SEMANTIC SEARCH FOR FIRM-LEVEL EMERGING RISK MEASUREMENT: EVIDENCE FROM CLIMATE RISK IN EARNINGS CALLS

Marleen de Jonge
Ana Mićković
Stevan Rudinac
Sweder van Wijnbergen

Working Paper

February 3, 2026

ABSTRACT

Measuring firm-level exposure to emerging risks from forward-looking disclosures requires methods that can capture evolving, multidimensional concepts from unstructured text. We develop a semantic search framework that measures firm-level attention to climate risks and opportunities from earnings call transcripts, overcoming a fundamental trade-off between scalability and precision that limited earlier approaches. Applied to S&P 500 and STOXX 600 firms (2009–2025), the framework produces granular measures of firm-level climate exposure with substantial time-series and cross-sectional variation. These measures predict real outcomes: firms emphasizing climate opportunities file 26% more green patents and convert R&D spending into green innovation more efficiently, while transition-risk exposure intensifies the market penalty for carbon emissions after the Paris Agreement. The approach provides a general, scalable tool for narrative-based measurement in economics and finance.

Keywords: Text as Data, Semantic Search, Representation Learning, Climate Finance, Firm-level Risk, Earnings Calls, TCFD, Economic Measurement

Acknowledgements

We acknowledge financial support from the Research Centre for Sustainable Investments and Insurance (RCSII) and the University of Amsterdam. We declare no conflicts of interest relevant to this research. All remaining errors are our own.

I Introduction

Understanding how individual firms are exposed to various sources of risk is a core input to asset pricing and corporate finance. A growing literature has demonstrated that firms' own disclosures—particularly in earnings calls and forward-looking statements—provide valuable, high-frequency information about perceived risk exposure that traditional backward-looking metrics cannot capture (Sautner et al., 2023; Hassan et al., 2025; Gentzkow et al., 2019). However, these text-based approaches have primarily focused on measuring aggregate exposure to broad risk categories, leaving fundamental questions about the underlying economic mechanisms unanswered. For instance, within climate transition risk, researchers cannot distinguish whether market reactions stem from policy uncertainty, technological substitution threats, or changing consumer preferences—each of which operates through different channels and requires distinct strategic responses.

This limitation stems from a core methodological constraint: existing approaches can-

not reliably decompose broad risk categories into their constituent economic channels while maintaining scalability across different risks, taxonomies, and time periods. Scalable methods like keyword searches cannot reliably distinguish between different channels within broad risk categories or adapt as terminology evolves, while more nuanced supervised approaches require costly retraining for each new application and cannot generalize across different risk taxonomies or time periods. Moreover, the limited contextual precision of existing scalable methods constrains their ability to capture the nuanced, firm-specific variation in risk narratives that is essential for studying belief updating, information shocks, and cross-sectional heterogeneity in risk perceptions.

In this paper, we propose a measurement framework for firm-level emerging risk exposure that overcomes the methodological limitations discussed above. Following the text-as-data literature (Gentzkow et al., 2019; Hassan et al., 2025; Sautner et al., 2023), we construct our measure from quarterly earnings call transcripts, adopting the core assumption that the share of a firm's discourse devoted to a topic proxies for its perceived exposure. Our key innovation is to replace traditional keyword-based retrieval with semantic search, which identifies relevant text based on conceptual similarity rather than lexical matching. The method works by encoding language into high-dimensional semantic vectors that place related phrases close together in a common vector space, regardless of the specific words used. This allows our framework to identify discussions of an economic concept by calculating the proximity between researcher-defined queries and text passages in this shared space.

Crucially, the approach requires no task-specific labeled data or model training, making it instantly adaptable to new concepts and taxonomies. While direct LLM classification of large corpora escalates into high financial and computational costs, our semantic search processes millions of paragraphs within seconds, making exploratory analysis and iterative text-based research on emerging risks feasible.

Further, by dramatically reducing the search space from millions to thousands of passages, this first stage makes it economically viable to apply more precise but expensive classification methods to the filtered subset. We therefore employ an optional second-stage LLM classifier that refines the retrieved passages when maximum precision is required. The modular two-stage architecture allows researchers to choose their preferred precision-cost trade-off: semantic search alone provides comprehensive coverage at minimal cost, while the

LLM refinement layer delivers maximum accuracy for applications that demand it.

While the methodology is general and can be adapted to other types of risks, we focus on climate risk as our primary application because it provides an ideal proving ground for the framework. Climate discourse is economically salient, features rapidly evolving and context-dependent terminology that challenges keyword approaches, and encompasses both risks and opportunities that require nuanced distinction. Moreover, constructing improved, high-frequency measures of firm-level climate exposure opens up important empirical applications in asset pricing and corporate finance, from testing climate risk pricing hypotheses to studying how narratives translate into actual investment decisions.

Concretely, we identify any discussion of climate change related business impacts based on the comprehensive Task Force on Climate related Financial Disclosures (TCFD) taxonomy. This framework organizes climate exposure into three overarching pillars—physical risks, transition risks, and climate-related opportunities—each containing a detailed taxonomy of specific impact channels (e.g., policy changes, technology shifts, acute physical events). For each granular subcategory, we design multiple semantic query variants that capture both the risk/opportunity source and its potential business impacts, yielding approximately 340 distinct queries across all TCFD categories. These queries enable measurement from broad transition risks down to specific mechanisms like carbon pricing policies, stranded asset exposure, and regulatory compliance costs. Applying this full set of queries to a global corpus of earnings call transcripts (S&P 500 and STOXX 600 firms, 2009–2025), we construct firm-quarter level attention scores for each specific channel, then aggregate these granular channels to obtain three composite scores: transition risk exposure, physical risk exposure, and climate-related opportunities.

We systematically assess performance through manual annotation and weighted performance metrics, benchmarking against state-of-the-art alternatives including ClimateBERT. Our two-stage method achieves substantially higher precision than ClimateBERT (0.93 vs. 0.47) while maintaining competitive recall, translating to superior overall performance (F1-score 0.86 vs. 0.58). Notably, even our semantic search layer alone outperforms ClimateBERT despite requiring no domain-specific training data.

Second, we establish that the resulting firm-quarter attention scores behave as credible, economically meaningful proxies for climate risk exposure. We show they possess substantial

firm-level variance orthogonal to sector or aggregate trends, exhibit a sparse, right-skewed distribution consistent with the concentration of material climate exposure, and align in the cross-section with observable indicators like carbon intensity and sector trends. Furthermore, we validate their timeliness by showing that unexpected, firm-level innovations in attention correspond clearly to identifiable climate-related news events.

Finally, we demonstrate the framework's practical value through two key applications: showing that climate attention predicts firms' future green patent filings and R&D investment, and revealing how narrative-based climate exposure conditions market responses to carbon emissions after major policy shifts like the Paris Agreement.

First, we show that our firm-level attention scores for climate-related opportunities and transition risks significantly predict a firm's future green patent filings and R&D investment intensity, even after controlling for sector, size, profitability, and existing green patents. Moreover, climate attention amplifies the productivity of R&D spending: firms emphasizing climate opportunities or transition risks convert R&D expenditures into green patents more efficiently than firms with similar R&D intensity but lower climate attention. This nonlinear relationship indicates that increased managerial attention to climate risks and opportunities, as captured in earnings calls, translates into both the scale and direction of innovative investment, suggesting that such discourse reflects genuine strategic reorientation rather than mere cheap talk.

Second, we employ the framework to study market pricing around a major structural shift: the Paris Agreement. Using our transition risk and physical risk scores to condition firm-level exposure, we find that the market penalty for high carbon intensity was significantly stronger for firms that were also perceived by investors as facing high transition risk. Specifically, the interaction of high emissions with high transition-risk attention is associated with a more negative post-agreement return differential than for high-emission firms with low perceived risk exposure. This result underscores how our measures can isolate the role of narrative and perceived exposure in asset pricing, going beyond static, accounting-based metrics.

This paper makes three main contributions. First, we introduce a scalable semantic search framework that enables economists to measure nuanced concepts from unstructured text without requiring labeled training data or model retraining. By replacing static keyword lists with dynamic semantic queries, the approach maintains the scalability of dictionary

methods while achieving the contextual accuracy of supervised classifiers. The two-stage architecture—combining fast semantic retrieval with optional LLM refinement—allows researchers to choose their preferred precision-cost trade-off while remaining instantly adaptable to new risk taxonomies as they emerge.

Second, we construct and validate granular, high-frequency measures of firm-level climate risk exposure that capture economically meaningful variation. Applied to the TCFD taxonomy and derived from earnings call transcripts covering S&P 500 and STOXX 600 firms (2009–2025), our measures exhibit substantial firm-specific variance, align with observable proxies like carbon intensity, and respond dynamically to climate news events. The validated dataset provides quarterly attention scores not just for broad categories, but for specific channels like carbon pricing mechanisms and technology substitution threats.

Third, we demonstrate how narrative-based measures enable new empirical research designs by isolating firm-level information shocks and testing specific economic channels. Our climate attention measures predict real outcomes—firms emphasizing opportunities file 26% more green patents and convert R&D spending more efficiently into green innovation. In asset pricing, we show that narrative-based climate exposure amplified the market penalty for carbon emissions after the Paris Agreement, a result that emerges only when conditioning on perceived risk rather than static firm characteristics. These applications illustrate how the framework can distinguish belief updating from baseline exposure in corporate decisions and market pricing.

The remainder of the paper proceeds as follows. Section II situates our work within the relevant literature. Section III describes the earnings call transcript data. Section IV details the semantic search architecture and the optional two-stage classification pipeline. Section VI reports benchmarking results and validates the economic properties of the constructed measures. Section VII demonstrates the measures' utility in forecasting green investment and asset pricing applications. Section VIII concludes.

II Related Literature

A. Text as Data in Economic Analysis

Economic decision-making fundamentally depends on forward-looking beliefs about risk,

opportunity, and uncertainty. Yet these beliefs—which drive investment, pricing, and strategic choices—are difficult to measure using conventional economic data. Surveys capture expectations infrequently and suffer from reporting biases (Manski, 2018). Market prices aggregate beliefs but obscure the underlying narratives, information updates, and reasoning processes. Administrative and financial data reveal realized outcomes but miss the evolving assessments that precede and shape those decisions.

A growing literature addresses this measurement gap by treating text as economic data (Gentzkow et al., 2019; Hassan et al., 2025). This approach exploits a key insight: language encodes information about beliefs, expectations, and perceived risks in real time as agents articulate their assessments through corporate disclosures, policy communications, and media coverage. Technological advances—the digitization of communication and developments in natural language processing—have made systematic analysis of these textual records both feasible and scalable.

