Corporate ESG disclosure has shifted toward narrative form, yet vendor ESG ratings that most empirical work depends on show substantial disagreement across providers, and encoders dominant in financial text analysis inherit a 512-token ceiling that discards most of a typical 10-K. HET-ESGFormer combines hierarchical long-document encoding, heterogeneity-aware sparse expert routing conditioned on industry sector and firm size, and a dual-channel explanation layer subject to faithfulness verification. The framework is evaluated on 3,182 U.S.-listed firms over 2015 to 2024 using 10-K filings, sustainability reports, earnings call transcripts, and ESG-tagged news. Against nine baselines spanning econometric, dictionary-based, contextual encoder, sparse-attention, and time-series approaches, HET-ESGFormer reduces return RMSE from 0.0417 to 0.0389 and QLIKE from 0.148 to 0.128, with Diebold-Mariano tests rejecting equal predictive accuracy against every baseline after correction for multiple comparisons. The text channel's marginal contribution to downside risk prediction is materially larger than its contribution to return prediction, widening at longer forecast horizons, indicating that ESG narratives carry more information about risk than about expected return. Ablations reveal a decoupling between predictive skill and explanation faithfulness: removing the consistency regularizer barely affects accuracy but sharply degrades explanation quality. The dual-channel explanation improves ERASER comprehensiveness by 50% over SHAP applied to XGBoost, and human raters significantly prefer the framework's rationales over SHAP-based rationales on trustworthiness. Decile-sorted long-short portfolios deliver a post-cost Sharpe ratio of 1.31 against 0.74 and 0.62 for portfolios built on Refinitiv and MSCI ratings, and pass VaR backtests at 95% coverage. The results support text-based ESG modeling as an alternative to vendor-rating-based approaches, particularly where downside risk management and traceable explanation are jointly required.

1. Introduction

The disclosure of environmental, social, and governance (ESG) information by enterprises has shifted significantly toward narrative form. Assets under management under sustainable investment strategies have grown substantially, and disclosure volume has also expanded. Mandatory reporting has further increased the flow of ESG-related text that enterprises release [1]. Annual reports, sustainability reports, calls, and press releases now generate large volumes of ESG text. However, most empirical studies rely on ESG ratings from commercial rating agencies as model inputs, despite large inconsistencies in ratings for the same company across agencies [2]. This inconsistency creates a pricing effect, under which companies with larger rating discrepancies earn larger return premiums [3]. Thus, vendor-rating inputs carry measurement error, whereas text can be mapped directly to market outcomes. Most ESG narratives contain tens of thousands of tokens, and current encoder models used to analyze financial texts have an upper bound of 512 tokens due to the quadratic cost of full self-attention [4]. Financially material sustainability issues vary across industries because companies that perform well on financially material issues specific to their industry perform better on such issues than companies that perform well on non-financially material issues [5]. If estimated with enough flexibility, the relationship between ESG aspects and returns is reversed for small versus large-capitalization firms [6]. Therefore, a unified worldwide calibration captures at most an

approximate compromise between these opposing systems. Financial organizations also operate under regulatory assumptions that require traceability of modeling results, a key unresolved issue in the area [7]. Post-hoc attribution is common, but ESG studies rarely test whether attributions faithfully reflect model computation. HET-ESGFormer addresses three linked weaknesses in ESG prediction: long disclosures exceed standard encoder limits, vendor ratings contain measurement disagreement, and model explanations are rarely tested for faithfulness. The framework hierarchically encodes disclosures, routes firms through sector- and size-conditioned sparse experts, and jointly predicts return and downside risk. Its contributions are ESG-domain adaptation, interpretable soft-regime routing, dual-channel explanation verified by ERASER metrics, and direct measurement of risk-return asymmetry in the text channel.

2. Related work

2.1 ESG rating divergence

The divergence across ESG rating agencies can be decomposed into three components: scope, measurement, and weighting. Among these, the measurement factor contributes the most. There is also a "rater effect", whereby a provider's overall assessment of an enterprise colors its scoring across individual subcategories, so residuals are correlated within the same provider rather than being independent [8]. This pattern strengthens with increased information disclosure rather than being resolved by it. The more ESG information an enterprise discloses, the greater the disagreement across rating agencies, and the disagreement itself is associated with higher return volatility, larger absolute price movements, and reduced access to external financing [2]. Therefore, empirical conclusions in the existing literature depend on which vendor provides the ratings, which motivates constructing labels based on market outcomes and reserving vendor ratings for benchmark comparison only.

2.2 Text mining in finance

Early financial text analysis mainly relied on carefully curated lexicons and bag-of-words representations, though their limitations have been pointed out by the developers themselves [9]. Contextual encoders removed this constraint. A finance-adapted BERT substantially outperformed dictionary and LSTM sentiment models, especially with small samples and domain-specific vocabulary. This advantage also extended to ESG classification [10]. ESG-domain pre-training improves environment-related tasks [11] but real disclosure remains noisy. When a domain-adapted encoder was applied to climate reporting, it was found that firms endorsing voluntary reporting frameworks did not measurably increase substantive disclosure, but rather preferentially reported on non-material categories [12]. ESG-text models must therefore identify uninformative boilerplate rather than treating all sentences as equally useful.

2.3 Long-document encoding in financial text

An average 10-K document typically contains approximately 37,400 words [13], and the ESG-related sections usually occupy the latter part of the document. When the input is truncated from the beginning, these contents are often lost. Common shortcuts encode selected 512-token windows or pooled segments, losing cross-segment context. Sparse attention combines local windowed attention with task-driven global attention to achieve linear scaling and can handle sequences of 4,096 tokens or longer [4]. Another solution would be hierarchical encoding, which surpasses the performance of sparse-attention models of similar size in classifying long documents while using 10-20% less GPU memory and processing documents 40-45% faster [14]. This hierarchy also fits ESG reports, whose sentence- and topic-level structure supports evidence retrieval.

2.4 Parameter sharing across firms and tasks

Materiality differs by industry. Meanwhile, the relationship between ESG score and return changes direction from small to large capitalization companies with controversial aspects exhibiting the highest explanatory power. An estimated globally parameterized model for such opposite regimes would fit none. This role is fulfilled by sparsely gated mixture-of-experts layers that send the input to a small number of experts chosen by the learned gate, with an additional load balancing objective to stop the system from favoring a certain set of experts [15]. Routing on observable firm features makes the learned regime assignment interpretable. Task-level sharing also poses a related problem on the output side. The returns and volatility have different magnitudes and signal-to-noise ratios, and there is evidence that ESG attributes can help lower both total and firm-specific risk via channels partially separate from those that affect expected returns [16]. Homoscedastic uncertainty weighting therefore lets tasks with different scales learn jointly without manual loss tuning [17].

2.5 Explainability and faithfulness

SHAP and LIME are the predominant agnostic methods used in finance, but their compliance with regulations is still an open question [7]. Integrated gradients have a stronger theoretical basis, being sensitive and invariant to the method of their computation, and do not require any changes in the neural network architecture [18]. Whether the attention mechanism itself can be considered an explanation is still an open question. There is also the perspective that if different attention weightings can lead to similar prediction outcomes, then the original attention weightings cannot be taken as an explanation for the model [19]. An opposing response suggests other approaches such as uniform weighting baseline tests, variance adjustment among different seeds, frozen weight tests, and adversarial training, whereby the attention model can be considered an explanation under some criteria [20]. This debate remains unsettled because it concerns the standard for treating saliency as explanation. A diagnostic comparison across techniques and datasets finds that no single method dominates on every faithfulness property, with gradient-based attributions performing most consistently but not uniformly [21]. Recent work also regularizes attribution agreement or human-rationale alignment, but ESG evidence remains limited.

Faithfulness metrics shift the emphasis from argument to measurement. Comprehensiveness refers to the degree to which model confidence drops in the absence of the identified reasoning process. Sufficiency refers to the degree to which confidence persists in the presence of the reasoning process alone. Faithfulness involves high comprehensiveness and low sufficiency [22]. Recent ESG applications have proceeded without this step. A greenwashing prediction framework combining hyperparameter-optimized XGBoost with SHAP attribution reports an R^2 of 0.979 on Chinese listed firms and interprets the resulting feature importances as regulatory guidance, without any check of whether these attributions faithfully reflect the model's computation [23]. This gap leaves the possibility open that the explanation diverges from the computation.

3. Methodology

3.1 Notation and framework

Let i index firms and t index time. For each firm–time pair, the model takes three input objects. The first is a set of ESG documents $D_{i,t}$ that includes the ESG-related sections of 10-K filings, standalone sustainability reports, earnings call transcripts, and ESG-tagged news. The second is a structured feature vector $x_{i,t} \in \mathbb{R}^{dx}$ containing accounting fundamentals and market factors. The third is a pair of heterogeneity attributes $z_{i,t}$ = *(sector i, size i, t)*. Two prediction targets are defined at horizon h. The return target $r_{i,t+h}$ is the cumulative market-adjusted return over *[t, t+h]*. The risk target has two parts: the realized volatility $\sigma_{i,t+h}$ is computed from daily returns within the same window, and the downside risk is measured by value-at-risk VaR at level α. All targets are constructed solely from market data rather than from vendor ESG ratings, so the framework does not inherit the rating-divergence problems documented earlier. The main notation and algorithms are reported in Appendix A. The framework is trained end-to-end against a composite objective,

$$\mathcal{L} = \lambda_{ret}\mathcal{L}_{ret} + \lambda_{risk}\mathcal{L}_{risk} + \lambda_{bal}\mathcal{L}_{bal} + \lambda_{agree}\mathcal{L}_{agree} + \lambda_{orth}\mathcal{L}_{orth} + \lambda_{share}\mathcal{L}_{share} \quad (1)$$

where L*ret and* L*risk* are the return and risk task losses, L_{bal} the expert load-balancing term, L*agree* the dual-channel consistency term, L*orth the E/S/G orthogonality term, and* L*share* the soft parameter-sharing prior across routed experts within a sector cluster. The task-loss weights λ_{ret} and λ_{risk} are set by the homoscedastic uncertainty-weighting scheme [17] rather than by grid search, and the regularizer weights λ_{bal}, λ_{agree}, λ_{orth}, λ_{share} are fixed constants. After representation, hierarchical encoding, disentanglement, fusion, and routing, the model predicts risk and return, and Figure 1 shows the architecture.

3.2 ESG-domain text representation

The original disclosure document undergoes two preprocessing stages before any learned representation is generated. The first is a boilerplate filter that removes paragraphs almost verbatim copied by firms across consecutive annual reports. Prior work has shown that much of this content is cheap talk that does not reflect substantive activity [12].

