UNIVERSITÀ
DI PAVIA

Department of Economics and Management

Master Programme in International Business &

Entrepreneurship - Digital Management

LARGE LANGUAGE MODELS AS ESG SENTIMENT GENERATORS: DATASET CONSTRUCTION AND ANALYSIS

Supervisor:

Prof.ssa Paola Cerchiello

Master Thesis by
Alice Martino

Matr. n. 54261

Academic Year 2025-2026

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREMISE

This thesis uses artificial intelligence, and in particular Large Language Models (LLMs), as tools for generating synthetic ESG sentiment proxies from unstructured textual input. The resulting dataset is not directly observed in external sources, but constructed through interaction with LLMs, which transform textual prompts and contextual cues into structured indicators. The analyses rely on the prompt-based construction of ESG metrics and should therefore be read as model-dependent reconstructions rather than objective observations. Their informational content may be partial, biased, distorted, or occasionally affected by hallucinated outputs. For this reason, the use of artificial intelligence is limited to the data generation phase, while research design, analysis and interpretation of the results remain entirely responsibility of the author.

ABSTRACT (ENGLISH VERSION)

This thesis examines the potential of Large Language Models as generators of synthetic ESG sentiment and evaluates whether the resulting signal can function as a meaningful informational variable in finance. The empirical analysis is based on a daily panel of 14,640 observations spanning eight European countries, France, Germany, Italy, Norway, Poland, Portugal, Spain and the United Kingdom, over the full calendar year 2024. Using a standardized zero-shot prompting framework, the study generates ESG sentiment scores with five different LLMs: GPT-5.2, Claude Sonnet 4.5, Grok 4.1, Sonar and Gemini 3 Pro. The resulting dataset is assessed through descriptive statistics, variance analysis, correlation analysis, ordinary least squares regressions, and principal component analysis in order to evaluate the internal coherence of the synthetic series and its relationship with financial market returns. The empirical evidence indicates that LLM-generated ESG sentiment is not homogeneous across providers, but instead varies substantially in terms of volatility, structural consistency, and latent cross-country organization. Furthermore, the association between aggregated ESG sentiment and financial market returns appears limited and loses explanatory strength once broader market dynamics are considered, suggesting that the signal captures a distinct informational dimension rather than a direct short-term market driver. The original contribution of the thesis lies in constructing and testing a comparative framework for synthetic ESG sentiment generation through LLMs, thereby offering a novel methodological perspective on the use of generative AI in sustainability-related financial research.

ABSTRACT (ITALIAN VERSION)

Questa tesi esamina il potenziale dei Large Language Models come generatori di ESG sentiment sintetico e valuta se il segnale risultante possa funzionare come una variabile informativa significativa in ambito finanziario. L'analisi empirica si basa su un panel giornaliero di 14.640 osservazioni, relativo ad otto Paesi europei, Francia, Germania, Italia, Norvegia, Polonia, Portogallo, Spagna e Regno Unito, lungo l'intero anno solare 2024. Attraverso un framework di prompting standardizzato in modalità zero-shot, lo studio genera punteggi di ESG sentiment con cinque diversi LLM: GPT-5.2, Claude Sonnet 4.5, Grok 4.1, Sonar e Gemini 3 Pro. Il dataset ottenuto viene analizzato mediante statistiche descrittive, analisi della varianza, analisi delle correlazioni, regressioni lineari OLS e analisi delle componenti principali, al fine di valutare sia la coerenza interna della serie sintetica, sia la sua relazione con i rendimenti dei mercati finanziari. Le evidenze empiriche indicano che l'ESG sentiment generato dai LLM non è omogeneo tra i diversi provider, ma varia in modo sostanziale in termini di volatilità, coerenza strutturale e organizzazione latente tra Paesi. Inoltre, l'associazione tra ESG sentiment aggregato e rendimenti di mercato appare limitata e perde forza esplicativa una volta considerata la dinamica più ampia del mercato, suggerendo che il segnale catturi una dimensione informativa distinta, piuttosto che un driver diretto delle variazioni di breve periodo. Il contributo originale della tesi consiste nella costruzione e nella verifica di un framework comparativo per la generazione di ESG sentiment sintetico tramite LLM, offrendo una prospettiva metodologica innovativa sull'impiego dell'intelligenza artificiale generativa nella ricerca finanziaria legata alla sostenibilità.