This text-as-data approach has generated influential measures across multiple domains. Baker et al. (2016) construct a widely used index of economic policy uncertainty from newspaper coverage, showing that textual signals predict investment and output volatility. Hassan et al. (2019) develop firm-level political risk measures from earnings calls, demonstrating that narrative-based exposure varies systematically across firms and correlates with real outcomes such as hiring and capital expenditure. Sautner et al. (2023) extend this framework to climate risk, distinguishing regulatory, physical, and opportunity exposures from corporate disclosures. Shapiro et al. (2022) extract sentiment from news to study its macroeconomic effects. These applications—alongside related studies on central bank communication (Silva et al., 2025), social media sentiment (Cookson et al., 2023), and narrative-driven macroeconomic fluctuations (Shiller, 2020)—demonstrate that text-derived measures carry substantial explanatory and predictive power for economic outcomes.

However, this success has been achieved despite significant methodological constraints. To identify topic-relevant content, existing studies predominantly rely on keyword dictionaries or handcrafted rules that search for predetermined terms or phrases. These dictionary-based approaches are rigid, sensitive to vocabulary choice, and unable to capture semantic meaning beyond exact word matches. This creates a bottleneck for measuring emerging or rapidly evolving concepts—precisely the domains where forward-looking textual measures

offer the greatest potential value. The challenge is particularly acute for risks like climate change, technological disruption, or geopolitical shifts, where terminology evolves faster than dictionary-based approaches can adapt.

B. Methods for extracting economic information from text

The central methodological challenge in text-as-data lies in accurately identifying topic-relevant content within large, unstructured corpora. Some approaches sidestep this challenge by using unsupervised methods like topic modeling to discover latent themes bottom-up, but these are typically suited for exploratory analysis rather than the precise measurement needs of downstream economic applications. For most text-based economic research, researchers must solve the fundamental problem of distinguishing passages that discuss a specific concept of interest from the vast majority of text that does not. This topic identification step is critical because it determines both the accuracy and scope of any subsequent textual measure.

While approaches to this identification problem have expanded, existing methods involve sharp trade-offs between transparency, flexibility, and scalability. The most established approach, widely used since the early text-as-data literature (Baker et al., 2016; Hassan et al., 2019; Sautner et al., 2023), relies on dictionaries or keyword-based rules to classify sentences or documents by topic or sentiment. Such dictionary-based methods essentially treat text as collections of discrete tokens—either a specific word appears or it does not—creating sparse, binary representations. These methods are transparent and computationally efficient, but fundamentally fragile: the same economic concept can be expressed through countless word combinations and phrasings, yet keyword approaches can only capture exact matches to predetermined terms. They miss synonyms, struggle with context, and degrade as language evolves.

A second class of methods addresses these limitations through advances in text representation learning. The key innovation is representing text as dense numerical vectors that capture underlying meaning rather than surface-level vocabulary. The core insight is that words and phrases with similar meanings can be positioned close together in a continuous mathematical space, allowing algorithms to recognize conceptually related content regardless of specific wording. Early approaches like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) demonstrated that words appearing in similar text contexts develop similar representations. More sophisticated transformer models like BERT (Devlin et al.,

2019) extend this logic to entire sentences, enabling context-dependent understanding where meaning depends on surrounding text rather than isolated words.

In economic applications, these representations are typically used as inputs to supervised machine learning classifiers trained for specific domains. de Araujo et al. (2025) incorporate embeddings from central bank communication to forecast inflation, while Gu et al. (2024) use embeddings of firm disclosures to predict post-announcement stock returns. Domain-specific models such as ClimateBERT (Webersinke et al., 2022) and FinBERT (Huang et al., 2023) achieve strong performance on sentiment and topic classification within their intended applications. However, this contextual sophistication comes at a cost: supervised approaches require expert-annotated training data and remain locked to their original domains, creating scalability constraints when researchers wish to measure new concepts or apply existing frameworks to different corpora.

While supervised approaches can achieve high accuracy, their domain specificity limits broader application. More recently, large language models have offered a potential solution by enabling zero-shot classification. These methods use natural language prompts to specify classification tasks, allowing researchers to define new categories through descriptive instructions rather than labeled examples. However, LLM-based approaches introduce their own limitations: they are computationally expensive at scale, often require costly API access, and produce non-deterministic results that can vary across model versions. For large-scale economic measurement involving millions of texts, the computational costs can become prohibitive while the lack of transparency makes validation and replication challenging.

We address this scalability-precision trade-off by applying semantic retrieval rather than classification. Our approach combines the computational efficiency of keyword methods with the semantic sophistication of dense embeddings, enabling both large-scale analysis and contextual understanding. Instead of training models to predict labels, we define economic concepts through query sets—short phrases that articulate the construct of interest—and retrieve passages whose embeddings exhibit high similarity to the query embeddings. This builds on advances in dense retrieval from information retrieval, particularly bi-encoder architectures such as Sentence-BERT (Reimers and Gurevych, 2019a) and dual-encoder systems like DPR (Karpukhin et al., 2020), which enable efficient similarity search over large corpora via approximate nearest-neighbor indexing. The approach offers decisive advantages for eco-

nomic measurement: new concepts can be specified without labeled data, applied instantly to any corpus without retraining, and validated through transparent similarity scores. Unlike supervised classifiers, it requires no annotation effort; unlike direct LLM application, it is deterministic and cost-effective at massive scale. Moreover, by efficiently filtering millions of passages to a manageable subset, semantic retrieval enables the optional deployment of precise but expensive methods—such as large language models—where maximum accuracy is needed, combining scalability with precision in a modular architecture. This hybrid approach makes it possible to track emerging risks and shifting narratives as they appear in language—before they crystallize into established categories—while maintaining the transparency and computational efficiency essential for rigorous economic research.

We address this scalability-precision trade-off by applying semantic retrieval rather than classification. Our approach combines the computational efficiency of keyword methods with the semantic sophistication of dense embeddings, enabling both large-scale analysis and contextual understanding. Instead of training models to predict labels, we define economic concepts through query sets—short phrases that articulate the construct of interest—and retrieve passages whose embeddings exhibit high similarity to the query embeddings. This builds on advances in dense retrieval from information retrieval, particularly bi-encoder architectures such as Sentence-BERT (Reimers and Gurevych, 2019a) and dual-encoder systems like DPR (Karpukhin et al., 2020), which enable efficient similarity search over large corpora via approximate nearest-neighbor indexing. The approach offers decisive advantages for economic measurement: new concepts can be specified without labeled data, applied instantly to any corpus without retraining, and validated through transparent similarity scores. Unlike supervised classifiers, it requires no annotation effort; unlike LLMs, it is deterministic, cost-effective at massive scale, and independent of external APIs. This combination makes it possible to track emerging risks and shifting narratives as they appear in language—before they crystallize into established categories—while maintaining the transparency and computational efficiency essential for rigorous economic research.

III Data

A. Earnings Call Transcripts

We obtain quarterly earnings call transcripts from FactSet for all firms included in the S&P 500 and STOXX 600 indices over the period 2010–2024. The dataset covers both prepared management remarks and question-and-answer sessions with financial analysts, providing a comprehensive record of forward-looking corporate communication across sectors and regions.

Earnings calls are a central disclosure channel through which firms communicate expectations, risks, and strategic priorities to investors. They occur at regular quarterly intervals, are subject to legal and reputational constraints, and are widely used in the literature to study firm-level narratives and beliefs. We link each transcript to firm identifiers and reporting dates, forming the textual basis for our analysis.

Table 2 reports descriptive statistics for the earnings call corpus by region and year. The final dataset contains 51,567 transcripts covering 1,486 unique firms, with 21,827 calls from STOXX 600 firms and 29,740 calls from S&P 500 firms. This processing yields approximately 4.4 million paragraphs, with an average of 85 paragraphs per earnings call transcript.

The scale of this textual corpus illustrates the computational challenge that motivates our scalable semantic search approach. Processing this corpus with direct LLM classification would cost thousands of dollars while local GPU-based alternatives demand specialized hardware and extensive processing periods. More fundamentally, emerging risks, by definition, evolve rapidly and often lack established taxonomies or validated measurement frameworks. Researchers studying these phenomena must be able to quickly test alternative conceptualizations, refine risk categories, and adapt to new terminology as understanding develops. Traditional supervised approaches require weeks or months of annotation for each new taxonomy, making them poorly suited for the fast-moving, exploratory analysis that emerging risk research demands. In contrast, our semantic search framework allows researchers to define new concepts through natural language queries and obtain results within minutes, enabling the rapid iteration essential for studying evolving phenomena like climate risk, technological disruption, or geopolitical threats.

The panel structure differs markedly across regions. The S&P 500 sample is highly stable over time, with more than 93% of firms appearing in both halves of the sample period and year-over-year retention rates typically exceeding 97%. As a result, changes in aggregate textual measures for U.S. firms are unlikely to be driven by mechanical sample entry or exit.

In contrast, the STOXX 600 sample exhibits substantially higher firm turnover despite a stable index size. Only about 69% of firms appear in both halves of the sample, and no firm is observed continuously throughout the full period. This higher measured entry and exit reflects both genuine index rebalancing and limitations of the current firm identifier construction in the European sample.

In the empirical analysis, we explicitly account for this unbalanced panel structure using firm fixed effects, robustness checks on restricted subsamples, and region-specific interpretations. Importantly, the main results do not rely on balanced panels and are robust to excluding firms with short sample presence.

Table 1: Earnings Call Dataset: Overview Statistics

Note: This table presents summary statistics for the earnings call transcript dataset covering S&P 500 and STOXX 600 firms from 2009–2025. The dataset contains 51,567 transcripts totaling 4.4 million paragraphs, illustrating the computational scale that motivates scalable measurement approaches. Transcripts are parsed into management discussion and Q&A sections. YoY retention rates show the percentage of firms appearing in consecutive years.

	EU (STOXX600)	**US (SP500)**
Total Transcripts	21,827	29,740
Unique Firms	811	590
Avg Words per Transcript	9,921	9,907
Avg Paragraphs per Transcript	82.4	86.7
Total Paragraphs (millions)	**1.8**	**2.6**
Avg Years per Firm	8.0	12.6
YoY Firm Retention Rate	93.3%	98.3%

Table 2: Earnings Call Dataset: Sample Composition by Year

Note: This table shows the temporal distribution of earnings call transcripts by year and region. The S&P 500 sample exhibits high stability with consistent firm participation, while the STOXX 600 sample shows more variation reflecting index rebalancing and data availability. Sample composition differences are addressed through firm fixed effects and region-specific analyses in empirical specifications.