Each candidate paragraph is decomposed into overlapping word n-grams and compressed into a fixed-length signature via the MinHash algorithm [24]. If the Jaccard similarity between paragraphs of two consecutive filings, estimated from signature agreement, exceeds 0.85, they are treated as near-duplicates, and the later occurrence is dropped. The filtered corpus is used for continued pre-training of a BERT-family encoder initialized with FinBERT weights, since domain-adapted initialization has been shown to outperform general BERT on both financial and ESG classification tasks [10]. Pre-training uses masked language modeling with ESG-aware span masking. Rather than randomly masking individual tokens, references to ESG-related items such as emission types, employee data, and governance systems are detected through a manually curated seed list of about 400 ESG terms expanded to their inflected forms, and each matched span is masked in its entirety with a probability raised by a factor of $\eta = 2$ relative to non-ESG spans, so that ESG spans are masked at a 30% rate against a 15% base rate. This concentrates the pre-training signal on words that matter for the downstream task while maintaining language generality.

3.3 Hierarchical long-document encoding

The typical 10-K report is about 37,400 words, with ESG information dispersed throughout [13]. To represent such a document, one must circumvent the 512-token limit of the encoder while still allowing the next layer to identify which sentences drive the prediction. The encoder is organized in three levels. At the sentence level, each sentence j is passed through the domain-pretrained encoder E and represented by its [CLS] embedding,

$$h_j^{(s)} = \mathcal{E}(\text{sentence}_j) \quad (2)$$

Neighboring sentences are grouped into topic segments by first splitting the document at its native section boundaries (10-K item headings and report section titles) and then dividing any section longer than *Lseg*=30 sentences at lexical-cohesion minima following the TextTiling criterion [25], with each segment capped at L_{seg} sentences. These segments correspond to disclosure themes such as environmental, workforce, and governance content.

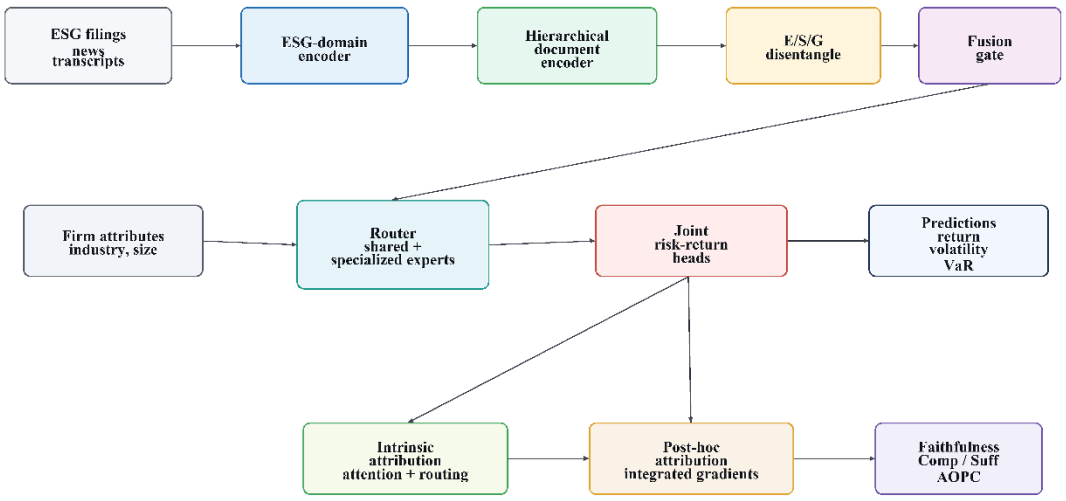

Figure 1. HET-ESGFormer architecture

A segment representation $g_m^{(seg)}$ is produced by attention pooling over its sentence embeddings,

$$a_j = \text{softmax}_j\left(q^\top \tanh\left(W_p h_j^{(s)}\right)\right), g_m^{(seg)} = \sum_{j\in S_m} a_j\ h_j^{(s)} \quad (3)$$

where S_m is the sentence set of segment m, W_p a learned projection, and q a learned query vector shared across segments; the weights a_j are reused by the explanation layer. At the document level a Transformer operates over the segment representations to yield the document embedding,

$$\hat{h}m = \text{Transformer}\left(g_m^{(seg)}\right), hd_k = \text{meanpool}_m\left(\hat{h}_m\right) \quad (4)$$

so that context can pass across sections, allowing an environmental commitment made in one section to be qualified by operations described in another. This hierarchy outperforms sparse-attention encoders such as Longformer and BigBird on long-document classification while using less GPU memory [14, 26].

At any time t, a firm has several documents open at once, including the latest 10-K, that year's sustainability report, and several ESG-tagged news items. These document embeddings are aggregated into a single firm–time embedding through time-decayed weighting. For document d_k released at time t_k,

$$h_{i,t} = \sum_k \frac{\exp(-\lambda(t-t_k))}{\sum_{k'}\exp(-\lambda(t-t_{k'}))}\ hd_k \quad (5)$$

where $\lambda \geq 0$ is a learned decay rate, initialized at a twelve-month half-life ($\lambda_0 ln2/365$ per day) and constrained to be non-negative. Recent documents receive higher weight while stale filings retain a residual influence rather than being discarded at an arbitrary cutoff. Only documents with a release timestamp $t_k \leq t$ enter the sum, so that no information dated after t influences the firm–time embedding. Figure 2 illustrates the encoder structure and the empirical difference between the hierarchical and truncation-based schemes.

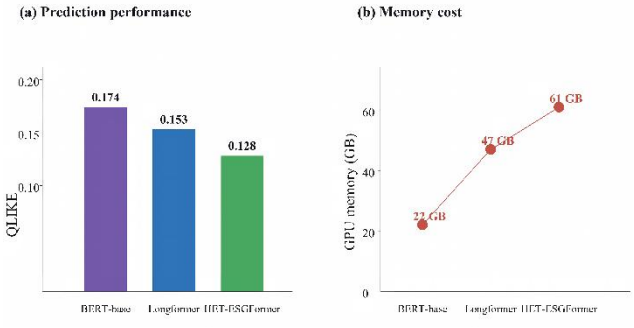

Figure 2. Hierarchical long-document encoder and truncation-based comparison

3.4 E/S/G disentanglement

The firm–time embedding $h_{i,t}$ integrates environmental, social, and governance information across all dimensions. Although convenient for prediction, this hinders per-pillar attribution. Therefore, before fusion, disentanglement is applied so that the explanation layer operates on axes carrying pillar semantics. Three learned projection matrices P_E, P_S, P_G map $h_{i,t}$ into three subspaces, and the concatenation $h_{i,t}^{ESG} = [P_E h_{i,t}; P_S h_{i,t}; P_G h_{i,t}]$ is passed forward. Without regularization, the three subspaces may overlap arbitrarily, so an orthogonality penalty drives their pairwise off-diagonal Gram matrices toward zero,

$$\mathcal{L}_{orth} = \sum_{p\neq q} \|P_p^\top P_q\|_F^2, \qquad p,q \in \{E,S,G\} \quad (6)$$

3.5 Cross-modal fusion and expert routing

Because structured features and text embeddings differ in scale and noise, gated fusion learns per-dimension modality weights,

$$g = \sigma\left(W_g\left[h_{i,t}^{ESG}; W_x x_{i,t}\right] + b_g\right), u_{i,t} = g \odot h_{i,t}^{ESG} + (1-g) \odot W_x x_{i,t} \quad (7)$$

where σ is the sigmoid and ⊙ the Hadamard product. Because the two modalities enter additively under a modality-specific gate, the marginal contribution of the text channel is recovered by zeroing $h_{i,t}^{ESG}$ and re-evaluating the prediction.

The fused representation $u_{i,t}$ is routed through a mixture-of-experts layer. The heterogeneity attributes $z_{i,t}$ are embedded into a dense vector e_z and concatenated with $u_{i,t}$ before a softmax router assigns probabilities p over experts, so routing depends on firm sector and size as well as fused content. Following DeepSeek-MoE [27], the expert pool is split into N_s shared experts, active for every input, and N_r routed experts, of which the top-k are active per input,

$$v_{i,t} = \sum_{j=1}^{N_s} \text{FFN}_j^{(s)}\left(u_{i,t}\right) + \sum_{j\in \text{Top}-k(p)} p_j\ \text{FFN}_j^{(r)}\left(u_{i,t}\right) \quad (8)$$

Shared experts absorb signals common across all firms so they are not learned redundantly by every routed expert. Routed experts specialize by heterogeneity regime, and because the router is conditioned on $z_{i,t}$, the vector p doubles as a soft regime label. To prevent collapse onto a few experts, a load-balancing loss is imposed following the Switch Transformer [15],

$$\mathcal{L}_{bal} = N_r \sum_{j=1}^{N_r} f_j\ P_j \quad (9)$$

where f_j is the fraction of tokens routed to expert j in a batch and P_j its average router probability; the product is minimized when both are uniform. The pool comprises $N_s = 2$ shared and $N_r = 16$ routed experts with top-k = 2 activation. Sectors with few firms would otherwise starve their specialized experts of training signal. The routed experts are therefore grouped into sector clusters $\mathcal{C}_c$, and a soft parameter-sharing prior [28] pulls the parameters $\theta_j^{(r)}$ of experts in the same cluster toward their cluster mean,

$$\mathcal{L}_{share} = \sum_c \sum_{j\in\mathcal{C}_c} \left\|\theta_j^{(r)} - \bar{\theta}_c\right\|_2^2, \qquad \bar{\theta}_c = \frac{1}{|\mathcal{C}_c|}\sum_{j\in\mathcal{C}_c}\theta_j^{(r)} \quad (10)$$

so that small-sample sectors benefit from the pooled statistical strength of their cluster while retaining expert-specific parameters. Figure 3 depicts the routing structure.

3.6 Complexity of encoding and routing

The computational cost is dominated by the encoder and the mixture-of-experts layer, both structured to scale sub-quadratically. For a document with L tokens divided into S sentences of average length L/S, sentence-level attention costs $O(S \cdot (L/S)^2) = O(L^2/S)$ and segment-level attention over segment vectors costs $O(S^2)$, giving total encoding cost $O(L^2/S + S^2)$, roughly two orders of magnitude below full self-attention's $O(L^2)$ for realistic 10-K sizes. For the mixture-of-experts layer, dense evaluation of all N_r routed experts on a batch of B inputs would cost $B \cdot N_r \cdot C_{FFN}$. Top-k routing activates only $k \ll N_r$ routed experts per input, so the actual cost is

$$C_{\text{MoE}} = B(N_s + k)C_{\text{FFN}} \quad (11)$$

with N_s small and k typically 2. Cost is decoupled from N_r, so expert capacity can be increased without a corresponding growth in FLOPs. The theoretical activation ratio $(N_s + k)/(N_s + N_r)$ is compared against measured activation in the empirical results.