INTRODUCTION

This thesis is situated within the growing debate on sustainability and its role in contemporary financial systems. In recent years, the increasing relevance of climate change, social inequality and corporate governance has made it clear that financial indicators alone are no longer sufficient to describe the behaviour and positioning of firms, institutions, and markets. Within this context, ESG has progressively emerged as one of the main frameworks through which sustainability is translated into observable information that can also be used in finance. At the same time, however, the diffusion of ESG has also highlighted its main limitations. ESG ratings and scores are not based on a single, universally shared methodology, but depend on different assumptions, selection criteria and weighting schemes. As a result, providers such as MSCI, Sustainalytics, and Refinitiv may generate evaluations that refer to the same object but are not fully comparable. This is not simply a technical issue, but a broader methodological problem, since it affects the way ESG is used in research, investment practice, and financial interpretation.

It is from these limitations that the present research takes shape. If traditional ESG measures provide only a partial and method-dependent picture, it becomes relevant to ask whether other instruments may capture sustainability-related information in a more flexible way. From this perspective, the thesis focuses on ESG sentiment, understood as a signal derived from textual and narrative information related to environmental, social, and governance issues rather than from static ratings alone. In this sense, ESG sentiment may offer a more dynamic representation of how sustainability-related information is

produced, interpreted, and absorbed over time. The growing development of Large Language Models makes this perspective particularly relevant. These models have expanded the possibility of processing large volumes of unstructured data, including news, reports, public documents, and other textual materials. The interest of this thesis therefore lies not only in their technical capacity to generate text, but also in their possible role as tools for constructing synthetic indicators from fragmented and widely dispersed information. The research asks whether Large Language Models can be used to generate an ESG sentiment indicator that is coherent, sufficiently stable, and informative from an empirical point of view.

On this basis, the study develops a comparative framework for the construction and analysis of synthetic ESG sentiment at the country level. Methodologically, it relies on a standardized zero-shot prompting framework applied to five different Large Language Models in order to generate daily ESG sentiment scores for eight European countries during the full year 2024. The resulting outputs are then organized into a final panel dataset and examined through statistical analysis.

The thesis is structured in four chapters. Chapter 1 presents the theoretical framework by defining ESG, clarifying its role in the broader sustainability debate, and discussing the main limitations of traditional ESG measurement, before introducing the shift from ESG metrics to ESG sentiment. Chapter 2 focuses on Large Language Models and their relevance for the processing of unstructured information. Chapter 3 presents the research design, the sentiment generation process, the sources considered, and the construction

of the final dataset. Finally, Chapter 4 develops the empirical analysis and discusses the main results and their implications.

CHAPTER 1 – ESG & ESG SENTIMENT: A THEORETICAL FRAMEWORK

1.1 - Definition and ESG Conceptual Framework

In recent decades, sustainability has become a central issue in economic and financial systems, partly in response to growing concerns about climate change, social inequality, and corporate governance practices.[1] Consequently, traditional methods of assessing corporate performance, based exclusively on financial indicators, have proven insufficient to capture the broader impact of business activities. In this context, the concept of Environmental, Social and Governance (ESG) has established itself as a fundamental framework for measuring non-financial performance and for integrating sustainability into investment decisions and corporate management.

The concept of Environmental, Social and Governance (ESG) was popularized in 2004 by the report "*Who Cares Wins: Connecting Financial Markets to a Changing World"* from the United Nations Global Compact, drafted in collaboration with major financial institutions at the invitation of Secretary-General Kofi Annan. There is no single, legally binding definition of ESG; however, a widely used working definition identifies it as a set of environmental, social, and governance criteria used to assess an organization's impact on the environment and society, as well as the quality of its governance system, alongside traditional financial indicators.[2]

---

[1] Boffo, R., & Patalano, R. (2020). *ESG investing: Practices, progress and challenges*. Paris: OECD Publishing. https://www.oecd.org/finance/ESG-Investing-Practices-Progress-and-Challenges.pdf