	EU (STOXX600)		**US (SP500)**	
Year	**Transcripts**	**Firms**	**Transcripts**	**Firms**
2010	1,024	327	1,902	442
2011	1,135	368	2,012	448
2012	1,172	378	1,755	454
2013	1,229	387	1,829	464
2014	1,280	407	1,867	475
2015	1,449	558	1,736	477
2016	1,346	435	1,879	477
2017	1,362	438	1,908	489
2018	1,398	448	1,705	483
2019	1,541	518	1,907	488
2020	1,621	523	1,916	489
2021	1,036	532	1,937	491
2022	1,699	545	1,923	493
2023	1,728	557	1,941	496
2024	1,393	566	1,735	487
2025	1,414	567	1,788	491
Total / Unique	**21,827**	**811**	**29,740**	**590**

B. Financial Market and Accounting Data

We obtain stock returns and firm-level financial characteristics from Bloomberg and Compustat. Monthly stock returns are constructed from Bloomberg's total return index, which incorporates both price changes and reinvested dividends. Excess returns are computed by subtracting the risk-free rate, measured using the 3-month Treasury bill rate for U.S. firms and the 3-month Euribor rate for European firms, which are also obtained from Bloomberg.

We further collect standard firm characteristics commonly used in asset pricing analyses, including firm size, book-to-market ratios, capital expenditures, and profitability measures from Compustat (S&P Capital IQ). To mitigate the influence of extreme values in accounting ratios, these variables are winsorized at the 1st and 99th percentiles of their cross-sectional distributions. All financial variables are aligned to the earnings call timeline.

C. Carbon Emissions Data

Firm-level carbon emissions data are obtained from MSCI. We use annual measures of Scope 1 and Scope 2 greenhouse gas emissions, expressed in metric tons of CO_2 equivalent. To facilitate comparability across firms of different sizes, we rely on MSCI's emissions intensity measure, which scales total Scope 1 and Scope 2 emissions by firm revenue.

MSCI reports emissions data with a one-year publication lag. To ensure consistency with investors' information sets, emissions measures are lagged accordingly in empirical specifications. In the regression analysis, we work with the logarithm of emissions intensity. Emissions data are used to benchmark text-based climate measures against conventional proxies of firms' carbon exposure.

D. Green Patent Data

To capture firm-level innovation related to climate mitigation technologies, we use patent data from the Google Patents Public Datasets, accessed via Google BigQuery. We focus on patents classified as environmentally related based on Cooperative Patent Classification (CPC) codes associated with climate change mitigation technologies (Y02). These classifications are widely used in the literature to identify green innovation and provide a technology-based measure of firms' environmental activities.

Patent records include information on filing years and harmonized assignee names. Because patent assignees are not uniquely identified using standard financial identifiers, we link patents to firms in our panel using fuzzy string matching between harmonized assignee names and company names in the earnings call dataset. Patent counts are then aggregated at the firm-year level based on filing year. Firms without matched green patents in a given year are assigned a value of zero.

IV Method

Figure 1 provides an overview of the semantic measurement pipeline, which transforms raw corporate transcripts into economically interpretable measures of climate risk exposure through four sequential phases. Phase I establishes the data infrastructure by processing earnings call PDFs into a structured paragraph-level corpus, embedding each text segment into a high-dimensional semantic vector space, and constructing an efficient similarity search index. Phase II applies semantic search to identify climate-relevant content using theory-

grounded query sets derived from the TCFD taxonomy, substantially reducing the search space while maintaining high recall. Phase III optionally refines these classifications using a large language model to distinguish between risk-oriented and opportunity-oriented narratives with maximum precision. Finally, Phase IV aggregates the classified paragraphs into firm-quarter economic indicators that capture both the intensity of climate attention and its directional framing. This modular architecture enables researchers to choose their preferred point on the cost-accuracy frontier while maintaining full scalability and reproducibility. The remainder of this section describes each component in detail.

Phase I: Data Infrastructure

First, we extract the full transcript text from each PDF and identify the two main parts of an earnings call: prepared management remarks and the subsequent question-and-answer session.[1] We harmonize call-level metadata, including reporting date, fiscal quarter, firm name, and ticker, using index-specific parsing rules that account for systematic formatting differences between US and EU transcripts. Within each call, speaker turns are identified based on standardized cues, yielding a structured representation that preserves the internal organization of the discussion across sections and speakers. To finalize data parsing and cleaning, we segment each call into individual paragraphs and construct a paragraph-level dataset in which each observation corresponds to a single coherent text block. For each paragraph, we retain firm and call identifiers, temporal information, section labels (management or Q&A), and speaker metadata.[2]

This processing pipeline produces a clean, structured corpus of earnings call text that serves as the input for semantic embedding, search, and indicator construction in subsequent steps. However, the raw textual data remains fundamentally unstructured for computational analysis. Traditional keyword-based approaches would struggle to capture the contextual nuance and evolving terminology that characterizes climate-related business discourse. To enable scalable identification of semantically related content regardless of surface-level wording differences, we transform each paragraph into a numerical representation that preserves semantic meaning.

[1] During this step, we also remove formatting noise and FactSet boilerplate statements.

[2] While speaker-level metadata are retained in the processed dataset, they are not exploited in the empirical analysis presented in this paper. The data processing pipeline preserves this information to facilitate future work on speaker and communication dynamics.

Figure 1: Semantic Measurement Pipeline

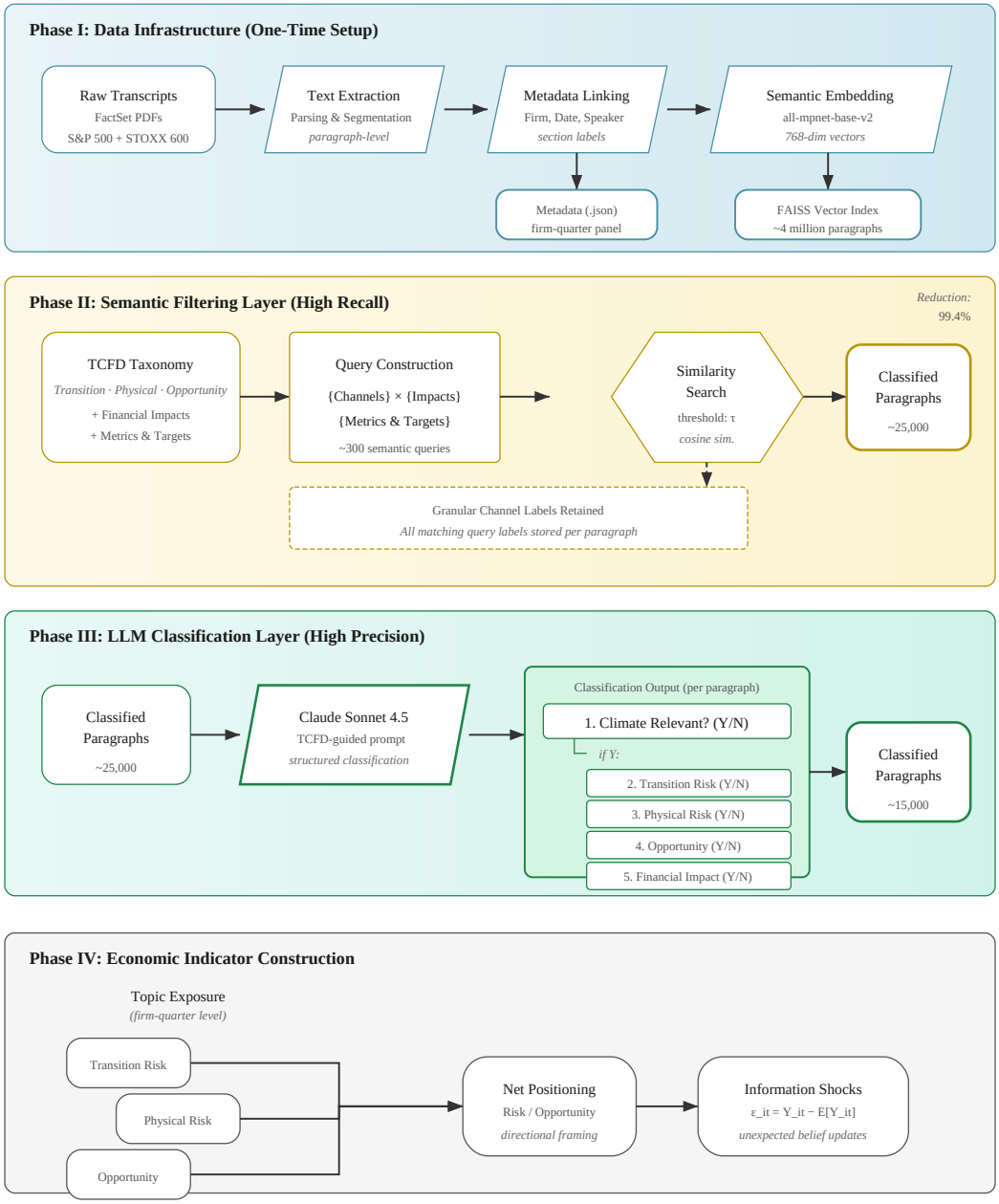


Figure 1: The pipeline converts raw earnings call transcripts into paragraph-level semantic representations, classifies climate-related risk and opportunity narratives, and aggregates them into firm-level economic indicators.

Hence, we embed each paragraph into a dense vector representation using `all-mpnet-base-v2`, a general-purpose sentence transformer designed for semantic similarity tasks (Reimers and Gurevych, 2019b). Each paragraph p is mapped to a vector $\mathbf{x}_p \in \mathbb{R}^{768}$. All embeddings are L2-normalized so that inner products correspond to cosine similarity, which depends on the angle between vectors in the embedding space. Paragraphs that convey similar semantic content are therefore mapped to nearby areas, even when they differ in vocabulary or phrasing. The embedding step is deterministic and applied once to the full corpus.

The resulting embeddings are stored in a FAISS index, which supports efficient approximate nearest-neighbor search over large vector collections (Johnson et al., 2019). This index enables similarity queries over several million paragraphs to be executed in sub-second time. The corpus contains approximately four million paragraphs. A brute-force similarity search would require computing millions of high-dimensional dot products per query. Approximate nearest-neighbor indexing reduces this cost by several orders of magnitude while preserving search accuracy.