3.7 Joint risk–return prediction

The routed representation $v_{i,t}$ feeds two task-specific heads. The return head is a two-layer MLP with a linear output predicting $r_{i,t+h}$ under the Huber loss, which is more robust than squared error to the heavy tails of financial returns. The risk head handles two components. Realized volatility is predicted by direct regression under squared-log error. Downside risk is estimated by quantile regression, in which a separate output predicts the α-quantile of the return distribution and minimizes the pinball loss [29],

$$\mathcal{L}_\alpha(y, \hat{q}) = \max\big(\alpha(y - \hat{q}), (\alpha - 1)(y - \hat{q})\big) \quad (12)$$

Setting $\alpha = 0.05$ yields a direct estimate of $VaR_{0.05}$ without assuming any parametric return distribution. The overall risk loss aggregates the volatility and quantile terms with equal weight, and the return–risk trade-off is set by learned homoscedastic uncertainty parameters following [17].

The gated fusion supplies the machinery for a distinct empirical question. Let M denote the trained model. The marginal contribution of the text channel to task $\tau \in$ {return,risk} at input (i, t) is:

$$\Delta_\tau(i,t) = M_\tau(u_{i,t}) - M_\tau(u_{i,t} \mid h_{i,t}^{ESG} = 0) \quad (13)$$

Averaging Δ_τ over the test cross-section gives a task-specific measure of the text contribution; comparing Δ_{return} with Δ_{risk} tests whether the two targets benefit from text to the same degree. All features are constructed on a point-in-time basis using data available at t and no later, with training strictly on $[\cdot, t)$ and evaluation on $[t, t + h]$. A document enters the aggregation of the time-decay equation only if its release timestamp $t_k \le t$, where t_k is the filing date for 10-K and 10-Q items, the disclosure date for sustainability reports, the call date for transcripts, and the event date for news, with each document treated as available only from the following trading day.

3.8 Dual-channel explanation

The explanation layer combines intrinsic and post-hoc attributions under a consistency constraint. The intrinsic channel uses sentence attention and expert-routing weights to identify influential sentences and firm regimes. Combining the two yields an attribution over source sentences fully traceable to the forward pass. The post-hoc channel applies integrated gradients [18], treating each input dimension's contribution as the integral of the model's gradient along a straight-line path from a baseline to the actual input; integrated gradients satisfy sensitivity and implementation invariance axiomatically and require no change to the network, so they serve as a clean external reference against which to check the intrinsic channel. Both channels produce an attribution vector over the same source-sentence space. Let $a_{int}(i,t)$ and $a_{ig}(i,t)$ denote the intrinsic and integrated-gradients attributions, each ℓ_1-normalized. A consistency term drives them toward agreement during training,

$$\mathcal{L}_{agree} = \mathbb{E}(i,t)\Big[\mathrm{KL}\Big(a_{int}(i,t) \;\|\; a_{ig}(i,t)\Big)\Big] \quad (14)$$

This penalizes the intrinsic channel for producing attributions the post-hoc channel cannot corroborate and, through backpropagation into the shared representation, shapes the model's internals so the two channels agree. The approach addresses the criticism that attention weights need not correspond to what drives predictions [19] while retaining the view that, under appropriate definitions and tests, attention can serve as an explanation [20]. The layer reports firm-level E/S/G decomposition, pillar-topic attribution, issue rankings, and source-sentence rationales.

Whether the two channels agree is not the same as whether either is faithful to the model's computation. Faithfulness is measured following ERASER [22]. For a top-k rationale R_k, comprehensiveness records how far the model's confidence falls when the rationale is removed and sufficiency measures how well the rationale alone reproduces the prediction; both are integrated over k using the area over the perturbation curve:

$$\text{Comp} = \frac{1}{|K|}\sum_{k\in K}[M(x) - M(x\backslash R_k)], \quad \text{Suff} = \frac{1}{|K|}\sum_{k\in K}[M(x) - M(R_k)]$$

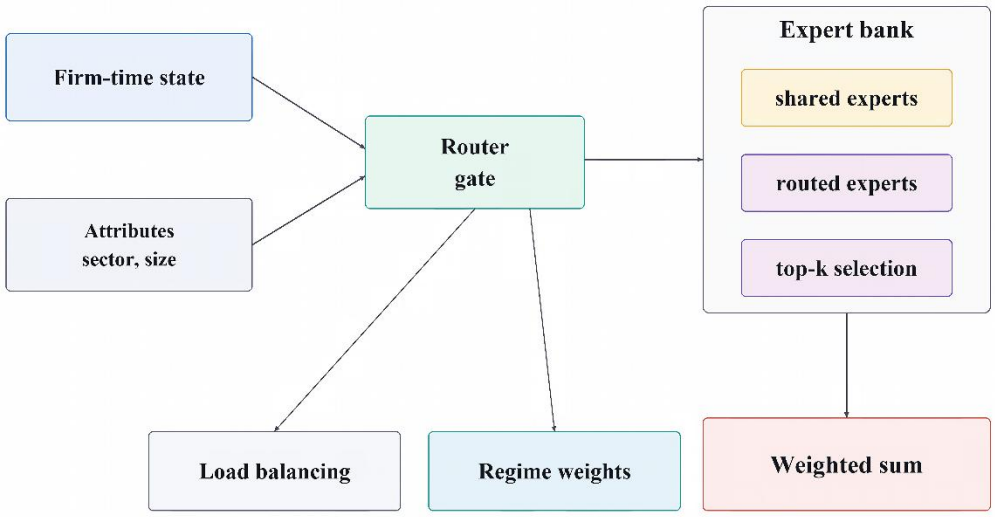

Figure 3. Heterogeneity-aware routing with shared and specialized experts

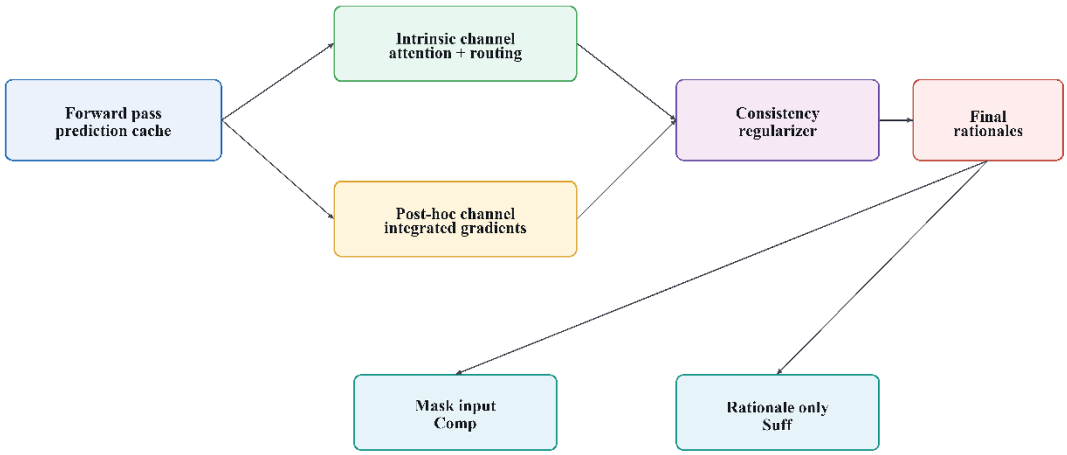

Figure 4. Dual-channel explanation pipeline

Faithful attribution has high comprehensiveness and low sufficiency. These metrics are reported for the framework and baseline attributions. Figure 4 summarizes the explanation pipeline, and Appendix A provides the detailed training and inference procedures.

4. Experimental setup

4.1 Data

Each data source used is publicly available or accessible through standard institutional subscriptions. Table 1 shows the eight data sources and their access methods.

Table 1. Data sources and access modes

Category	Source	Access
10-K and 10-Q ESG sections	SEC EDGAR Full-Text Search	Fully public via official API
CSR and sustainability reports	Company websites; GRI Sustainability Disclosure Database	Public
Earnings call transcripts	Seeking Alpha; Refinitiv Eikon	Partial public; institutional subscription
ESG news	GDELT Project	Fully public
Prices and returns	CRSP via WRDS; Yahoo Finance	University subscription; public
Financial fundamentals	Compustat via WRDS	University subscription
Risk factors	Kenneth French Data Library	Fully public
ESG ratings (benchmark only)	Refinitiv; MSCI	Institutional subscription

The prediction labels are constructed solely from market data. The return targets use CRSP-adjusted returns, while the risk targets use daily realized volatility. Vendor ESG ratings are never included in the main framework as labels or features, since treating vendor ratings as ground truth would contradict the rating-divergence evidence that motivates the design; vendor ratings are used only in the portfolio-comparison benchmark. The sample covers 2015 to 2024, restricted to U.S.-listed firms with available EDGAR and CRSP coverage during the period. The heterogeneity attributes are two-digit SIC industry codes and the market-capitalization decile at t.

Cross-market extension is left for future research, as within-market industry heterogeneity is already sufficient to test the validity of the framework. The sample statistics are presented in Table 2.

Table 2. Sample composition

Statistic	Value
Firms	3,182
Firm-year observations	28,647
10-K filings	27,913
Sustainability reports	11,405
Earnings call transcripts	62,338
ESG news items	1,842,617
Median 10-K length (tokens)	42,183
Median sustainability report length (tokens)	24,776
Industries covered (2-digit SIC)	63
Firms in bottom size decile	318
Firms in top size decile	316

Because the firm–time embedding aggregates several document types with different release cadences, each document is admitted at time t only when its release timestamp $t_k \leq t$: the EDGAR acceptance date for 10-K and 10-Q items, the GRI database disclosure date (or the company release date when the former is missing) for sustainability reports, the call date for earnings transcripts, and the event date for GDELT news, with availability lagged to the following trading day. Under this strict publication-date filter, 2.3% of firm–quarter observations contained at least one document whose original assignment preceded its availability date; re-timestamping altered the aggregated embedding for 0.6% of observations and shifted full-sample return RMSE by less than 0.0003 and QLIKE by less than 0.002. All reported results use the re-timestamped sample.

4.2 Baselines

Baseline models are chosen so that each addresses one distinct claim of the framework. The Fama–French five-factor regression checks whether text is needed at all [30]. XGBoost on TF-IDF features verifies whether deep semantic representation improves on shallow term counts. LM dictionary sentiment [9] and LDA topic weights [31] test whether curated financial NLP suffices. FinBERT [10] and ESG-BERT [11] evaluate whether ESG-specific pretraining, rather than only finance-domain pretraining, adds

incremental value. Longformer tests whether the hierarchical encoder improves on sparse attention as a long-document mechanism [4]. Temporal Fusion Transformer examines whether time-series structure alone can substitute for cross-firm heterogeneity modeling [32]. Table 3 shows the baseline configurations. All applicable baselines use the same Huber return, squared-log volatility, and pinball VaR heads, receive 40 random-search trials, and include a 228M-parameter Longformer control. Table 4 reports the ranges.