[2] IBM. (2024, January 23). What is environmental, social, and governance (ESG)? Retrieved from https://www.ibm.com/think/topics/environmental-social-and-governance

1.1.1 – ESG Pillars: Environmental, Social and Governance Dimensions

ESG performance is usually broken down into three main pillars or dimensions:

- The environmental (E) dimension focuses on how a company affects and is affected by the natural environment. It typically includes issues such as climate change, greenhouse-gas emissions, air and water pollution, waste management, energy and resource efficiency and biodiversity loss.
- The social (S) dimension concerns how a company manages its relationships with key stakeholder groups, including employees, customers, suppliers and local communities. It covers labour standards and human rights, occupational health and safety, fair and non-discriminatory treatment, training and human-capital development, consumer protection and community engagement.
- The governance (G) dimension relates to how a company is overseen and controlled. It encompasses board composition and independence, the separation of roles between chair and CEO, shareholder rights, executive remuneration, internal control and audit systems, as well as business ethics, transparency and anti-corruption policies.

Taken together, these three pillars define the substantive content of ESG and provide the basis on which data providers and rating agencies construct ESG indicators and scores, by aggregating multiple environmental, social and governance metrics into synthetic measures of a company's overall ESG performance.[3]

[3] CFA Institute. (2015). *Environmental, social, and governance issues in investing: A guide for investment professionals.* CFA Institute; ESG BC. (2025, 28 aprile). *What is ESG? Explore environmental, social, governance.* https://www.esgbc.ca/what-is-esg/

1.1.2 – ESG Frameworks and Reporting Infrastructure

ESG dimensions are operationalised through reporting frameworks that define which information firms disclose, how it is measured and how it is communicated to stakeholders. As highlighted by EcoVadis, ESG reporting can be broadly understood as the process through which firms disclose information on their environmental, social and governance performance to provide transparency on risks, impacts and non-financial outcomes.

A useful distinction can be drawn between:

- **Voluntary (non-binding) frameworks**: including the Global Reporting Initiative (GRI), Sustainability Accounting Standards Board (SASB), Task Force on Climate-related Financial Disclosures (TCFD) and the International Sustainability Standards Board (ISSB). These frameworks reflect different conceptualisations of materiality: GRI adopts a stakeholder-oriented perspective, whereas SASB and ISSB prioritise financially material information for investors and TCFD focuses specifically on climate-related governance, strategy, and risk management.
- **Mandatory (legally binding) frameworks**: most notably the Corporate Sustainability Reporting Directive (CSRD) and the European Sustainability Reporting Standards (ESRS). These introduce detailed and enforceable disclosure requirements based on the principle of double materiality, significantly expanding both the scope and depth of ESG reporting within the EU regulatory perimeter.[4]

[4] EcoVadis. (2025). *ESG reporting – Key frameworks, compliance & best practices*. EcoVadis. https://ecovadis.com/glossary/esg-reporting/

The parallel existence of these initiatives has generated a dynamic interaction of standard layering and competition, characterised by partial convergence alongside persistent divergences in objectives, scope and underlying notions of materiality. As argued by Afolabi et al. (2022), sustainability reporting regulation can be conceptualised as a "*contested arena*", in which key institutions engage in strategic interactions to shape the emerging global infrastructure of non-financial reporting.[5] Despite ongoing efforts towards harmonisation, empirical evidence indicates that ESG reporting and the ratings derived from it remain fragmented and only partially comparable. Gibson et al. (2021) show that a substantial share of ESG rating divergence stems from differences in the scope of categories included by rating agencies, in the methodologies used to evaluate similar dimensions and, in the weighting, schemes applied when aggregating indicators into composite scores. These discrepancies are closely linked to the underlying disclosure frameworks and data sources on which ratings rely.[6]

Overall, these findings highlight the need for caution when interpreting ESG scores, particularly when they are employed as informational inputs in financial models, as their meaning and comparability remain contingent on the specific reporting standards and evaluation methodologies adopted.