Paragraph-level metadata, including firm identifiers, call dates, sections, and speaker labels, are stored separately and linked to the index through stable keys, allowing flexible filtering and aggregation after topic search.

A key advantage of this design is that embedding and indexing are performed only once. New topics can be analyzed by embedding a small set of query texts and querying the existing index, without reprocessing the corpus. This sharply contrasts with supervised classifiers or large language model pipelines that require retraining or repeated full-corpus inference for each new construct, making semantic search particularly well suited for scalable and extensible economic measurement.

Phase II: Semantic Search Layer

Building on the embedded corpus from Phase I, the next step is to identify climate-relevant content through semantic similarity search, commonly called semantic search. The challenge is to construct query sets that comprehensively capture climate-related business discourse while maintaining theoretical grounding and minimizing false positives. Rather than relying on ad-hoc keyword selection, we ground our query construction in the Task Force on Climate-related Financial Disclosures (TCFD) taxonomy, which provides a systematic

framework for identifying the channels through which climate-related risks and opportunities may materially affect firms (Task Force on Climate-related Financial Disclosures, 2017).

The TCFD framework organizes climate-related business impacts into three categories: transition risks (policy, technology, market, and reputation changes), physical risks (acute and chronic climate impacts), and climate-related opportunities (resource efficiency, energy sources, products/services, markets, and resilience). Each category contains a hierarchical taxonomy that extends from broad categories down to specific impact channels. This structure enables systematic query construction that ensures comprehensive coverage while maintaining theoretical grounding.

We construct queries using a bottom-up approach that begins with the most granular level of the TCFD taxonomy. For each lowest-level risk or opportunity channel, we create multiple query variants that capture both the source of exposure and its potential business impacts. This dual approach recognizes that firms may discuss climate factors either in terms of their underlying drivers or their anticipated effects on operations and financial performance.

Table 3 illustrates this construction process with representative examples across the three TCFD pillars. For transition risks, a policy risk such as carbon pricing mechanisms generates queries that capture both the regulatory instrument itself ("carbon tax implementation") and its business impact ("increased operational costs from carbon pricing"). Similarly, physical risks like extreme weather events produce queries covering both the hazard ("flooding and storm damage") and its operational consequences ("supply chain disruption from weather events"). Climate opportunities follow the same pattern, producing queries combining the opportunity source ("energy efficiency technologies") and its financial benefit ("cost savings from reduced energy consumption").

Beyond risk and opportunity sources, we use the TCFD pillar "Metrics & Targets" to construct an additional query set capturing ways in which firms may disclose and evaluate climate-related risks and opportunities. These queries may identify passages discussing carbon footprints, emissions reduction targets, climate-related financial disclosures, exposed assets, and other exposure or performance indicators.

In total, this systematic approach yields approximately 340 distinct semantic queries across all TCFD categories, each designed to capture a specific aspect of climate-related

Table 3: Query Construction Examples

TCFD Pillar	Risk/Opportunity Channel	Financial Impact	Example Queries
Transition Risk	Carbon pricing mechanisms	Increased costs	"carbon tax implementation", "emissions trading scheme costs", "carbon pricing regulatory compliance"
Transition Risk	Technology shifts	Capital expenditure	"renewable energy transition costs", "stranded fossil fuel assets", "clean technology investment requirements"
Physical Risk	Extreme weather events	Supply chain disruption	"flooding impact on operations", "hurricane supply chain disruption", "weather-related facility damage"
Physical Risk	Sea level rise	Asset impairment	"coastal facility flooding risk", "infrastructure climate adaptation", "sea level rise asset exposure"
Opportunity	Energy efficiency	Cost reduction	"energy efficiency cost savings", "operational efficiency improvements", "reduced energy consumption benefits"
Opportunity	Renewable energy	Revenue generation	"clean energy revenue opportunities", "renewable energy market growth", "green product demand increase"
Metrics & Targets	Emissions measurement	Risk management	"carbon footprint reporting", "scope 3 emissions tracking", "net zero commitment progress"

business impact. The comprehensive query set ensures broad coverage of climate-related discourse while maintaining clear conceptual boundaries between different types of exposure.

Semantic search produces a continuous measure of similarity between each paragraph and the concept-defining queries. Each economic concept c is represented by a set of query texts $Q^c = \{q_1^c, \ldots, q_M^c\}$. Each query q_j^c is embedded using the same sentence transformer model (`all-mpnet-base-v2`) that produced the paragraph embeddings, yielding query vectors $\mathbf{q}_j^c \in \mathbb{R}^{768}$ that are L2-normalized to unit length.

For every paragraph p with embedding $\mathbf{x}_p$, we compute its similarity to concept c as:

$$\text{sim}(p, c) = \max_{j \in \{1,\ldots,M\}} \mathbf{x}_p^T \mathbf{q}_j^c \tag{1}$$

This maximum operation ensures that a paragraph is considered relevant to concept c if it exhibits high similarity to any query in the concept's query set, capturing the diverse ways that economic concepts may be expressed in corporate discourse. We apply this process separately for each of the three TCFD categories, yielding distinct similarity scores for physical risks, transition risks, and climate opportunities for every paragraph in the corpus.

The similarity computation leverages the FAISS approximate nearest neighbor index to efficiently identify high-similarity paragraphs without computing exhaustive pairwise com-

parisons. For each query embedding, FAISS returns the top-k most similar paragraph embeddings along with their similarity scores, where k is set sufficiently high to capture all paragraphs above the relevance threshold.

To convert continuous similarity scores into binary relevance classifications, we impose concept-specific thresholds τ^c. A paragraph is classified as relevant to concept c if $\text{sim}(p, c) > \tau^c$. The choice of threshold represents a fundamental trade-off between recall (capturing all relevant content) and precision (avoiding false positives).

We optimize these thresholds through systematic validation using performance-gradient adaptive sampling. This approach concentrates annotation effort in similarity score regions where threshold choice has the greatest impact on classification performance, enabling efficient assessment of precision-recall trade-offs across the full range of potential thresholds. The validation protocol, including stratified sampling methodology and performance evaluation across different threshold values, is detailed in Section VI.

Phase III: LLM Classification Layer

The semantic search layer provides a complete set of classified paragraphs suitable for many downstream analyses. However, semantic queries are intentionally permissive to maximize recall: a query designed to capture transition risks may also retrieve passages discussing related opportunities or neutral climate content due to similar semantic patterns. When applications require maximum precision in distinguishing risk from opportunity sentiment, we deploy an optional second-stage classification layer that refines the semantic search results.

This two-stage architecture confines computationally expensive LLM processing to the subset of paragraphs already identified as climate-relevant, dramatically reducing costs while preserving accuracy. The 45,000 paragraphs classified by semantic search represent a 99.4% reduction from the original corpus, making precise LLM classification economically feasible.

We implement the precision layer using Claude Sonnet 4 via Anthropic's Batch API, which processes paragraphs in groups of 40 with structured prompts based on the TCFD taxonomy. Each paragraph receives six classifications: climate relevance, transition risk, physical risk, opportunity, financial impact channel, and impact direction. The structured prompt guides the model to identify specific climate-related business impacts while distinguishing between risks and opportunities that may be semantically similar but economically distinct.

All subsequent empirical analyses employ the precision layer classifications, while robustness checks using semantic search scores alone are provided in the appendix. This design allows researchers to choose their preferred point on the cost-accuracy frontier while maintaining full transparency about methodological trade-offs.

This two-stage architecture confines computationally expensive LLM processing to the subset of paragraphs already identified as climate-relevant, dramatically reducing costs while preserving accuracy. The 45,000 paragraphs classified by semantic search represent a 99% reduction from the original corpus, making precise LLM classification economically feasible. Our implementation costs approximately $43 to process the full 4 million paragraph corpus using Claude Sonnet 4 Batch API with caching optimization. In contrast, applying LLM classification directly to all paragraphs without semantic pre-filtering would cost over $3,800, a 100-fold increase that renders such analyses impractical for most research budgets. Local GPU-based alternatives require specialized hardware and weeks of processing time, while offering no flexibility for exploratory analysis or pattern discovery. The two-stage architecture therefore transforms expensive precision tools into practical research instruments, enabling systematic exploration of corporate climate discourse that would otherwise be economically infeasible.

This cost differential becomes even more pronounced at larger scales. While our study covering S&P 500 and STOXX 600 firms represents a substantial corpus, direct LLM approaches quickly become prohibitively expensive for broader firm universes or extended time horizons. However, the scalability of FAISS indexing, which can efficiently handle corpora exceeding 100 million paragraphs on standard hardware, means our two-stage approach can accommodate much larger research designs. A study covering all publicly traded firms globally over multiple decades would cost thousands of dollars using our framework, compared to hundreds of thousands using direct LLM classification. This scalability advantage transforms narrative-based empirical research from a niche methodology constrained to small samples into a broadly applicable tool for large-scale economic analysis.

Local GPU-based alternatives require specialized hardware and weeks of processing time, while offering no flexibility for exploratory analysis or pattern discovery. The two-stage architecture therefore transforms expensive precision tools into practical research instruments, enabling systematic exploration of corporate climate discourse that would otherwise be eco-

nomically infeasible.

Phase IV: Economic Indicator Construction

The objective of our measurement framework is to construct firm-level indicators that capture the intensity of discussion around climate-related topics. We view an earnings call as a finite allocation of managerial attention across competing issues, and we use the distribution of textual content within the call to quantify this attention.

Let i index firms and t index quarters. Each earnings call is represented as a collection of paragraphs $\{(p_{it})_j\}_{j=1}^{P_{it}}$, where P_{it} denotes the total number of paragraphs in firm i's earnings call in quarter t. Let L_j denote the number of sentences in paragraph j, and $N_{it} = \sum_{j=1}^{P_{it}} L_j$ denote the total number of sentences in the call. The classification pipeline produces paragraph-level labels indicating relevance to specific climate-related categories and channels. We can construct indicators using either the semantic search classifications alone (Phase II) or the refined LLM classifications (Phase III), depending on their precision requirements.

For each of the three primary TCFD categories $c \in \{\text{Physical Risk}, \text{Transition Risk}, \text{Opportunity}\}$, we define $\mathcal{S}_{it}^c \subset \{(p_{it})_j\}$ as the subset of paragraphs classified as relevant to concept c. The primary exposure measure is the share of the conference call (measured in sentences) devoted to concept c:

$$\text{TopicExposure}_{it}^c = \frac{\sum_{j \in \mathcal{S}_{it}^c} L_j}{N_{it}} \tag{2}$$

This normalization yields a scale-free attention measure that is comparable across firms and time periods regardless of transcript length. The multi-label nature of our classification means that $\sum_c \text{TopicExposure}_{it}^c$ may exceed the total share of climate-relevant discussion, as paragraphs can contribute to multiple exposure measures simultaneously.