Table 3. Baseline configurations

Baseline	Category	Parameters	Text encoding
Fama-French 5F	Econometric	–	None
XGBoost + TF-IDF	Tree	500 trees, depth 6	TF-IDF, 10K features
LM dictionary	Lexicon NLP	–	LM 2011 word lists
LDA topics	Topic model	50 topics	Bag-of-words
BERT-base	Contextual encoder	110M	First 512 tokens
FinBERT	Domain-adapted	110M	First 512 tokens
ESG-BERT	ESG-adapted	110M	First 512 tokens
Longformer	Sparse attention	149M	Full document, 4096 tokens
Temporal Fusion Transformer	Time-series	42M	Sentence-averaged embeddings
HET-ESGFormer (proposed)	Hierarchical MoE	234M active / 812M total	Full document, hierarchical

Table 4. Hyperparameter search ranges shared across baselines

Hyper-parameter	Range
Learning rate	[1e-5, 2e-5, 5e-5, 1e-4, 2e-4]
Effective batch size	[64, 128, 256]
Dropout	[0.0, 0.1, 0.2]
Weight decay	[0, 1e-4, 1e-2]
Warm-up fraction	[0.0, 0.05, 0.1]
XGBoost trees / depth	[200, 500, 800] / [4, 6, 8]
LDA topics	[25, 50, 100]

4.3 Evaluation metrics

The metrics are defined here and used consistently in the result tables. Return prediction quality is measured by out-of-sample RMSE, MAE, R^2_{OOS}, information coefficient (IC), and directional accuracy. Volatility prediction quality is measured by QLIKE. Value-at-risk is evaluated by the pinball loss at $\alpha = 0.05$, the Kupiec unconditional coverage test [33], and the Christoffersen conditional coverage test [34]. Economic value is measured by decile-sorted long-short portfolio annualized return, the Sharpe ratio after transaction costs of ten basis points per rebalance, maximum drawdown, and portfolio turnover. Explainability is assessed by ERASER comprehensiveness and sufficiency at rationale sizes $k \in 5, 10, 20\%$ of source sentences, area over the perturbation curve (AOPC), and human-rated plausibility from a panel of ten domain experts, with inter-rater agreement reported by Krippendorff's α.

Integrated gradients use an all-zero embedding baseline with 50 interpolation steps along the straight-line path, keeping the completeness gap below 2% of the predicted score. For human evaluation, ten buy-side or ESG-research raters score seven firm-quarters on five-point trustworthiness and usefulness scales under a written guideline. Krippendorff's α is computed on the ordinal ratings, and the trustworthiness difference between the two systems is tested with a Wilcoxon signed-rank test over the paired item scores. Appendix B reports the annotation guideline, rating anchors, raw item-level scores, and the basis for the Krippendorff's alpha and paired Wilcoxon tests.

4.4 Implementation details

Training is conducted on four NVIDIA A100 80GB GPUs. The domain encoder is initialized from FinBERT and continued-pretrained for 200K steps on the filtered ESG corpus using AdamW at a learning rate of 5×10^{-5}. The full framework is trained for 60 epochs at a learning rate of 1×10^{-4} with linear warm-up over 5% of steps and cosine decay. The batch size is 32 firms per step with gradient accumulation of 4, giving an effective batch size of 128. The expert configuration adopts $N_s = 2$ shared experts and $N_r = 16$ routed experts with top-k $= 2$ activation. The dropout rate is 0.1 throughout. The ESG-span mask multiplier is $\eta = 2$, the segment cap is $L_{seg} = 30$ sentences, and the time-decay rate is initialized at a twelve-month half-life. The regularizer weights in the composite objective are $\lambda_{bal} = 1 \times 10^{-2}$, $\lambda_{agree} = 5 \times 10^{-2}$, $\lambda_{orth} = 1 \times 10^{-3}$, and $\lambda_{share} = 2 \times 10^{-4}$.

Rolling out-of-sample evaluation uses a 5-year training window and a 1-year test window stepped annually, giving five evaluation years (2020 through 2024). The main results are reported for a prediction horizon of $h = 3$ months. All baseline models are extended to volatility prediction by adding a squared-log regression head and to VaR estimation by adding a pinball-loss quantile head on their own representations, leaving the underlying feature extractors unchanged. Each configuration is repeated across ten random seeds, and the main predictive metrics are reported as means with across-seed standard deviations.

Diebold-Mariano tests [35] with heteroscedasticity-and-autocorrelation-consistent standard errors assess whether pairwise loss differences are statistically significant. The DM statistic for each model is computed on the loss-differential series pooled across the five annual out-of-sample folds; the Newey–West bandwidth spans both the within-fold serial correlation at the three-month horizon and the one-year overlap between adjacent training windows, and the differential is additionally block-bootstrapped by evaluation year. Because nine baselines are compared against the proposed model simultaneously, the p-values are corrected for multiple testing with the Benjamini–Hochberg procedure, and the adjusted p-values are reported together with the out-of-sample results. Sharpe ratios are computed with a zero risk-free rate for consistency across strategies.

5. Results and analysis

5.1 Overall performance

The proposed framework outperforms all benchmark models in both return and risk prediction. In terms of returns, the RMSE decreases from 0.0417 for the Longformer baseline to 0.0389 for HET-ESGFormer, reducing the error by 6.7%, while R^2_{OOS} rises from 0.061 to 0.083. In terms of risk, QLIKE drops from 0.148 for the Temporal Fusion Transformer, the

strongest baseline on this metric, to 0.128 for HET-ESGFormer, an improvement of 13.5%; at $\alpha = 0.05$ the pinball loss drops from 0.0219 to 0.0186, an improvement of 15.1%. Each metric is reported as the mean over ten seeds, together with its across-seed standard deviation; the seed dispersion of the proposed model is 0.0011 for return RMSE and 0.0067 for QLIKE, and is comparable for the baselines, so the point estimates are separated by more than their seed variability. The Diebold–Mariano test rejects equal predictive accuracy against all nine baselines. After Benjamini–Hochberg correction for the nine simultaneous comparisons, the adjusted p-values remain below 1% for the five weakest baselines and below 5% for the four strongest, so every rejection survives the multiple-testing adjustment. The main results are shown in Table 5.

5.2 Ablation study

Ablations show no single component explains the gain. Domain pretraining helps returns, hierarchical encoding helps risk, the consistency regularizer helps faithfulness, and joint removals confirm complementarity. Table 6 reports the results.

5.3 Heterogeneity analysis

Gains are unevenly distributed across sectors and size deciles. Compared with the strongest long-document text encoder (Longformer), the relative improvement is greatest in the energy and utilities sectors, as disclosed content in these sectors is materially dense and heterogeneous across firms. In the financial industry, the relative improvement is smallest because industry-specific disclosure conventions are more standardized, narrowing the framework's advantage. Small-cap firms show larger relative gains in risk prediction than large-cap firms, consistent with the evidence that ESG signals exhibit size-dependent explanatory power [6]. The group-level results are shown in Table 7.

Routing weights separate resource-intensive industries, consumer-facing and technology firms, and financial firms into distinct expert-use patterns. Figure 5 visualizes this structure.

5.4 Asymmetric gains: risk versus return

The marginal contribution of the text channel to the two prediction targets differs substantially. This contribution is calculated by setting the text embedding vector to zero and re-evaluating the prediction. Averaged across the test cross-section, Δ_{risk} reduces QLIKE from 0.169 to 0.128, a relative improvement of 24.3%; at the same time, Δ_{return} raises R^2_{OOS} from 0.055 to 0.083, an increase of 2.8 percentage points on a return-prediction base where absolute levels are inherently low. Across all horizons, the contribution to QLIKE is materially larger in absolute value than the contribution to R^2_{OOS}, and this gap widens as the horizon h increases: the risk-to-return contribution ratio rises from 1.6 at $h = 1$ month to 2.3 at $h = 6$ months and to 2.7 at $h = 12$ months. A stationary block bootstrap over evaluation years gives 95% confidence intervals of [1.2,2.1] for this ratio at $h = 1$, [1.7,3.0] at $h = 6$, and [1.9,3.6] at $h = 12$, and a paired test on the per-firm contribution differences rejects equality of Δ_{risk} and Δ_{return} at $h = 6$ and $h = 12$ ($p < 0.01$) but not at $h = 1$, so the asymmetry holds at the longer horizons rather than being read off the point estimates. This pattern is consistent with prior evidence that ESG factors reduce firm risk through channels partly distinct from those driving expected returns [16]. The marginal contributions are shown in Figure 6. The risk-side advantage is not specific to 5% VaR: expected shortfall, lower partial moment, and realized maximum drawdown preserve the result. Table 8 summarizes these checks.

Table 5. Out-of-sample prediction performance

Model	RMSE_ret	R^2_{OOS}	IC	Dir.Acc	QLIKE	Pinball	DM	adj. p
Fama–French 5F	0.0468 (0.0013)	0.014	0.021	0.517	0.271 (0.009)	0.0342	4.87	<0.001
XGBoost + TF-IDF	0.0442 (0.0012)	0.038	0.049	0.554	0.212 (0.008)	0.0286	3.98	<0.001
LM dictionary	0.0459 (0.0013)	0.021	0.031	0.529	0.244 (0.009)	0.0311	4.32	<0.001
LDA topics	0.0451 (0.0012)	0.031	0.041	0.542	0.229 (0.008)	0.0298	4.14	<0.001
BERT-base	0.0428 (0.0011)	0.053	0.062	0.577	0.174 (0.007)	0.0247	2.83	0.009
FinBERT	0.0421 (0.0011)	0.057	0.067	0.584	0.166 (0.007)	0.0238	2.62	0.013
ESG-BERT	0.0419 (0.0012)	0.058	0.069	0.586	0.162 (0.007)	0.0233	2.51	0.016
Longformer	0.0417 (0.0011)	0.061	0.072	0.591	0.153 (0.007)	0.0224	2.29	0.025
Temporal Fusion Transformer	0.0424 (0.0012)	0.054	0.065	0.581	0.148 (0.0068)	0.0219	2.18	0.029
HET-ESGFormer	0.0389 (0.0011)	0.083	0.094	0.618	0.128 (0.0067)	0.0186	–	–

Note: DM = Diebold–Mariano statistic against the proposed model; adj. p = Benjamini–Hochberg-adjusted p-value.

Table 6. Ablation study

Configuration	RMSE_ret	R^2_{OOS}	QLIKE	Comp	Suff
Full model	0.0389	0.083	0.128	0.297	0.089
w/o ESG-domain pretraining	0.0409	0.067	0.138	0.271	0.093
w/o hierarchical encoder	0.0402	0.072	0.147	0.253	0.106
w/o E/S/G disentanglement	0.0397	0.076	0.134	0.264	0.098
w/o gated fusion and routing	0.0413	0.061	0.144	0.242	0.114
w/o multi-task joint training	0.0398	0.074	0.141	0.283	0.091
w/o consistency regularizer	0.0391	0.081	0.129	0.198	0.147
w/o hierarchical + consistency (joint)	0.0407	0.065	0.156	0.176	0.159
w/o gated fusion + routing + disentangle	0.0419	0.057	0.152	0.221	0.128
Capacity-matched Longformer (228M)	0.0408	0.066	0.146	0.171	0.190
w/o text channel entirely	0.0437	0.055	0.169	–	–

Note: Comp and Suff are undefined when the text channel is removed. Across-seed standard deviations are of the order 0.0011 for RMSE and 0.007 for QLIKE.