---

[5] Afolabi, H., Ram, R., & Rimmel, G. (2022). *Harmonization of sustainability reporting regulation: Analysis of a contested arena. Sustainability*, 14(9), 5517. https://doi.org/10.3390/su14095517

[6] Gibson, R., Krueger, P., & Schmidt, P. S. (2021). *ESG rating disagreement and stock returns. Review of Finance*, 25(5), 1315–1356 (ECGI Finance Working Paper No. 651/2020)

1.1.3 – Core Principles of Responsible Investment

At an investment practice level, the fundamental principles that form the basis of the ESG framework are defined by the Principles for Responsible Investment (PRI), an UN-backed initiative that serves as a global benchmark for incorporating ESG aspects in investment and ownership decisions. The PRI are based on six principles that that invite signatories to:

1. integrate ESG issues into investment analysis and decision-making processes;
2. be active owners and incorporate ESG issues into ownership policies and practices;
3. request appropriate disclosure on ESG issues from the entities in which they invest;
4. advance the acceptance and implementation of the principles throughout the investment industry.
5. collaborate in enhancing their effectiveness in implementing the principles.
6. communicate with the public on their activities and progress towards implementation. [7]

The PRI more generally reflects the recognition that ESG factors may have an impact on portfolio performance and as such, they should be accounted for within a long-term fiduciary view. This echoes research from CFA Institute highlighting how sustainability is becoming more integrated into investment management due to client demand and the increasing materiality of E, S and G factors in investment processes.[8]

---

[7] Principles for Responsible Investment. *What are the Principles for Responsible Investment*?. https://www.unpri.org/about-PRI/what-principles-for-responsible-investment

[8] CFA Institute (2020). *Sustainability in investment management at a pivotal juncture*. https://www.cfainstitute.org/about/press-room/2020/sustainability-in-investment-management-report

In short, the ESG pillars, reporting frameworks and PRI principles together form the framework that translates sustainability issues into visible ESG indicators and ratings that can be used in financial models as informational variables.

1.2 – Measurement of ESG Data

ESG ratings are produced by a set of specialized data providers that aggregate non-financial information into composite scores designed to inform investment decisions. These agencies differ profoundly in their conceptual foundations and scoring logics with direct consequences for the comparability of their outputs. Among the best-known are MSCI, Sustainalytics and Refinitiv (LSEG), each of which represents a distinct paradigm for thinking about what ESG performance means and how it should be measured.

MSCI is one of the largest index and financial data providers in the world and its ESG rating product is built on the premise that environmental, social and governance factors matter primarily insofar as they represent financially material risks or opportunities for a company. The fundamental question MSCI's methodology is designed to answer is not "*how responsible is this company?"* but rather *"how well positioned is this company relative to its industry peers in managing ESG-related financial risks?"*. MSCI builds its ratings around a library of 33 ESG issues, grouped into 10 thematic categories across the 3 pillars. For each industry sub-group, MSCI identifies between two to seven Key Issues, those ESG themes estimated most likely to create significant risk or opportunity for a typical company operating in that sector (see Figure 1). The selection of Key Issues

is revised annually, ensuring that industry-level materiality assessments remain aligned with evolving regulatory and market conditions.

3 Pillars	10 Themes	33 ESG Key Issues
Environment	Climate Change	Carbon Emissions
		Climate Change Vulnerability
		Financing Environmental Impact
		Product Carbon Footprint
	Natural Capital	Biodiversity & Land Use
		Raw Material Sourcing
		Water Stress
	Pollution & Waste	Electronic Waste
		Packaging Material & Waste
		Toxic Emissions & Waste
	Environmental Opportunities	Opportunities in Clean Tech
		Opportunities in Green Building
		Opportunities in Renewable Energy
Social	Human Capital	Health & Safety
		Human Capital Development
		Labor Management
		Supply Chain Labor Standards
	Product Liability	Chemical Safety
		Consumer Financial Protection
		Privacy & Data Security
		Product Safety & Quality
		Responsible Investment
	Stakeholder Opposition	Community Relations
		Controversial Sourcing
	Social Opportunities	Access to Finance
		Access to Health Care
		Opportunities in Nutrition & Health
Governance	Corporate Governance	Board
		Pay
		Ownership & Control
		Accounting
	Corporate Behavior	Business Ethics
		Tax Transparency

Figure 1 - MSCI ESG Rating Framework.