Beyond these aggregate measures, our framework enables construction of more granular exposure indicators for specific TCFD subchannels, such as carbon pricing mechanisms, climate litigation, or renewable energy opportunities. However, our downstream empirical analyses focus primarily on the three aggregate category measures, which provide sufficient granularity for most economic research applications while maintaining statistical power.

V Validation of the classifier

This section establishes the technical validity of our two-stage classification pipeline by evaluating performance against manually annotated data and established benchmarks. We demonstrate that our approach achieves high precision and recall across all TCFD categories while offering decisive advantages in scalability and adaptability over existing methods. The validation employs density-weighted sampling to ensure reported performance metrics accurately reflect expected classification accuracy when applied to the full corpus.

A. Threshold Selection and Validation Methodology

Converting continuous similarity scores to binary classifications requires setting appropriate thresholds for each TCFD category. This critical step determines the recall-precision trade-off of our measurement framework. Similarity scores between earnings call paragraphs and individual climate queries exhibit highly right-skewed distributions with most content concentrated in very low similarity ranges and sparse tails extending toward higher similarities. When aggregating across larger query sets within each TCFD category, distributions shift toward middle ranges (0.2–0.5) as more paragraphs achieve moderate similarity to at least one query, but maintain their right-skewed character.

We employ different threshold optimization strategies depending on intended use. For standalone semantic search classification, we optimize thresholds to maximize F1-scores, balancing precision and recall. For the first stage of our two-stage approach, we deliberately lower thresholds to maximize recall, accepting reduced precision since the subsequent LLM layer filters false positives. This exploits each layer's comparative advantage: semantic search provides comprehensive coverage at scale while the LLM delivers nuanced classification on the filtered subset.

We validate threshold choices using density-weighted sampling. This approach divides similarity score space into 0.05-width bins and samples paragraphs from each bin proportional to population density. The method ensures adequate sample sizes for accurate performance estimation, particularly in high-density, low-to-medium similarity regions where most irrelevant paragraphs cluster despite low positive rates. Manual annotation of sampled paragraphs as true/false positives and negatives enables computation of performance metrics weighted by bin density.

B. Classification Performance and Benchmarking

To benchmark our approach, we compare against ClimateBERT (Webersinke et al., 2022), a transformer model specifically fine-tuned on climate-related financial and business disclosures. ClimateBERT represents the current state-of-the-art for supervised climate content detection, having been trained on expert-labeled datasets to identify climate-relevant passages in corporate communications. This makes it an ideal benchmark for evaluating our unsupervised semantic search approach.

Table 4 demonstrates the performance of our approach across different configurations and benchmarks. Panel A presents the most key result: our two-stage method achieves substantially higher precision than ClimateBERT (0.93 vs. 0.47) while maintaining competitive recall (0.70 vs. 0.79). This translates to a significantly higher F1-score (0.86 vs. 0.58), indicating that our approach provides superior overall classification performance. Notably, even our semantic search layer alone outperforms ClimateBERT in F1-score (0.61 vs. 0.58), despite requiring no domain-specific training data.

The performance breakdown across TCFD categories in Panels B and C reveals several important patterns. First, the semantic search layer shows moderate but balanced performance across categories, with precision ranging from 0.59 to 0.78 and recall from 0.40 to 0.63. Physical risk classification proves most challenging for the semantic layer, achieving lower recall (0.40) than the other categories. This likely reflects the lower prevalence of physical risk in discussions, which reduces the effective sample size for this category in our density-weighted performance calculations. Second, the LLM refinement layer delivers substantial precision gains across all categories, with precision consistently above 0.90 while maintaining recall between 0.69 and 0.77. The results demonstrate the value of the two-stage architecture: the semantic layer provides comprehensive coverage while the LLM layer achieves the precision needed for high-quality economic measurement.

The superior performance relative to ClimateBERT is particularly noteworthy given the fundamental differences in approach. While ClimateBERT requires extensive labeled training data and retraining for new domains, our method generalizes immediately to new taxonomies through simple query specification. This combination of higher accuracy and greater flexibility positions our framework as a more practical solution for large-scale textual measurement in economics and finance.

Table 4: Classification Performance: Semantic Search and LLM Layers

Note: This table reports classification performance for the two-stage measurement pipeline. Panel A shows weighted performance metrics for the semantic search layer (Phase II) across TCFD categories, with thresholds optimized for maximum F1-score. Panel B shows performance for the LLM classification layer (Phase III) applied to semantically filtered paragraphs. Metrics are weighted by bin population density from stratified sampling validation. Sample size indicates manually annotated paragraphs.

Panel A: Benchmark Comparison (general climate)				
Method	**Precision**	**Recall**	**F1**	**Threshold**
Our Method (Two-stage)	0.93	0.70	0.86	—
Semantic Search Only	0.79	0.49	0.61	0.58
ClimateBERT	0.47	0.79	0.58	0.90

Panel B: Semantic Search Layer				
Category	**Precision**	**Recall**	**Threshold**	**Sample Size**
Physical Risk	0.78	0.40	0.58	400
Transition Risk	0.62	0.58	0.62	400
Opportunity	0.59	0.63	0.61	400
Climate (All)	0.79	0.56	0.58	400

Panel C: Two-stage method (Semantic + LLM)				
Category	**Precision**	**Recall**	**Threshold**	**Sample Size**
Physical Risk	0.94	0.69	0.55	400
Transition Risk	0.90	0.74	0.59	400
Opportunity	0.93	0.77	0.59	400
Climate (All)	0.93	0.70	0.56	400

To provide additional validation, we also evaluate both methods on the ClimateBERT detection dataset, the expert-labeled corpus on which ClimateBERT was originally trained and evaluated. This provides a direct comparison on the supervised model's home turf. On this dataset, ClimateBERT achieves superior performance with precision of 0.98, recall of 0.98, and F1-score of 0.98. Our semantic search approach achieves precision of 0.96, recall of 0.85, and F1-score of 0.90. While ClimateBERT outperforms our method on this specialized dataset, our approach delivers competitive results without requiring any domain-specific training data or fine-tuning.

VI Validation of the economic properties

This section evaluates whether the semantic measures behave like economically meaningful firm-level information variables rather than mechanical artifacts of language or disclosure conventions. The goal is not to exhaustively document correlations, but to assess whether the measures exhibit a set of properties that are necessary for their use in applied empirical research.

We organize the validation around five such properties. First, the measures should capture selective attention to climate issues rather than boilerplate climate disclosure. If climate attention is meaningful, most firms should rarely discuss climate in their earnings calls, while a small subset should devote substantial discussion at particular times, resulting in sparse and right-skewed distributions. Second, meaningful variation should be predominantly firm-specific rather than driven by industry or aggregate time effects. Third, cross-sectional patterns should align with observable economic exposure and behavior, such as firms' carbon intensity and innovation activity. Fourth, unexpected changes in attention should respond to identifiable information arrivals, consistent with an interpretation of attention innovations as information shocks. Finally, the measurement framework should generalize beyond the climate domain without domain-specific retraining or tuning.

The subsections below assess these properties in turn, using descriptive evidence, variance decompositions, event responses, and links to observable firm outcomes.

A. Descriptive Statistics

We begin by assessing whether climate-related attention is selective rather than driven by routine disclosure language. Table 5 reports summary statistics for the climate-related semantic measures. The average firm devotes 0.60% of earnings call content to climate-related discussion, with substantial right-skew: the median is near zero while the 95th percentile reaches 3.87%. This pattern—many firms with minimal discussion and a tail of high-attention firms—is consistent across all climate dimensions. Physical risk discussion is rarest (mean 0.03%), while climate opportunities are discussed more frequently than risks across all categories. The skewness suggests that climate attention remains concentrated among a subset of firms even as aggregate attention has grown.

Table 5: Descriptive Statistics for Climate-Related Semantic Measures

Note: This table reports descriptive statistics for semantic climate indicators extracted from earnings call transcripts (2009–2025). Values represent percentages of total sentences in each call. Measures are constructed using semantic similarity retrieval with LLM precision filtering. Main categories (Physical Risk, Transition Risk, Opportunity) are shown in bold, with subcategories indented below each. All statistics are computed over the full sample ($N = 51,567$ earnings calls).

Variable	Mean	Std.Dev.	Min	P95	Max	% Zero
Physical Risk	0.15	0.80	0.00	0.83	23.24	92.7
Acute	0.07	0.51	0.00	0.00	18.28	95.9
Chronic	0.04	0.36	0.00	0.00	11.85	97.6
Transition Risk	0.14	0.86	0.00	0.52	30.78	94.3
Policy	0.04	0.45	0.00	0.00	30.78	98.0
Legal	0.00	0.07	0.00	0.00	5.08	99.7
Technology	0.03	0.33	0.00	0.00	12.64	98.1
Market	0.04	0.37	0.00	0.00	17.24	97.9
Reputation	0.00	0.11	0.00	0.00	11.47	99.8
Opportunity	0.59	2.05	0.00	3.64	37.81	80.6
Energy Source	0.27	1.20	0.00	1.54	31.19	88.4
Products & Services	0.13	0.69	0.00	0.88	26.65	91.7
Markets	0.06	0.48	0.00	0.00	18.68	96.3
Resource Efficiency	0.12	0.62	0.00	0.74	36.62	92.5
Resilience	0.17	1.01	0.00	0.76	28.25	93.3

B. Variance Decomposition

We next examine whether variation in climate attention is predominantly firm-specific rather than driven by industry or aggregate time effects. Table 6 decomposes variation in climate attention across time, industry, and firm dimensions. Panel A shows that year fixed effects explain 5% of variation in general climate attention—reflecting the aggregate upward trend visible in Figure 2—while industry fixed effects explain an additional 4%. The bulk of variation (88%) occurs at the firm level, indicating that climate attention is primarily a firm-specific characteristic rather than an industry or time phenomenon. Panel B examines within-firm variation: firm fixed effects explain 53% of within-firm variance for general climate attention, implying that roughly half of a firm's climate discussion reflects persistent firm characteristics while the other half varies over time in response to events or changing circumstances. This concentration of variation at the firm level is desirable: it indicates that the measures capture meaningful heterogeneity across firms within the same industry, rather than merely reflecting sector-wide characteristics that could be proxied by industry dummies.