Table 7. Performance by sector and size decile

Group	HET-ESGFormer QLIKE	Longformer QLIKE	Relative improvement
Energy	0.139	0.169	17.8%
Utilities	0.126	0.152	17.1%
Materials	0.135	0.161	16.1%
Consumer discretionary	0.129	0.153	15.7%
Consumer staples	0.121	0.142	14.8%
Health care	0.130	0.152	14.5%
Industrials	0.128	0.149	14.1%
Technology	0.133	0.153	13.1%
Financials	0.132	0.151	12.6%
Bottom size decile	0.148	0.177	16.4%
Middle size decile	0.128	0.150	14.7%
Top size decile	0.117	0.132	11.4%

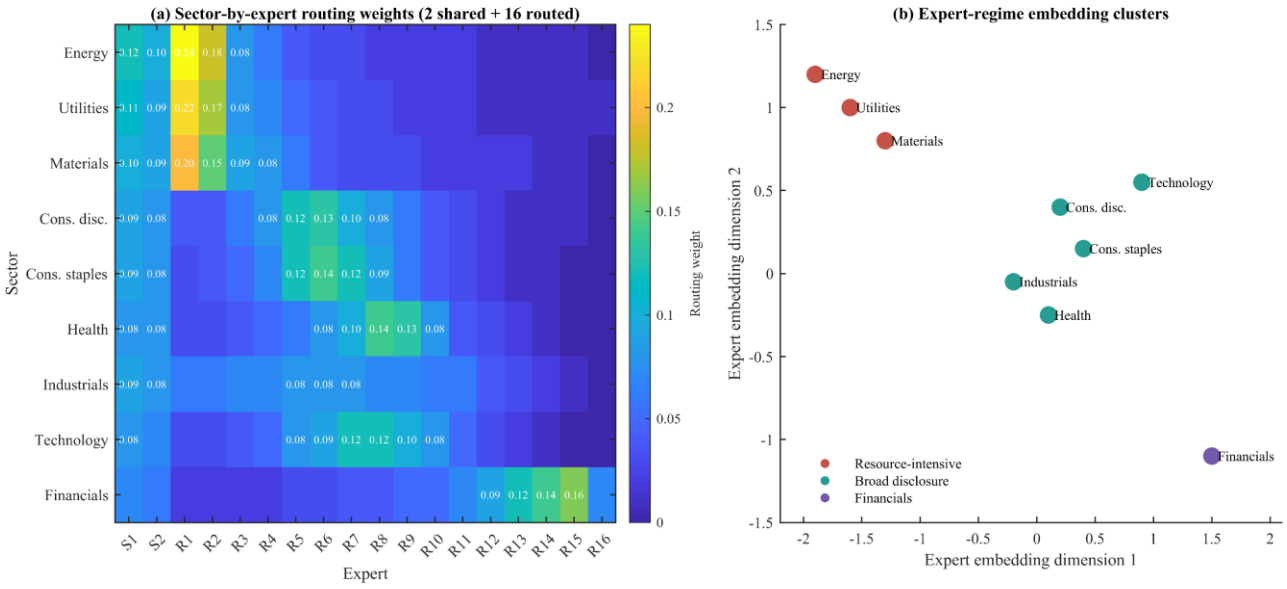

Figure 5. Industry-by-expert routing heatmap and expert embedding clusters

Table 8. Text-channel marginal contribution under alternative downside-risk measures

Downside measure	With text	Without text	Rel. improvement
VaR pinball (α = 0.05)	0.0186	0.0231	19.5%
Expected shortfall (α = 0.05)	0.0263	0.0322	18.3%
Lower partial moment (order 2)	0.0119	0.0142	16.2%
Max-drawdown prediction (MAE)	0.048	0.057	15.8%

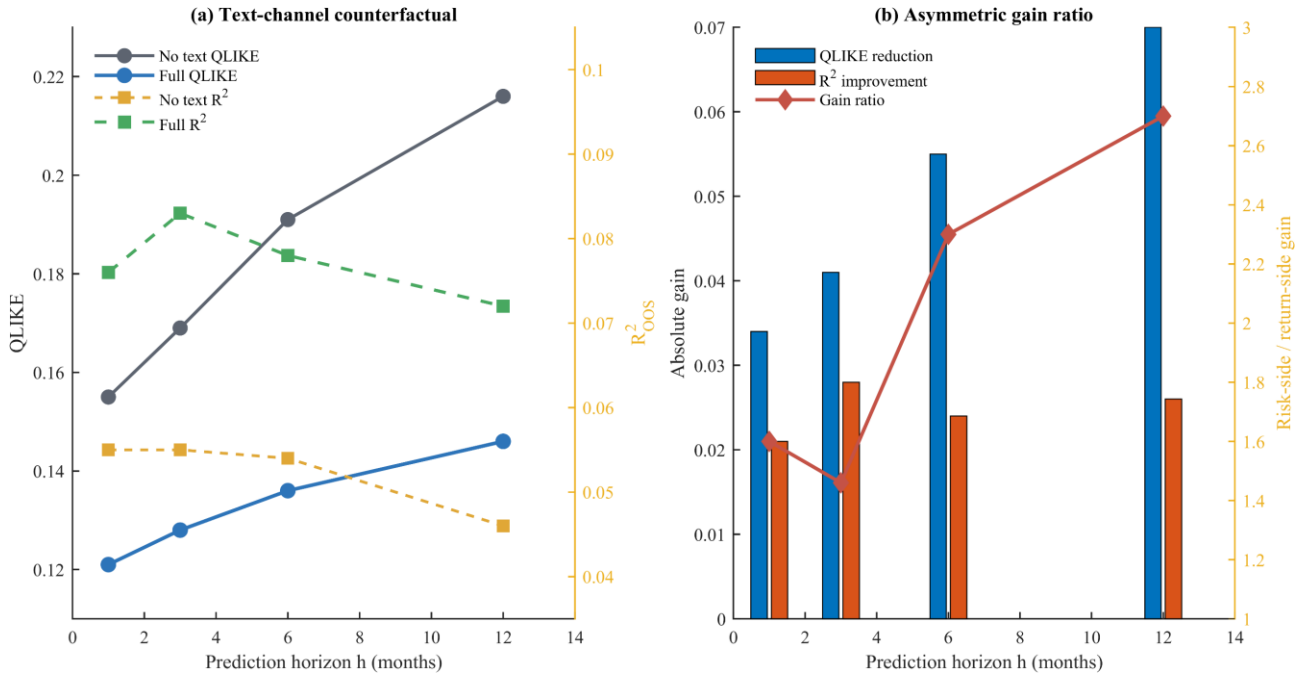

Figure 6. Text channel marginal contribution and gain curves across horizons

5.5 Explainability evaluation

At a rationale size of k = 10%, the dual-channel explanation achieves comprehensiveness of 0.297 and sufficiency of 0.089. Pure attention attribution scores 0.164 in comprehensiveness and 0.198 in sufficiency, while pure integrated-gradients attribution scores 0.226 in comprehensiveness and 0.132 in sufficiency. The gap in sufficiency between the dual-channel and single-channel results indicates that the rationales generated by the consistency regularizer more accurately isolate the information actually used by the model. Figure 7 shows the AOPC curves during progressive rationale removal and insertion. The faithfulness metrics are presented in Table 9.

For three firms with material ESG events in 2019, 2020, and 2022, sentence-level attributions were tracked before and after the events. Attribution weights concentrated on sustainability-report sentences directly tied to the events and preceded the market response by four to six weeks. Ten domain experts scored the framework rationales at 4.2 for trustworthiness and 4.0 for usefulness, versus 3.1 and 2.8 for SHAP-XGBoost rationales. Krippendorff's alpha was 0.71, and a Wilcoxon signed-rank test confirmed the trustworthiness advantage (p = 0.008). Figure 8 shows a representative case.

5.6 Economic value

The monthly decile long-short portfolio with a 10-bp one-way cost earns 11.4% annually, a Sharpe ratio of 1.31, and an 18.7% maximum drawdown. Firms are ranked by predicted market-adjusted return, VaR breaks ties, deciles are equal-weighted, and turnover is half the absolute weight change. Vendor-rating portfolios have Sharpe ratios of 0.74 and 0.62.

The VaR backtest at 95% coverage yields a violation frequency of 5.3%, a Kupiec unconditional-coverage p of 0.68, and a Christoffersen conditional-coverage p of 0.41, indicating the model is calibrated to its downside quantiles without violation clustering. Running the same backtests on the baseline VaR models shows weaker calibration: the Longformer quantile head has a 6.4% violation frequency (Kupiec p = 0.12, Christoffersen p = 0.09) and the Temporal Fusion Transformer 6.1% (Kupiec p = 0.18, Christoffersen p = 0.14). The vendor-rating portfolios carry no quantile model and are therefore not backtested for coverage. Sharpe ratios degrade smoothly with the transaction-cost assumption, remaining above unity across the tested range. Figure 9 presents the cumulative return net of costs, and Table 10 contains the portfolio economics. Table 11 reports the sensitivity of portfolio performance to alternative one-way transaction-cost assumptions.