Source: MSCI (2024), MSCI ESG Ratings Methodology, Exhibit 1

For each Key Issue, a company receives two distinct scores: an *Exposure Score*, derived from over 80 business and geographic vulnerability factors (including revenue concentration in high-risk activities, geographic exposure to climate-sensitive regions or reliance on scarce natural resources) and a *Management Score*, based on more than 150 policy, programme and performance indicators assessing how effectively the company addresses that exposure. The final Key Issue score is computed as the net of exposure and management and Key Issue scores are then aggregated into a final pillar score and an overall weighted ESGE score, which is ultimately mapped into a seven-point letter scale ranging from AAA (leader) to CCC (laggard), as shown in Figure 2.

Letter Rating	Leader/Laggard	Final Industry-Adjusted Company Score
AAA	Leader	8.571* - 10.0
AA	Leader	7.143 – 8.571
A	Average	5.714 – 7.143
BBB	Average	4.286 – 5.714
BB	Average	2.857 – 4.286
B	Laggard	1.429 – 2.857
CCC	Laggard	0.0 – 1.429

Figure 2 - Mapping of the Final Industry-Adjusted Company Score into the MSCI seven-point letter rating scale

Source: MSCI (2024), MSCI ESG Ratings Methodology, Exhibit 2

A defining structural feature of MSCI's methodology is *peer normalisation*: company scores are not evaluated against an absolute standard of conduct but are benchmarked relative to the distribution of scores within each industry peer group. This means that a company rated AA in one sector may exhibit objectively worse ESG practices than a company rated BB in another sector with a higher average level of ESG management. This design choice reflects a deliberate portfolio management logic, but it simultaneously prevents cross-industry comparisons and makes MSCI ratings a relative, rather than absolute, measure of ESG performance.[9]

Sustainalytics, a subsidiary of Morningstar since 2020, takes a structurally different approach. Where MSCI asks how well a company manages its ESG risks relative to peers, Sustainalytics asks how much unmanaged ESG risk a company still carries in absolute terms. The conceptual shift is significant: Sustainalytics' metric is not a performance ranking but an estimate of residual financial exposure. This distinction is further clarified in Figure 3, which illustrates the three building blocks of the Sustainalytics ESG Risk Rating.

[9] MSCI (2024). *MSCI ESG Ratings Methodology*. https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Ratings+Methodology.pdf

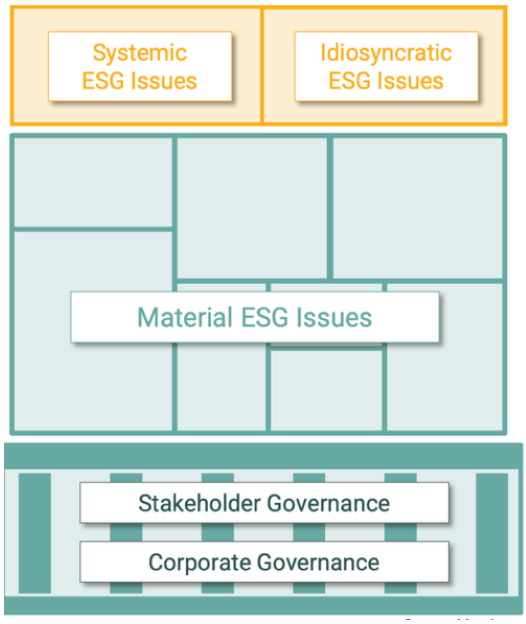


Figure 3 - The Three Building Blocks of the Sustainalytics ESG Risk Rating

Source: Sustainalytics (2024), ESG Risk Ratings Methodology v3.1, Exhibit 1.

The methodology is built around a two-component model. For each company, Sustainalytics first estimates Total ESG Risk Exposure, the aggregate level of risk to which the business model is inherently subject, prior to any management action. Exposure is driven by both Corporate Governance Risk (a baseline component applied uniformly across all companies, reflecting structural governance vulnerabilities) and Material ESG Risk (a company-specific and industry-specific component determined by 20–30 Material ESG Issues relevant to the firm's operating context). Exposure is thus a function of what a company does and where it operates, independently of how it is managed. In this sense, exposure depends on what a company does and where it operates, rather than on how it is managed, as illustrated in Figure 4.