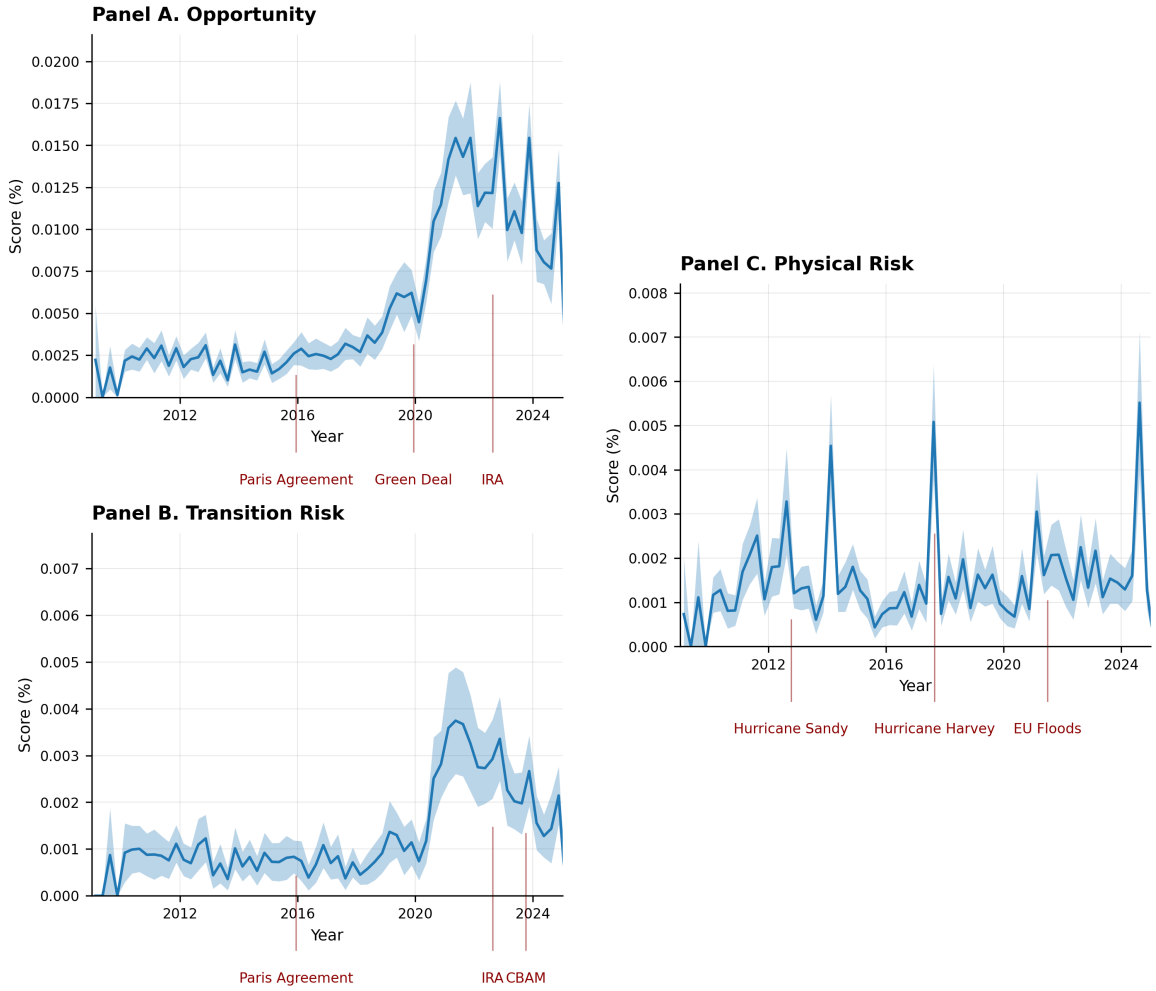


Figure 2: Evolution of Climate Risk Attention in Corporate Earnings Calls, 2009-2024. This figure displays the time series evolution of climate-related attention across four dimensions measured from earnings call transcripts. Panel A shows climate regulation attention, Panel B shows climate innovation opportunity attention, Panel C shows climate physical risk attention, and Panel D shows climate transition risk attention. All measures are expressed as percentages of total sentences in earnings calls. The sample includes S&P 500 and STOXX 600 firms. Vertical red lines indicate key climate policy milestones. The shaded areas represent 95% confidence intervals.

Table 6: Variance Decomposition of Climate-Related Semantic Measures

Note: This table decomposes variation in climate-related semantic measures across firms, sectors, and time. Panel A reports incremental R^2 from regressions including year, industry, industry×year, country, and firm fixed effects. Panel B shows the share of within-firm variation explained by firm fixed effects.

Panel A: Incremental R^2

Variable	Year FE	Industry FE	Industry×Year FE	Country FE	Firm-level	Sum
General	5.0%	3.9%	2.2%	0.6%	88.2%	100%
Regulation	0.9%	1.7%	2.0%	0.0%	95.4%	100%
Transition	5.2%	3.9%	2.3%	0.8%	87.7%	100%
Physical	0.6%	0.5%	3.7%	0.3%	94.8%	100%

Panel B: Within-Firm Fraction of Variation

Variable	Firm FE	Residual	Sum
General	53.1%	46.9%	100%
Regulation	34.2%	65.8%	100%
Transition	52.2%	47.8%	100%
Physical	20.2%	79.8%	100%

C. Cross-sectional Validity: Sector Patterns

If the semantic measures capture economically meaningful climate exposure, firms in sectors with higher emissions or greater regulatory exposure should exhibit systematically higher climate attention. If the semantic measures capture genuine climate exposure, sectors with higher emissions or greater regulatory exposure should exhibit elevated attention. Table 7 reports sector-level averages. Utilities rank first in overall climate attention (2.09%), followed by Energy (1.11%) and Automobiles (0.95%)—sectors with substantial carbon footprints and direct exposure to transition policies. For transition risk specifically, Utilities and Energy again lead, consistent with their exposure to decarbonization pressures. Physical risk attention is highest in Insurance (0.04%) and Automobiles (0.05%), reflecting exposure to weather-related claims and supply chain disruptions. These cross-sectional patterns align with economic priors about which sectors face the greatest climate-related challenges.

Table 7: Sector-Level Climate Attention and Risk

Note: This table reports sector-level descriptive statistics of climate-related semantic measures derived from earnings call transcripts. Values represent the share of topic-relevant paragraphs per call, expressed in percentages. Only the top-ranked sectors by mean exposure are shown for each panel. Sample spans 2009–2025.

Panel A: Climate Attention

Sector	Mean (%)	Std.Dev. (%)	P95 (%)	N Calls	N Firms
Utilities	2.09	4.36	10.77	1,155	32
Energy	1.11	2.93	7.01	1,097	25
Automobiles & Components	0.95	2.66	6.22	725	15
Real Estate Management & Development	0.89	2.44	5.00	559	15
Media & Entertainment	0.82	2.39	5.48	833	22
Commercial Services & Supplies	0.80	2.11	5.02	606	16
Materials	0.73	2.60	4.85	2,877	67
Capital Goods	0.69	2.31	4.60	4,632	121
Semiconductors & Semiconductor Equipment	0.63	2.16	4.40	1,008	22
Consumer Services	0.62	2.34	4.28	1,142	30

Panel B: Climate Transition Risk

Sector	Mean (%)	Std.Dev. (%)	P95 (%)	N Calls	N Firms
Utilities	0.18	1.14	0.67	1,155	32
Energy	0.14	0.95	0.00	1,097	25
Real Estate Management & Development	0.09	0.66	0.00	559	15
Materials	0.07	0.65	0.00	2,877	67
Food & Staples Retailing	0.06	0.55	0.00	1,034	25
Media & Entertainment	0.06	0.46	0.00	833	22
Automobiles & Components	0.06	0.44	0.00	725	15
Equity REITs	0.05	0.64	0.00	785	26
Consumer Services	0.05	0.52	0.00	1,142	30
Consumer Durables & Apparel	0.05	0.49	0.00	1,828	45

Panel C: Climate Physical Risk

Sector	Mean (%)	Std.Dev. (%)	P95 (%)	N Calls	N Firms
Automobiles & Components	0.05	0.50	0.00	725	15
Insurance	0.04	0.33	0.00	4,315	103
Real Estate Management & Development	0.03	0.39	0.00	559	15
Utilities	0.03	0.34	0.00	1,155	32
Media & Entertainment	0.03	0.25	0.00	833	22
Consumer Services	0.03	0.26	0.00	1,142	30
Health Care Equipment & Services	0.02	0.31	0.00	770	17
Diversified Financials	0.02	0.30	0.00	2,153	62
Capital Goods	0.02	0.28	0.00	4,632	121
Materials	0.02	0.17	0.00	2,877	67

Panel D: Climate Regulation Risk

Sector	Mean (%)	Std.Dev. (%)	P95 (%)	N Calls	N Firms
Utilities	0.18	1.13	0.74	1,155	32
Energy	0.09	0.79	0.00	1,097	25
Automobiles & Components	0.08	0.63	0.00	725	15
Real Estate Management & Development	0.07	0.60	0.00	559	15
Media & Entertainment	0.07	0.54	0.00	833	22
Semiconductors & Semiconductor Equipment	0.07	0.61	0.00	1,008	22
Consumer Durables & Apparel	0.07	0.55	0.00	1,828	45
Consumer Services	0.06	0.58	0.00	1,142	30
Insurance	0.06	0.61	0.00	4,315	103
Materials	0.05	0.58	0.00	2,877	67

Table 5–continued

Panel E: Climate Innovation Opportunity					
Sector	**Mean (%)**	**Std.Dev. (%)**	**P95 (%)**	**N Calls**	**N Firms**
Utilities	0.85	2.61	5.30	1,155	32
Automobiles & Components	0.47	1.74	3.19	725	15
Commercial Services & Supplies	0.46	1.47	3.90	606	16
Energy	0.43	1.55	2.93	1,097	25
Real Estate Management & Development	0.37	1.36	2.28	559	15
Media & Entertainment	0.37	1.47	2.35	833	22
Software & Services	0.34	1.62	1.64	1,005	21
Materials	0.32	1.56	1.51	2,877	67
Capital Goods	0.29	1.35	1.75	4,632	121
Semiconductors & Semiconductor Equipment	0.26	1.19	1.58	1,008	22

VII Green Innovation, R&D and Capital Expenditures

Does corporate climate discourse predict green innovation, or is it cheap talk? This section tests whether firms that emphasize climate-related risks and opportunities in earnings calls subsequently produce more green patents than sector peers with comparable innovative capacity. The central question is whether climate attention reflects genuine strategic reorientation toward green technologies, or merely performative disclosure.