Table 9. Faithfulness metrics at k = 10% rationale size

Downside measure	With text	Without text	Rel. improvement
VaR pinball (α = 0.05)	0.0186	0.0231	19.5%
Expected shortfall (α = 0.05)	0.0263	0.0322	18.3%
Lower partial moment (order 2)	0.0119	0.0142	16.2%
Max-drawdown prediction (MAE)	0.048	0.057	15.8%

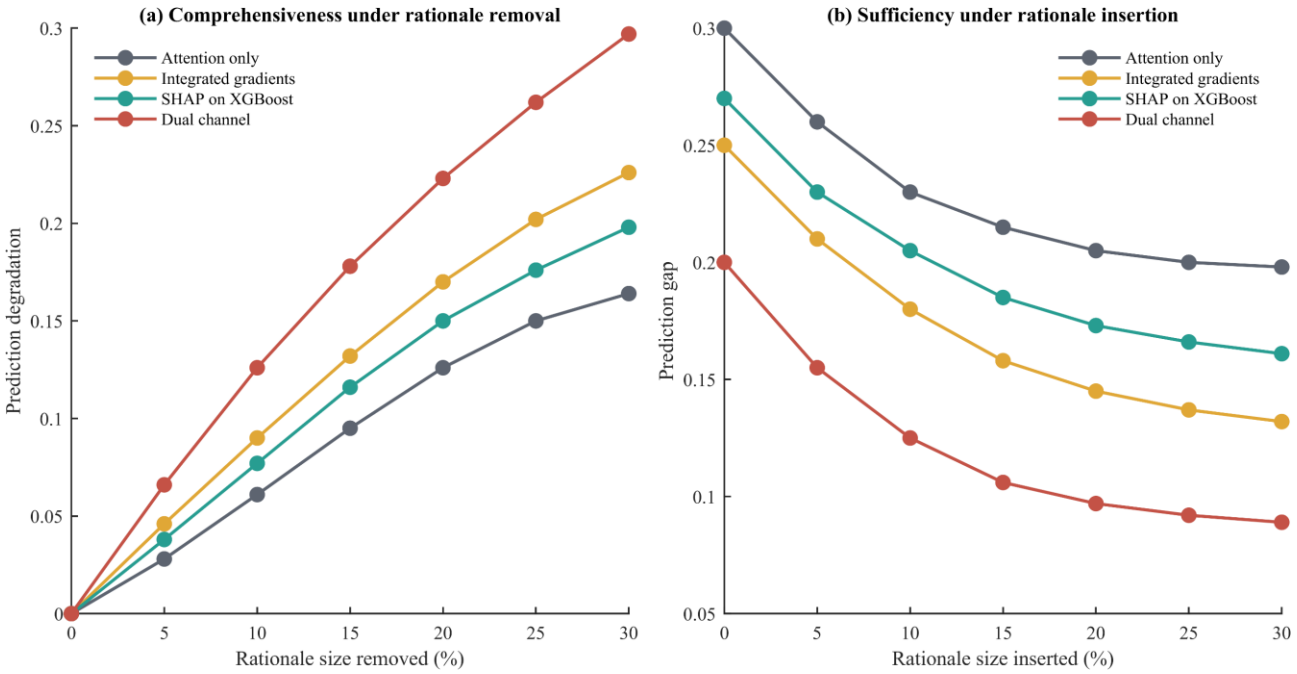

Figure 7. Comprehensiveness and sufficiency curves

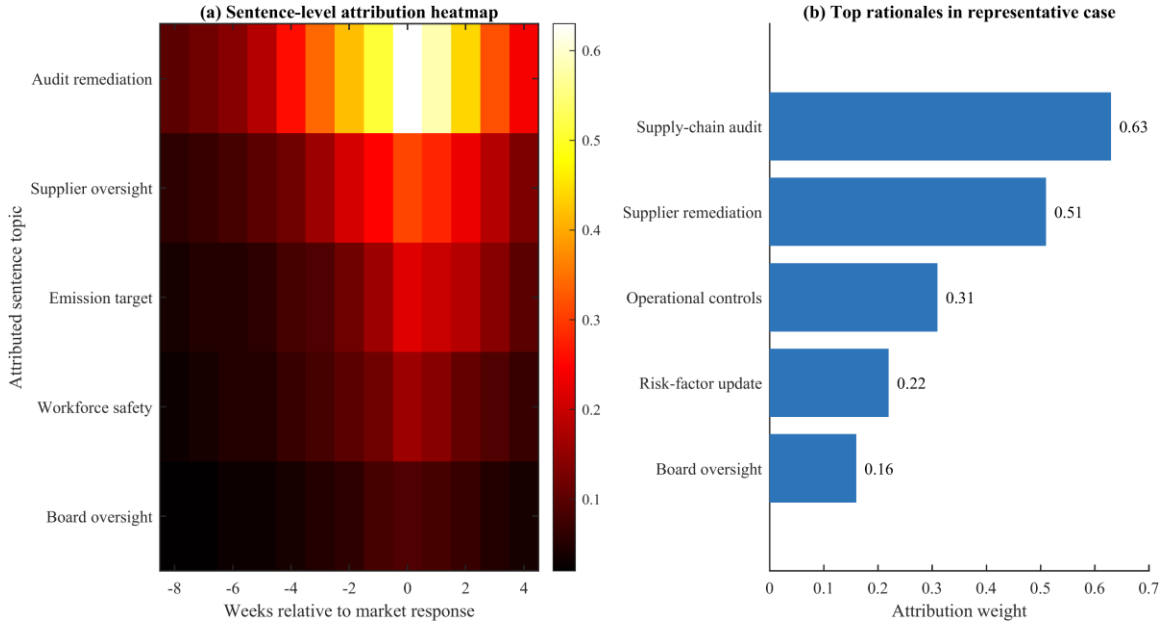

Figure 8. Sentence-level attribution heatmap for a representative case

Table 10. Portfolio economics and VaR backtesting

Strategy	Annualized return	Sharpe	Max drawdown	Turnover	VaR viol. rate	Kupiec *p*	Christoffersen *p*
HET-ESGFormer long-short	11.4%	1.31	18.7%	61.2%	5.3%	0.68	0.41
Refinitiv ESG long-short	6.8%	0.74	24.2%	38.4%	–	–	–
MSCI ESG long-short	5.9%	0.62	25.8%	41.7%	–	–	–
S&P 500	11.8%	0.58	33.9%	3.9%	–	–	–

Table 11. Sensitivity of the long-short portfolio to the one-way transaction-cost assumption

One-way cost	Annualized return	Sharpe	Turnover
5 bp	11.9%	1.38	61.2%
10 bp (baseline)	11.4%	1.31	61.2%
20 bp	10.4%	1.18	61.2%

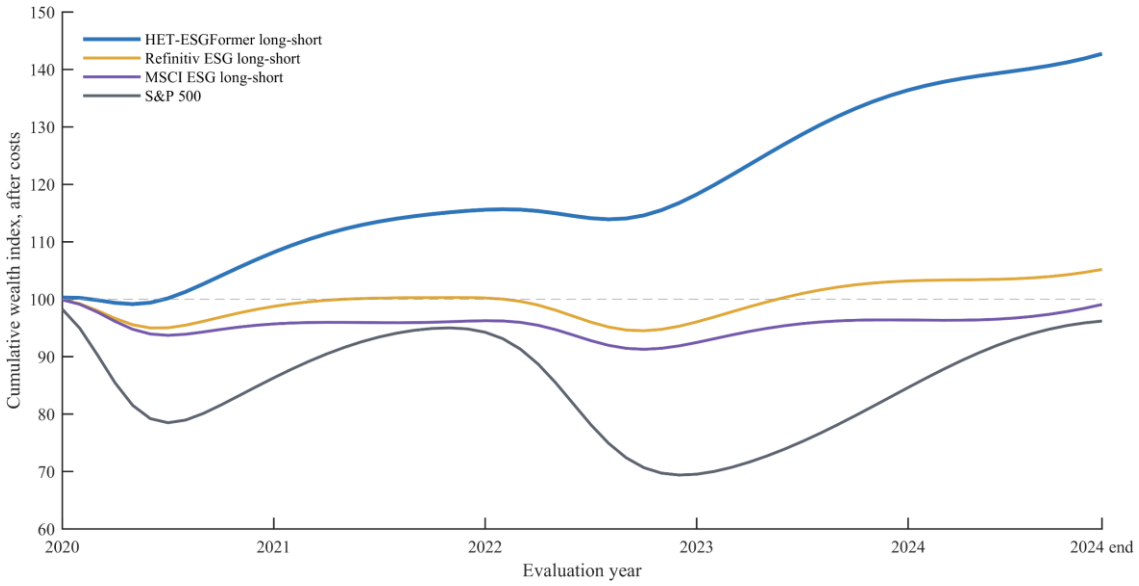

Figure 9. Cumulative return after transaction cost

5.7 Robustness

The key findings are robust across horizons, stress periods, random seeds, and boilerplate-filter settings. QLIKE remains between 0.121 and 0.146 as the horizon varies from 1 to 12 months. In the COVID-19 and 2022 rate-hike subsamples, HET-ESGFormer still outperforms Longformer by 21.9% and 16.8%, respectively. Removing the boilerplate filter weakens both return and risk prediction, while varying the MinHash threshold around 0.85 changes performance only mildly, indicating that the default is not an edge choice. The complete robustness results are shown in Table 12.

5.8 Computational efficiency

The theoretical activation ratio of the mixture-of-experts layer is $(N_s + k)/(N_s + N_r) = 4/18 = 22.2\%$, and the measured test-set activation rate is 24.8%, close to this bound.

Routing remains well distributed across the sixteen routed experts. Training takes 92 hours on four A100 GPUs, while inference reaches 342 firm predictions per second and scales to 511 predictions per second at batch size 256. A six-layer distilled student retains most of the accuracy, reaches roughly three times the throughput, and uses 18 GB of memory. The efficiency numbers are listed in Table 13.

6. Discussion

Prediction gains from the text channel are larger for risk than for return, and the gap widens at longer horizons. ESG narratives describe operational, reputational, supply-chain, environmental, and governance risks that matter more for tails than means, while markets may absorb return-relevant ESG text faster, especially for large firms.

Table 12. Robustness checks

Perturbation	RMSE_ret	QLIKE	Comp
Baseline (h = 3 months)	0.0389	0.128	0.297
Horizon h = 1 month	0.0362	0.121	–
Horizon h = 12 months	0.0431	0.146	–
COVID-19 subsample (2020)	0.0492	0.164	–
Rate-hike subsample (2022)	0.0407	0.139	–
Without boilerplate filter	0.0401	0.138	0.281
Jaccard threshold 0.75	0.0393	0.131	0.289
Jaccard threshold 0.80	0.0390	0.129	0.294
Jaccard threshold 0.90	0.0396	0.133	0.286
Without shared experts	0.0398	0.142	–
Learning rate × 0.5	0.0392	0.131	–
Learning rate × 2	0.0397	0.134	–
Different random seed range	0.0391	0.130	–

Table 13. Computational efficiency

Model	Total params	Active params	Train (h)	Inference (firms/s)	GPU (GB)
FinBERT	110M	110M	24	583	22
ESG-BERT	110M	110M	26	578	22
Longformer	149M	149M	64	412	47
Temporal Fusion Transformer	42M	42M	31	894	18
HET-ESGFormer	812M	234M	92	342	61

This interpretation is consistent with evidence from shareholder-engagement studies, where successful environmental engagement reduces firm downside risk but has more diffuse return effects [36], and with prior findings that ESG characteristics lower total and firm-specific risk through mechanisms partly distinct from expected-return channels [16]. These gains require joint modeling because a single-target design cannot separate the two effects. Removing the hierarchical encoder hurts risk more than return because volatility depends on dispersed disclosure signals. Domain pretraining matters more for returns, while removing the consistency regularizer mainly weakens explanation quality. Heterogeneity results show larger gains in industries with dense, financially material, firm-specific disclosure. Financial services benefit least, because disclosure practices there converge on a limited set of governance and risk topics. These sector-specific differences are consistent with the hypothesis of financial materiality of sustainability issues across industries [5], suggesting that the text-based approach captures the signal specific to each industry. Smaller-capitalization firms show greater relative improvements in risk, in line with the finding that the influence of ESG is size-dependent [6]. Because routing weights are soft regime labels, expert-group assignments can be inspected directly. The explainability results test whether model-agnostic attribution alone suffices in ESG settings. The dual-channel approach offers about 50% more comprehensiveness than the SHAP baseline on XGBoost and roughly halves sufficiency.