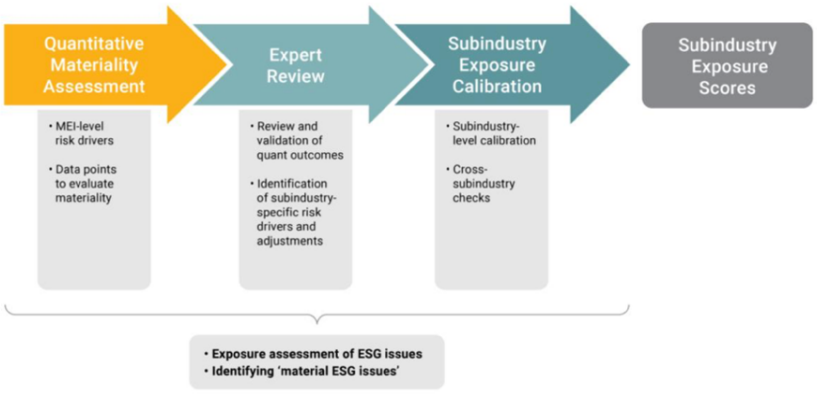


Figure 4 - The Sustainalytics Subindustry Exposure Score Assessment Process

Source: Sustainalytics (2024), ESG Risk Ratings Methodology v3.1, Exhibit 3

The second component is ESG Risk Management, which captures the extent to which the company has implemented policies, programmes and verified performance indicators that reduce its exposure. Management scores are derived from disclosed data, third-party audits and controversy signals. The final Unmanaged Risk score, the metric that Sustainalytics publishes, is computed as the portion of total exposure that management programmes have failed to address; higher unmanaged risk implies greater residual financial vulnerability. Scores are placed on a continuous scale from 0 to 40+ and then classified into five categorical risk levels: Negligible (0–10), Low (10–20), Medium (20–30), High (30–40) and Severe (40+). Because Sustainalytics' output is an absolute risk estimate rather than a peer-relative ranking, it is particularly suited to fixed-income analysis, credit risk assessment and engagement-oriented investment strategies where the concern is not relative positioning but the identification of firms with material uncovered liabilities. However, this absolute framing also means that comparisons across sectors are more direct, at the cost of not rewarding companies that outperform peers that operate in high-risk industries.[10]

[10] Sustainalytics (2024). *ESG Risk Ratings Methodology, Version 3.1.* https://www.sustainalytics.com/docs/knowledgehublibraries/default-document-library/sustainalytics_-esg-risk-ratings_-version-3-1_-methodology-abstract_-june-2024.pdf

Refinitiv, now part of the London Stock Exchange Group (LSEG), operates on a third and distinct logic. Its ESG scoring framework is primarily driven by data availability and controversy exposure, making it structurally closer to a transparency and reputational assessment than to a risk-adjusted financial measure. In Figure 6, Refinitiv's ESG score architecture is presented, highlighting how multiple datapoints are aggregated into the final score and mapped into a scale ranging from ESG laggards to ESG leaders.

Score range	Grade	Description
0.0 <= score <= 0.083333	D -	'D' score indicates poor relative ESG performance and insufficient degree of transparency in reporting material ESG data publicly.
0.083333 < score <= 0.166666	D	
0.166666 < score <= 0.250000	D +	
0.250000 < score <= 0.333333	C -	'C' score indicates satisfactory relative ESG performance and moderate degree of transparency in reporting material ESG data publicly.
0.333333 < score <= 0.416666	C	
0.416666 < score <= 0.500000	C +	
0.500000 < score <= 0.583333	B -	'B' score indicates good relative ESG performance and above- average degree of transparency in reporting material ESG data publicly.
0.583333 < score <= 0.666666	B	
0.666666 < score <= 0.750000	B +	
0.750000 < score <= 0.833333	A -	'A' score indicates excellent relative ESG performance and high degree of transparency in reporting material ESG data publicly.
0.833333 < score <= 0.916666	A	
0.916666 < score <= 1	A +	

Figure 5 - Refinitiv (LSEG) ESG Grade Scale

Source: LSEG (2024), LSEG ESG Scores Methodology

Refinitiv collects more than 630 individual datapoints across ten thematic pillars: three environmental pillars, three social pillars, three governance pillars and an ESG Controversies component. For each pillar, raw datapoints are normalized into percentile scores relative to the global industry peer group and then combined into three pillar group scores for E, S, and G (see Figure 6).