We document three main findings. First, attention to climate opportunities and transition risks is positively associated with future green patenting, while attention to physical climate risk is not. Within the same industry and year, firms that identify and emphasize climate opportunities or transition risks file more green patents in the subsequent year, conditional on long-run innovative capacity, emission intensity, and other firm characteristics. Second, climate attention amplifies the green patent output associated with R&D spending: firms that both identify climate opportunities or risks and invest heavily in R&D produce disproportionately more green patents than either factor alone would predict. This nonlinear relationship suggests that climate narratives accompany genuine reallocation or expansion of innovative effort toward green technologies, rather than cheap talk. Third, exploiting within-firm variation over time, we find that elevated climate attention precedes increased R&D intensity, consistent with climate positioning reflecting forward-looking investment commitments rather than ex-post rationalization.

A. Baseline Specification: Climate Attention and Green Patenting

Green innovation is measured using firm-level counts of yearly green patent applications filed in year $t + 1$, with climate attention measured in year t. This timing reflects that

patenting is a lagged outcome of research activity initiated in response to strategic priorities revealed in corporate disclosures. Since patent data are available only annually, we aggregate quarterly climate attention scores to firm-year measures by taking the mean across all earnings calls within each calendar year.

Because the dependent variable is a non-negative count with many zeros and a right-skewed distribution, we estimate the model using Poisson pseudo-maximum likelihood (PPML). [3]

A distinct challenge is distinguishing shifts in the direction of innovation toward green technologies from differences in firms' overall innovative capacity. More innovative firms produce more patents of all types, including green patents. Additionally, firms that patented green technologies in the past are more likely to do so again, both because research programs exhibit path dependence and because green innovation capacity accumulates over time. To address these characteristics of the data, all specifications control for both lagged green patenting, which captures persistence and path dependence, and patent stock, a perpetual-inventory measure of accumulated innovative capacity. Together, they allow us to ask whether climate attention predicts additional green innovation conditional on a firm's history of green patenting and its overall innovative capacity.

Patent stock is constructed by applying the perpetual-inventory method to total patent applications, both green and non-green. For firm i:

$$\text{PatentStock}_{i,t} = (1 - \delta)\,\text{PatentStock}_{i,t-1} + \text{TotalPatents}_{i,t},$$

where $\text{TotalPatents}_{i,t}$ denotes the flow of all patent applications filed in year t, and $\delta = 0.15$ is an annual depreciation rate capturing obsolescence of technological knowledge, following Hall (2007). This measure summarizes accumulated technological capabilities, research infrastructure, and innovative know-how, independent of the environmental orientation of innovation.

The baseline specifications further include industry-by-year fixed effects, which absorb sector-level technological trajectories, regulatory shocks, and time-varying market conditions

[3] PPML provides consistent estimates under the weaker assumption that the conditional mean is correctly specified, without requiring that the data follow a true Poisson distribution (Silva and Tenreyro, 2006). This robustness is valuable given that patent counts typically exhibit overdispersion. PPML also handles zeros without requiring ad hoc transformations, which can introduce bias and avoids the retransformation problem of log-linearized OLS.

affecting green innovation incentives. Identification thus relies on within-industry-year variation: among firms in the same sector facing identical external conditions in a given year, do those emphasizing climate opportunities or risks innovate more in green technologies? This design addresses the concern that green patenting and climate attention could both respond to common industry-level drivers, such as anticipated regulation or technological breakthroughs, without one causing the other. By comparing firms within narrow industry-year cells, we isolate firm-specific variation in climate attention from confounding sectoral trends.

For each attention dimension $A_{i,t} \in \{\text{Opportunity}, \text{Transition Risk}, \text{Physical Risk}\}$, the baseline specification is:

$$E[\text{GreenPatents}_{i,t+1} \mid \mathcal{I}_{i,t}] = \exp\big(\beta A_{i,t} + \rho\, \text{GreenPatents}_{i,t} + \kappa\, \text{PatentStock}_{i,t} + \delta' \mathbf{X}_{i,t} + \gamma_{j\times t}\big)\,, \tag{3}$$

where $\mathcal{I}_{i,t} = \{A_{i,t}, \text{GreenPatents}_{i,t}, \text{PatentStock}_{i,t}, \mathbf{X}_{i,t}, \gamma_{j\times t}\}$ and $\gamma_{j\times t}$ denotes industry-by-year fixed effects.

The vector of firm-level controls $\mathbf{X}_{i,t}$ includes R&D intensity (R&D/assets), capital expenditure intensity (CAPEX/assets in USD), cash holdings (cash/assets in USD), headline return on equity (%), firm size (log total assets in USD), and carbon emission intensity (log Scope 1–2 emissions per million USD revenue). These controls absorb variation in investment intensity, financial slack, profitability, size, and carbon risk exposure that might independently predict both climate attention and green patenting.

Coefficients on the climate attention variables represent semi-elasticities: a coefficient of β on a standardized attention measure implies that a one-standard-deviation increase is associated with a $100 \times (e^{\beta} - 1)$ percent change in expected green patent counts. Standard errors are clustered at the firm level to account for serial correlation in patenting outcomes within firms.

Table 8 reports baseline results. Columns (2) through (4) introduce each climate attention dimension separately. Opportunity attention enters with a coefficient of 0.23, implying that a one-standard-deviation increase is associated with approximately 26 percent more green patent applications in the following year. Transition risk attention is also positive and significant, though smaller in magnitude. Physical risk attention, by contrast, enters

negatively, a result we explore further below.

B. R&D Expenditures and the Direction of Investment

The baseline results establish that climate attention predicts green patenting conditional on innovative capacity. But do firms that emphasize climate opportunities and risks actually allocate investment differently, or does climate attention simply proxy for unobserved firm characteristics correlated with both disclosure and innovation? This subsection tests whether climate attention moderates the relationship between investment inputs and green innovation outputs.

The intuition is straightforward. If climate attention reflects cheap talk, then the mapping from R&D and capital expenditures to green patents should be similar across firms regardless of their climate narratives. If, however, climate attention signals genuine strategic reorientation, then a given level of investment spending should translate into more green innovation among firms that emphasize climate opportunities or transition risks. We test this by interacting climate attention with investment intensity.

We augment the baseline PPML specification with R&D intensity (standardized) and its interaction with climate attention:

$$E[\text{GreenPatents}_{i,t+1} \mid \mathcal{I}_{i,t}] = \exp\big(\beta_1 A_{i,t} + \beta_2 \text{R\&D}_{i,t} + \beta_3 A_{i,t} \times \text{R\&D}_{i,t} + \mathbf{Z}'_{i,t} + \gamma_{j\times t}\big)\,,$$

where $A_{i,t}$ denotes either opportunity or transition risk attention, $\mathbf{Z}_{i,t}$ includes lagged green patenting, patent stock, and standard firm controls, and $\gamma_{j\times t}$ denotes industry-by-year fixed effects. The coefficient β_2 captures the baseline effect of R&D intensity on green patenting, while β_3 tests whether this effect varies with climate attention.

Table 9 reports results. The interaction between opportunity attention and R&D is positive and significant (0.18, s.e. 0.06), indicating that the marginal product of R&D spending in generating green patents is higher among firms that emphasize climate opportunities. The interaction with transition risk attention is even larger (0.50, s.e. 0.07). These estimates imply economically meaningful heterogeneity: among firms at the 75th percentile of opportunity attention, a one-standard-deviation increase in R&D is associated with roughly 65 percent more green patents, compared to 37 percent for firms at the 25th percentile.

Table 8: Semantic Climate Scores and Green Innovation: Baseline Results

Notes: Poisson pseudo-maximum likelihood estimates on an unbalanced firm-year panel (2009–2025). The dependent variable is the count of green patent applications filed in year $t+1$. Climate attention scores are standardized to have mean zero and unit variance. All specifications include industry-by-year fixed effects. Standard errors clustered at the firm level are reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.10$.

Dependent variable	Green Patents$_{t+1}$			
	(1) Baseline	(2) Opportunity	(3) Transition	(4) Physical
Opportunity score (std.)		0.23***		
		(0.07)		
Transition risk score (std.)			0.11***	
			(0.03)	
Physical risk score (std.)				-0.29**
				(0.14)
Green Patents (1000s)	1.30**	1.30**	1.30**	1.30**
	(0.61)	(0.61)	(0.61)	(0.61)
Log Assets	0.38**	0.38**	0.38**	0.37**
	(0.16)	(0.16)	(0.16)	(0.16)
Scope 1–2 intensity (log)	-0.12	-0.12	-0.11	-0.11
	(0.11)	(0.11)	(0.11)	(0.11)
CAPEX / Assets	8.60***	8.60***	8.58***	8.54***
	(3.22)	(3.22)	(3.20)	(3.24)
Cash / Assets	-0.85	-0.85	-0.87	-0.96
	(0.87)	(0.87)	(0.88)	(0.91)
Cash / Assets (missing)	1.97*	1.97*	1.96*	1.94*
	(1.08)	(1.08)	(1.08)	(1.07)
Total Patents (1000s)	0.02	0.02	0.02	0.00
	(0.07)	(0.07)	(0.07)	(0.07)
R&D missing	-2.19***	-2.19***	-2.19***	-2.19***
	(0.25)	(0.25)	(0.24)	(0.25)
Headline ROE	-0.00	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Intercept	-0.42	-0.42	-0.47	-0.45
	(1.72)	(1.72)	(1.73)	(1.73)
Observations	9,859	9,859	9,859	9,859
Industry-year FE	Yes	Yes	Yes	Yes
Estimator	Poisson Pseudo-Maximum Likelihood			

This pattern is difficult to reconcile with cheap talk. If climate narratives were merely performative, we would expect no systematic relationship between attention and the productivity of R&D in generating green innovation. The positive interactions suggest instead that climate attention accompanies, genuine differences in how firms allocate research effort across green and non-green technologies.

A complementary question is whether climate attention predicts changes in overall R&D intensity, not just its composition. Table 10 reports within-firm regressions of R&D intensity on climate attention, exploiting variation over time within the same firm. Opportunity and transition risk attention both positively predict subsequent R&D spending, with coefficients of 0.03 and 0.02 respectively. Physical risk attention shows no association. These results suggest that firms increase total R&D investment when they perceive climate-related opportunities or transition risks, reinforcing the interpretation that climate attention reflects substantive strategic shifts rather than disclosure management.