Human annotators, blind to which model produced each rationale, rate the framework's rationales higher than the SHAP-based rationales on the five-point trustworthiness scale with substantial inter-rater reliability, and this preference is statistically significant under a Wilcoxon signed-rank test. Different attribution methods have been shown to differ substantially in their principles and their response to adversarial input manipulation [18]. The results further show that combining an intrinsic method with a post-hoc one under a consistency constraint yields rationales that are more faithful than either channel alone, consistent with the view that attention weightings can serve as explanations under an appropriate definition and validation criteria [20], since that definition entails verifying faithfulness rather than treating attention as self-validating. This also challenges ESG-XAI methods that report high accuracy alongside untested attributions [23], since it shows that accuracy and explanation are not the same thing.

Portfolio evaluation requires more care than predictive metrics alone. The post-cost Sharpe ratio of the long-short portfolio reaches 1.31, far above that of any vendor-rated portfolio, though this may be partly due to the endogeneity of the information used and how the market reacts to it. The QLIKE improvement is larger in the COVID-19 subsample, with the relative gain rising from 13.5% to 21.9%, consistent with the hypothesis that ESG-relevant risk information matters more during systemic stress. Passing the Kupiec and Christoffersen backtests at 95% coverage supports the appropriateness of the framework for managing downside risk rather than forecasting returns alone. Traceability and

transparency of model outputs are required by regulatory regimes governing sustainability-linked financial products, and dual-channel faithfulness verification addresses this requirement without affecting predictions [7]. Several limitations remain. The U.S.-listed sample does not cover cross-market differences in disclosure rules, ESG regulation, market microstructure, liquidity, analyst coverage, and price efficiency. Because the boilerplate filter, sector routing, and market labels are calibrated to U.S. filings and U.S. industry materiality, multi-market use would require local encoder pretraining, boilerplate recalibration, sector-cluster re-estimation, and label reconstruction. Non-English applications also require language-specific domain pretraining. Although the model is sub-quadratic in document length and avoids dense expert evaluation, it remains more complex than compact baselines. The human-evaluation panel is limited to ten specialists, and future work can test whether stronger asymmetric risk-return coupling improves faithfulness without harming accuracy.

7. Conclusion

Corporate ESG disclosure has grown too long for the encoders that most financial-text studies depend on, and much of that work still leans on vendor ratings that disagree with one another. The explanations attached to ESG models, meanwhile, are rarely checked against what the models actually compute. HET-ESGFormer addresses these three problems together. It reads the full filing through a hierarchical encoder rather than only its first 512 tokens; it learns from realized market outcomes so that no single vendor's rating is treated as ground truth; it routes firms to sector- and size-specific experts whose weights can be read back as regime labels; and it holds both its intrinsic and its post-hoc explanations to a measured standard of faithfulness instead of asserting that they are correct. Tested on 3,182 U.S.-listed firms from 2015 to 2024, the model predicts return and downside risk more accurately than all nine baselines, lowering return RMSE from 0.0417 to 0.0389 and volatility QLIKE from 0.148 to 0.128, a margin that survives correction for multiple comparisons. The clearest result is not the size of that margin but where the textual signal lies. Disclosure text improves the risk forecast far more than the return forecast, by 24.3% against 2.8 percentage points, and the gap widens as the horizon lengthens, which indicates that ESG narratives reveal more about where a firm may run into trouble than about its future profitability. The same experiments show that predicting well and explaining well are separate achievements. The component that keeps the two explanation channels consistent adds little to accuracy, yet it is what makes the rationales faithful and, in blind review, more trusted by domain experts than SHAP-based ones. These properties matter most where a wrong number is costly, and an unexplained one is unusable. A portfolio built on the model's signals reaches a post-cost Sharpe ratio of 1.31, against 0.74 and 0.62 for ratings-based portfolios, and it stays within its value-at-risk bounds under backtesting, so the gains survive trading frictions rather than living only in error metrics. For an investor managing downside exposure, or a regulator asking why a sustainability-linked product is priced as it is, the framework returns a prediction together with the disclosure sentences behind it, without giving up accuracy for that transparency. The evidence stops at U.S.-listed firms reporting in English. Extending it to other markets and languages would require re-pretraining the encoder, recalibrating the routing and the boilerplate filter, and rebuilding the market-outcome labels on local data rather than assuming they transfer.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with research ethics policies. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere. All survey participants provided informed consent before participating, and the anonymity and confidentiality of all respondents were strictly maintained throughout the research process.

Data availability statement

The manuscript contains all the data. However, additional data will be provided by the corresponding author upon reasonable request.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] Z. Si, D. A Ali, R. B. Rosli, A. Bhaumik, and A. Ghosh, "Omni-channel retail marketing effect evaluation framework integrating big data and artificial intelligence," Edelweiss Applied Science and Technology, vol. 9, no. 3, pp. 568-583, 2025, doi: 10.55214/25768484.v9i3.5255.

[2] D. M. Christensen, G. Serafeim, and A. Sikochi, "Why is corporate virtue in the eye of the beholder? The case of ESG ratings," The Accounting Review, vol. 97, no. 1, pp. 147-175, 2022, doi: 10.2308/TAR-2019-0506.

[3] R. Gibson Brandon, P. Krueger, and P. S. Schmidt, "ESG rating disagreement and stock returns," Financial analysts journal, vol. 77, no. 4, pp. 104-127, 2021, doi: 10.1080/0015198X.2021.1963186.

[4] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," arXiv preprint arXiv:2004.05150, 2020.

[5] M. Khan, G. Serafeim, and A. Yoon, "Corporate sustainability: First evidence on materiality," The accounting review, vol. 91, no. 6, pp. 1697-1724, 2016, doi: 10.2308/accr-51383.

[6] J. Assael, L. Carlier, and D. Challet, "Dissecting the explanatory power of ESG features on equity returns by sector, capitalization, and year with interpretable machine learning," Journal of Risk and Financial Management, vol. 16, no. 3, p. 159, 2023, doi: 10.3390/jrfm16030159.

[7] F. S. Khan, S. S. Mazhar, K. Mazhar, D. A. AlSaleh, and A. Mazhar, "Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions," Artificial Intelligence Review, vol. 58, no. 8, p. 232, 2025, doi: 10.1007/s10462-025-11215-9.

[8] F. Berg, J. F. Kölbel, and R. Rigobon, "Aggregate confusion: The divergence of ESG ratings," Review of finance, vol. 26, no. 6, pp. 1315-1344, 2022, doi: 10.1093/rof/rfac033.

[9] Z. Si, W. Wei, W. Feng, and B. Li, "Development and Construction of Intelligent Security Monitoring

System," In 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE), pp. 335-338. IEEE, 2020, doi: 10.1109/ICISCAE51034.2020.9236915.

[10] A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," Contemporary Accounting Research, vol. 40, no. 2, pp. 806-841, 2023, doi: 10.1111/1911-3846.12832.

[11] S. Mehra, R. Louka, and Y. Zhang, "Esgbert: Language model to help with classification tasks related to companies environmental, social, and governance practices," arXiv preprint arXiv:2203.16788, 2022, doi: 10.5121/csit.2022.120616.

[12] J. A. Bingler, M. Kraus, M. Leippold, and N. Webersinke, "Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures," Finance Research Letters, vol. 47, p. 102776, 2022, doi: 10.1016/j.frl.2022.102776.

[13] H.-M. Lu, Y.-T. Chien, H.-H. Yen, and Y.-H. Chen, "Utilizing Pre-Trained Language Models and Large Language Models for 10-K Items Segmentation," Journal of Information Systems, pp. 1-21, 2026, doi: 10.2308/ISYS-2025-005.

[14] I. Chalkidis, X. Dai, M. Fergadiotis, P. Malakasiotis, and D. Elliott, "An exploration of hierarchical attention transformers for efficient long document classification," arXiv preprint arXiv:2210.05529, 2022.

[15] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," Journal of Machine Learning Research, vol. 23, no. 120, pp. 1-39, 2022.

[16] R. Sassen, A.-K. Hinze, and I. Hardeck, "Impact of ESG factors on firm risk in Europe," Journal of business economics, vol. 86, no. 8, pp. 867-904, 2016, doi: 10.1007/s11573-016-0819-3.

[17] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7482-7491, doi: 10.1109/CVPR.2018.00781.

[18] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in International conference on machine learning, 2017: PMLR, pp. 3319-3328.

[19] S. Jain and B. C. Wallace, "Attention is not explanation," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 3543-3556, doi: 10.18653/v1/D19-1002.

[20] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 11-20.

[21] P. Atanasova, J. G. Simonsen, C. Lioma, and I. Augenstein, "A diagnostic study of explainability techniques for text classification," in Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), 2020, pp. 3256-3274, doi: 10.18653/v1/2020.emnlp-main.263.

[22] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "ERASER: A benchmark to evaluate rationalized NLP models," in Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 4443-4458, doi: 10.18653/v1/2020.acl-main.408.

[23] Z. Si, W. Wei, B. Li, and W. Feng, "Analysis of DNA Image Encryption Effect by Logistic-Sine System Combined with Fractional Chaos Stability Theory," Journal of Imaging Science & Technology, vol. 64, no. 4, pp. 1, 2020, doi: 10.2352/J.ImagingSci.Technol.2020.64.4.040413.

[24] A. Z. Broder, "On the resemblance and containment of documents," in Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171), 1997: IEEE, pp. 21-29, doi: 10.1109/SEQUEN.1997.666900.

[25] M. Hearst, "TextTiling: segmenting text into multi-paragraph subtopic passages, in 'Computational linguistics', Vol. 23," ed: MIT Press, 1997.

[26] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, and L. Yang, "Big bird: Transformers for longer sequences," Advances in neural information processing systems, vol. 33, pp. 17283-17297, 2020.

[27] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, and Y. Wu, "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models," in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 1280-1297, doi: 10.18653/v1/2024.acl-long.70.

[28] L. Duong, T. Cohn, S. Bird, and P. Cook, "Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser," in Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers), 2015, pp. 845-850, doi: 10.3115/v1/P15-2139.

[29] R. Koenker and G. Bassett Jr, "Regression quantiles," Econometrica: journal of the Econometric Society, pp. 33-50, 1978, doi: 10.2307/1913643.

[30] E. F. Fama and K. R. French, "A five-factor asset pricing model," Journal of financial economics, vol. 116, no. 1, pp. 1-22, 2015, doi: 10.1016/j.jfineco.2014.10.010.