Taken together, the interaction and within-firm results point to the same conclusion: climate attention is associated with both the scale and direction of innovative investment. Firms that emphasize climate opportunities and transition risks spend more on R&D and convert that spending into green patents more efficiently. Physical risk attention, by contrast, neither predicts R&D intensity nor moderates the R&D-to-patents relationship, consistent with the baseline finding that physical risk operates primarily as a contextual factor rather than a direct innovation driver.

C. Summary of Findings

Three findings emerge from this analysis.

First, climate attention predicts green innovation, but not all dimensions equally. Opportunity and transition risk attention are positively and significantly associated with subsequent green patenting, while physical risk attention is not. A one-standard-deviation increase in opportunity attention corresponds to approximately 26 percent more green patent applications in the following year; for transition risk attention, the figure is roughly 12 percent. These estimates are conditional on industry-year fixed effects and controls for innovative capacity, so they reflect within-sector, within-year variation across firms with comparable patent stocks and R&D intensity. Physical risk attention, by contrast, enters negatively and

Table 9: Climate Attention, R&D Investment, and Green Innovation

	(1) Opportunity × R&D	(2) Transition Risk × R&D
Dependent variable	Green Patents$_{t+1}$	
Opportunity score (std.)	0.24***	
	(0.07)	
Transition risk score (std.)		0.12***
		(0.03)
R&D (std.)	0.32***	0.37***
	(0.11)	(0.10)
Opportunity × R&D	0.18***	
	(0.06)	
Transition risk × R&D		0.50***
		(0.07)
Lagged green patents	0.00***	0.00***
	(0.00)	(0.00)
Patent stock	-0.00	-0.00
	(0.00)	(0.00)
Log Assets	0.48***	0.48***
	(0.16)	(0.16)
CAPEX / Assets	6.43*	6.51*
	(3.80)	(3.81)
Scope 1–2 intensity (log)	-0.09	-0.08
	(0.09)	(0.10)
Headline ROE	-0.00	-0.00
	(0.00)	(0.00)
Intercept	-2.03	-2.09
	(1.65)	(1.65)
Observations	9,859	9,859
Industry-year FE	Yes	Yes
Estimator	Poisson Pseudo-Maximum Likelihood	

Notes: Poisson pseudo–maximum likelihood estimates. Standard errors clustered at the firm level are reported in parentheses. Opportunity, transition risk, and R&D variables are standardized to have mean zero and unit variance. The main effect of R&D intensity is omitted when collinear with the interaction term. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 10: Within-Firm Effects of Climate Attention on R&D Investment

	(1) Opportunity	(2) Transition Risk	(3) Physical Risk
Dependent variable	R&D / Assets		
Opportunity score	0.03***		
	(0.01)		
Transition risk score		0.02**	
		(0.01)	
Physical risk score			-0.00
			(0.01)
Lagged green patents	0.00**	0.00**	0.00**
	(0.00)	(0.00)	(0.00)
Patent stock	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
Log Assets	-0.01***	-0.01***	-0.01***
	(0.00)	(0.00)	(0.00)
Cash / Assets	-0.01**	-0.01**	-0.01**
	(0.01)	(0.01)	(0.01)
CAPEX / Assets	0.04***	0.04***	0.04***
	(0.01)	(0.01)	(0.01)
Scope 1–2 intensity (log)	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)
Headline ROE	0.00***	0.00***	0.00***
	(0.00)	(0.00)	(0.00)
Intercept	0.08***	0.08***	0.08***
	(0.01)	(0.01)	(0.01)
Observations	9,084	9,084	9,084
Firm fixed effects	Yes	Yes	Yes
Estimator	Linear regression (within-firm)		

Notes: Linear fixed-effects regressions exploiting within-firm variation only. All specifications include firm fixed effects, and standard errors are clustered at the firm level. The dependent variable is R&D expenditures scaled by total assets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

does not predict green innovation directly.

Second, climate attention affects both the scale and direction of innovative investment. Firms that emphasize climate opportunities or transition risks not only patent more green technologies but also spend more on R&D and convert that spending into green patents more efficiently. The interaction between climate attention and R&D intensity is positive and significant for both opportunity and transition risk dimensions, indicating that the marginal product of research spending in generating green patents is higher among firms with stronger climate narratives. This pattern is inconsistent with cheap talk: if climate attention were merely performative, the productivity of R&D in generating green innovation should not vary systematically with disclosure.

Third, different attention dimensions interact in economically meaningful ways. Opportunity and transition risk attention appear to operate as substitutes rather than complements: the marginal effect of opportunity attention diminishes among firms that also emphasize transition risks. One interpretation is that these represent alternative strategic framings of the same underlying green innovation activity. A firm repositioning toward clean technologies might frame this as capturing new market opportunities or as managing regulatory and competitive risks, but the innovation itself is similar. Physical risk attention, while not directly associated with green patenting, amplifies the effect of transition risk attention. Firms exposed to tangible climate threats appear to translate regulatory and market concerns into technological investment more readily than firms for whom climate change remains abstract.

The overall picture is one in which corporate climate narratives contain economically meaningful information about innovation trajectories. Firms that discuss climate opportunities and transition risks subsequently produce more green patents and invest more in R&D, even after conditioning on existing innovative capacity and absorbing industry-year shocks. The association is predictive rather than causal, but it suggests that earnings call disclosures reveal genuine heterogeneity in climate-oriented strategy rather than uniform greenwashing.

VIII Conclusion

This paper introduces a semantic search framework that overcomes a fundamental trade-off between scalability and precision in text-as-data economics. Our two-stage architecture combines efficient semantic retrieval with optional LLM refinement, enabling researchers to

measure nuanced economic concepts from unstructured text without requiring labeled training data. The framework delivers granular, hierarchical measurement—distinguishing not just transition risk from physical risk, but specific channels like carbon pricing mechanisms, technology shifts, and regulatory compliance within each category. While direct LLM classification of large corpora quickly escalates in computational and financial costs, making it unsuitable for exploratory analysis or emerging domains, our approach processes millions of text passages into structured firm-quarter panel data with minimal resource constraints. Moreover, the underlying FAISS index architecture can efficiently handle corpora exceeding 100 million paragraphs, making the framework suitable for comprehensive analyses of entire firm universes over extended time horizons.

Applied to climate risk measurement, the framework produces economically meaningful firm-level indicators with substantial cross-sectional and temporal variation. Our climate attention measures predict real outcomes: firms emphasizing climate opportunities file 26% more green patents, while those discussing transition risks convert R&D spending into green innovation more efficiently. These findings move beyond static ESG scores to capture dynamic, forward-looking risk perceptions that translate into tangible corporate actions. The measures also reveal how narrative-based exposure conditions market responses, as demonstrated by the amplified carbon penalty for high-emission firms perceived to face significant transition risk after the Paris Agreement.

The framework's generalizability extends far beyond climate applications. By replacing rigid keyword dictionaries with flexible semantic queries, researchers can instantly adapt the methodology to any emerging risk—whether artificial intelligence adoption, geopolitical tensions, or supply chain vulnerabilities—without costly retraining or annotation. This capability transforms narrative-based empirical research from a niche methodology constrained to small samples into a broadly applicable tool for large-scale economic analysis. The framework enables new research designs, including firm-level information shock identification and narrative-driven event studies that isolate the role of belief updating in corporate decisions and asset pricing.

Several promising directions emerge for future methodological development. Query design, while requiring initial care, can increasingly be automated through clustering-based discovery, dynamic expansion based on retrieved content patterns, and temporal adaptation

mechanisms that flag when revalidation is needed. These advances would further reduce the domain expertise required for effective deployment while enabling systematic exploration of risk dimensions that researchers have not explicitly considered. Extensions to other languages, smaller firms, and different document types would broaden the framework's applicability across diverse economic contexts.

References

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. The quarterly journal of economics, 131(4):1593–1636.

Cookson, J. A., Engelberg, J. E., and Mullins, W. (2023). Echo chambers. The Review of Financial Studies, 36(2):450–500.

de Araujo, D. K. G., Bokan, N., Comazzi, F. A., and Lenza, M. (2025). Word2Prices: embedding central bank communications for inflation prediction. Number 3047. ECB Working Paper.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J., Doran, C., and Solorio, T., editors, 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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3):535–574.

Gu, W., hao Zhong, Y., zun Li, S., song Wei, C., ting Dong, L., yue Wang, Z., and Yan, C. (2024). Predicting stock prices with finbert-lstm: Integrating news sentiment analysis. In Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing, pages 67–72.

Hall, B. H. (2007). Measuring the returns to r&d: The depreciation problem.

Hassan, T. A., Hollander, S., Kalyani, A., van Lent, L., Schwedeler, M., and Tahoun, A.

(2025). Text as data in economic analysis. Journal of Economic Perspectives, 39(3):193–220.

Hassan, T. A., Hollander, S., Van Lent, L., and Tahoun, A. (2019). Firm-level political risk: Measurement and effects. The Quarterly Journal of Economics, 134(4):2135–2202.

Huang, A. H., Wang, H., and Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. Contemporary Accounting Research, 40(2):806–841. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1911-3846.12832.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with gpus. IEEE Transactions on Big Data, 7(3):535–547.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781. Conference Name: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) Place: Online Publisher: Association for Computational Linguistics.

Manski, C. F. (2018). Survey measurement of probabilistic macroeconomic expectations: progress and promise. NBER Macroeconomics Annual, 32(1):411–471.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs].

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In Moschitti, A., Pang, B., and Daelemans, W., editors, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019a). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019b). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Sautner, Z., Van Lent, L., Vilkov, G., and Zhang, R. (2023). Firm-level climate change exposure. The Journal of Finance, 78(3):1449–1498.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2022). Measuring news sentiment. Journal of econometrics, 228(2):221–243.

Shiller, R. J. (2020). Narrative economics: How stories go viral and drive major economic events.

Silva, J. S. and Tenreyro, S. (2006). The log of gravity. The Review of Economics and statistics, pages 641–658.

Silva, T. C., Moriya, K., Veyrune, M. R. M., and Veyrune, R. (2025). From text to quantified insights: a large-scale llm analysis of central bank communication. International Monetary Fund.

Task Force on Climate-related Financial Disclosures (2017). Recommendations of the task force on climate-related financial disclosures. Technical report, Financial Stability Board, Basel, Switzerland. Final Report.

Webersinke, N., Kraus, M., Bingler, J., and Leippold, M. (2022). CLIMATEBERT: A Pretrained Language Model for Climate-Related Text. SSRN Electronic Journal.