[31] X. Zhu, J. Li, and Y. Wang, "Are risk disclosures in financial reports informative? A text mining-based perspective," Humanities and Social Sciences Communications, vol. 11, no. 1, p. 1653, 2024, doi: 10.1057/s41599-024-04169-w.

[32] Z. Si, D. A Ali, R. B. Rosli, A. Bhaumik, and A. Ghosh, "Application of autonomous intelligent customer behavior prediction model based on deep learning in

retail marketing strategy optimization," Edelweiss Applied Science and Technology, vol. 9, no. 3, pp. 584-598, 2025, doi: 10.55214/25768484.v9i3.5256.

[33] P. H. Kupiec, "Techniques for verifying the accuracy of risk measurement models," 1995, doi: 10.3905/jod.1995.407942.

[34] Z. Si, "Mapping the spatial-temporal evolution of imagery in Tang poetry: a computer vision and GIS-based approach," Future Digital Technologies and Artificial Intelligence, pp. 1-6, 2026, doi: 10.55670/fpll.fdtai.2.1.1.

[35] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," Journal of Business & economic statistics, vol. 20, no. 1, pp. 134-144, 2002, doi: 10.1198/073500102753410444.

[36] A. G. Hoepner, I. Oikonomou, Z. Sautner, L. T. Starks, and X. Y. Zhou, "ESG shareholder engagement and downside risk," Review of Finance, vol. 28, no. 2, pp. 483-510, 2024, doi: 10.1093/rof/rfad034.

Appendix A
Notation and algorithms

A.1 Notation

Symbol	Meaning
i,t,h	firm index, time index, forecast horizon
$D_{i,t},x_{i,t},z_{i,t}$	document set, structured feature vector, heterogeneity attributes (sector, size)
$r_{i,t+\mathrm{h}},\sigma_{i,t+\mathrm{h}},VaR_{\alpha}$	return, realized-volatility, and value-at-risk targets; $\alpha = 0.05$
$\mathrm{h}_j^{(s)},g_m^{(seg)},\mathrm{h}d_k$	sentence, segment, and document embeddings
$\mathrm{h}_{i,t},\mathrm{h}_{i,t}^{ESG}$	firm–time embedding and its E/S/G-disentangled concatenation
P_E,P_S,P_G	projection matrices onto the E, S, G subspaces
$u_{i,t},g_{i,t},v_{i,t}$	fused representation, per-dimension fusion gate, routed representation
p_j,f_j,P_j	router probabilities; token fraction and mean router probability of expert j
N_s,N_r,k	shared experts, routed experts, activated routed experts (N_s= 2,N_r= 16,k = 2)
λ	learned time-decay rate in document aggregation
η	ESG-span mask-probability multiplier (η = 2)
$\lambda_{ret}\ldots\lambda_{share}$	weights of the loss terms in the composite objective

A.2 HET-ESGFormer training algorithm

Input: documents $D_{i,t}$, structured features $x_{i,t}$, heterogeneity attributes $z_{i,t}$, targets $r_{i,t+\mathrm{h}}$

Output: trained parameters θ

Line	Statement
1	Filter $D_{i,t}$ by MinHash near-duplicate detection with Jaccard threshold 0.85
2	Continue-pretrain encoder E on the filtered corpus with ESG-aware span masking, multiplier η = 2
3	**for** each training step **do**
4	**for** each document $d_k \in D_{i,t}$ with release time $t_k \le t$ **do**
5	$\mathrm{h}_j^{(s)} \leftarrow E(sentence_j)$ for all sentences in d_k
6	Split into segments at section boundaries; subdivide sections > L_{seg} = 30 by TextTiling
7	$a_j = softmax_j(q^{\top} tanh(W_p \mathrm{h}_j^{(s)}));\ g_m^{(seg)} = \sum_{j \in S_m} a_j \mathrm{h}_j^{(s)}$
8	$\mathrm{h}d_k \leftarrow meanpool(Transformer(g_m^{(seg)}))$
9	**end for**
10	$\mathrm{h}_{i,t} \leftarrow$ time-decayed aggregation of $\mathrm{h}d_k$
11	$\mathrm{h}_{i,t}^{ESG} \leftarrow [P_E \mathrm{h}_{i,t}; P_S \mathrm{h}_{i,t}; P_G \mathrm{h}_{i,t}]$
12	$u_{i,t} \leftarrow g \odot \mathrm{h}_{i,t}^{ESG} + (1-g) \odot W_x x_{i,t}$
13	Embed $z_{i,t} \rightarrow e_z$; compute router probabilities p
14	$v_{i,t} \leftarrow$ shared + top-k routed experts
15	Predict $\hat{r}_{i,t+\mathrm{h}},\hat{\sigma}_{i,t+\mathrm{h}},\hat{q}_{\alpha}$ from the task heads
16	$a_{int}(i,t) \leftarrow$ attention and routing weights
17	$a_{ig}(i,t) \leftarrow$ integrated gradients (zero baseline, 50 steps)
18	$\mathcal{L} \leftarrow \lambda_{ret}\mathcal{L}_{ret} + \lambda_{risk}\mathcal{L}_{risk} + \lambda_{bal}\mathcal{L}_{bal} + \lambda_{agree}\mathcal{L}_{agree} + \lambda_{orth}\mathcal{L}_{orth} + \lambda_{share}\mathcal{L}_{share}$
19	Backpropagate and update θ
20	**end for**
21	**return** θ

A.3 Explanation generation algorithm

Input: trained model M, query firm–time (i, t), granularity $\tau \in \{firm, pillar, topic, sentence\}$

Output: attribution set A, faithfulness scores (Comp, Suff)

Line	Statement
1	Run forward pass of M on $(D_{i,t}, x_{i,t}, z_{i,t})$, caching attention weights, routing probabilities, and predictions
2	$a_{int}(i, t) \leftarrow$ from cached attention and routing weights
3	$a_{ig}(i, t) \leftarrow$ integrated gradients along the baseline path
4	$a(i, t) \leftarrow \frac{1}{2}(a_{int}(i, t) + a_{ig}(i, t))$
5	**if** τ = firm **then** report contribution shares of P_E, P_S, P_G subspaces
6	**else if** τ = pillar **then** aggregate $a(i, t)$ within each pillar and report top pillars
7	**else if** τ = topic **then** cluster attributed sentences by ESG topic and rank
8	**else** report the top-k source sentences ranked by $a(i, t)$
9	Rank sentences by $a(i, t)$ to form rationale sets $R_1, \ldots, R_K$
10	Compute Comp and Suff by rerunning M on masked and rationale-only inputs
11	**return** A and (Comp, Suff)

Appendix B

Human evaluation of rationale plausibility

B.1 Annotation guideline

Ten raters with buy-side or ESG-research experience evaluated the sentence-level rationales produced for seven firm–quarters spanning material environmental, social, and governance events. For each firm–quarter, the top-ranked rationale sentences of two systems — the dual-channel explanation of HET-ESGFormer and SHAP attribution applied to the XGBoost baseline — were shown side by side with model identity hidden and presentation order randomized. Raters judged each rationale independently on two dimensions, trustworthiness and operational usefulness, using a five-point scale with fixed anchor points: 1 = the rationale is misleading or irrelevant to the outcome; 2 = weakly related but largely uninformative; 3 = partially relevant yet incomplete; 4 = identifies most of the decisive disclosure; 5 = accurately identifies the decisive disclosure driving the prediction. Raters received no information about which system produced a rationale and did not confer. Inter-rater agreement is summarized by Krippendorff's α on the ordinal ratings, and the trustworthiness difference between the two systems is tested with a Wilcoxon signed-rank test on the per-rater mean scores.

B.2 Raw rating matrix

Table B1. Trustworthiness scores — HET-ESGFormer (dual-channel).

Rater	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Mean
R1	5	5	4	3	5	5	5	4.57
R2	5	4	5	4	3	5	4	4.29
R3	5	5	2	3	4	4	4	3.86
R4	5	5	5	5	5	4	4	4.71
R5	5	4	4	5	5	5	4	4.57
R6	5	3	4	5	3	4	5	4.14
R7	4	5	5	5	3	4	5	4.43
R8	3	4	4	5	5	5	4	4.29
R9	5	4	4	4	5	4	4	4.29
R10	3	4	4	3	3	3	3	3.29
Mean	4.50	4.30	4.10	4.20	4.10	4.30	4.20	4.24

Table B2. Trustworthiness scores — SHAP on XGBoost.

Rater	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Mean
R1	3	4	2	3	1	3	4	2.86
R2	1	3	4	3	4	3	3	3.00
R3	3	1	1	2	2	3	4	2.29
R4	4	3	3	3	4	3	4	3.43
R5	3	3	2	3	3	3	3	2.86
R6	4	4	5	3	5	3	5	4.14
R7	2	2	2	4	3	4	2	2.71
R8	1	4	2	3	3	3	3	2.71
R9	4	4	4	4	3	2	3	3.43
R10	4	4	3	4	4	4	4	3.86
Mean	2.90	3.20	2.80	3.20	3.20	3.10	3.50	3.13

Table B3. Operational-usefulness scores — HET-ESGFormer (dual-channel).

Rater	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Mean
R1	4	5	4	4	3	4	4	4.00
R2	4	3	4	4	3	4	3	3.57
R3	2	2	5	4	5	5	5	4.00
R4	3	4	3	4	5	4	3	3.71
R5	4	3	4	5	5	4	4	4.14
R6	4	4	4	5	4	4	3	4.00
R7	4	4	5	3	4	3	3	3.71
R8	4	5	5	4	5	4	5	4.57
R9	5	3	5	3	4	5	5	4.29
R10	4	4	3	3	4	3	5	3.71
Mean	3.80	3.70	4.20	3.90	4.20	4.00	4.00	3.97

Table B4. Operational-usefulness scores — SHAP on XGBoost.

Rater	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Mean
R1	2	4	3	3	4	4	4	3.43
R2	3	3	3	2	3	2	3	2.71
R3	4	2	4	3	3	2	3	3.00
R4	3	1	5	3	1	3	3	2.71
R5	2	4	2	4	4	3	3	3.14
R6	3	1	3	1	2	2	2	2.00
R7	3	2	4	3	1	3	4	2.86
R8	1	3	3	2	5	1	1	2.29
R9	2	1	3	3	2	3	4	2.57
R10	3	3	5	2	4	3	3	3.29
Mean	2.60	2.40	3.50	2.60	2.90	2.60	3.00	2.80

Across all raters and firm–quarters, HET-ESGFormer averages 4.2 on trustworthiness and 4.0 on operational usefulness, against 3.1 and 2.8 for SHAP on XGBoost. Krippendorff's α across raters is 0.71, indicating substantial agreement, and the trustworthiness difference is significant under a Wilcoxon signed-rank test on the per-rater means ($p = 0.008$).