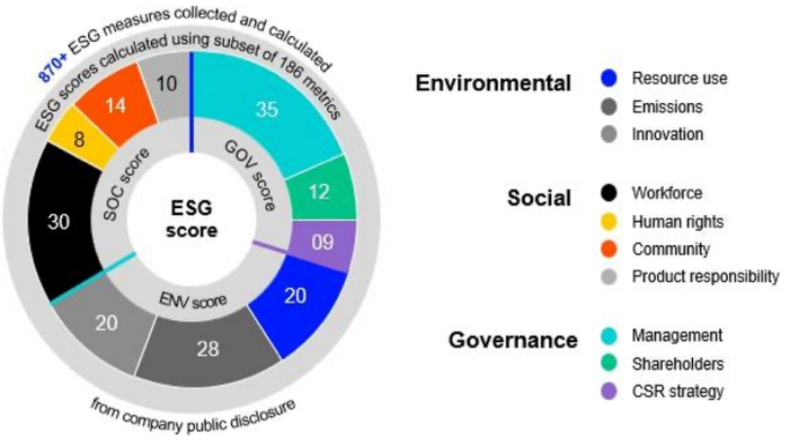


Figure 6 - Refinitiv (LSEG) ESG Score Architecture

Source: LSEG (2024), LSEG ESG Scores Methodology

The three pillar group scores are aggregated using disclosed weights that differ by pillar group into an overall ESG score on a continuous scale from 0 to 100. The methodologically distinctive element of Refinitiv's framework is the ESG Controversies Score, which monitors whether a company has been involved in significant adverse events during the assessment period including environmental disasters, regulatory sanctions, human rights violations, product recalls, or major legal disputes through real-time media and NGO tracking across 23 controversy topics. When a controversy is identified, it can materially reduce the company's final ESGC score (the ESG score adjusted for controversies) below its underlying ESG score, regardless of the quality of its policies. This creates an asymmetric correction mechanism that can sharply penalize firms for specific incidents even when their long-term practices are otherwise well-rated. An additional structural feature is the prominent role of disclosure: because Refinitiv constructs its scores primarily from reported data, companies that disclose more information are, all else equal, likely to receive higher scores. This introduces a systematic bias toward disclosure volume: a company may improve its Refinitiv score by reporting additional data without changing its actual ESG practices, a dynamic that

Drempetic et al. (2020) have identified as a structural conflation of transparency and performance in disclosure-based rating systems.[11][12]

In essence, these three models are designed to answer three fundamentally different questions:

1. How well positioned is this firm relative to peers in managing material ESG risks? (MSCI);
2. How much unmanaged ESG risk does this firm carry in absolute terms? (Sustainalytics);
3. How transparent and controversy-free is this firm's ESG conduct? (Refinitiv).

The outputs are therefore non-comparable by construction, not merely by implementation. Berg et al. (2022) formalize this divergence empirically, decomposing disagreement across six major ESG rating agencies into three sources. They find that 56% of total divergence is attributable to differences in measurement, the way the same underlying attribute is operationalized and quantified; 38% to differences in scope, which attributes are selected for inclusion in the assessment and only 6% to differences in weighting, the relative importance assigned to each attribute once included. The dominance of the measurement component is particularly significant for the purposes of this thesis: it implies that the lack of convergence between rating providers is not primarily a matter of competing analytical priorities, but of fundamentally inconsistent data construction for the same underlying construct. Under these conditions, using ESG

[11] London Stock Exchange Group (2024). *LSEG ESG Scores Methodology.* https://www.lseg.com/content/dam/data-analytics/en_us/documents/methodology/lseg-esg-scores-methodology.pdf

[12] Drempetic, S., Klein, C., & Zwergel, B. (2019). The Influence of Firm Size on the ESG Score: Corporate Sustainability Ratings Under Review. Journal of Business Ethics, 167(2), 333–360. https://doi.org/10.1007/s10551-019-04164-1

ratings interchangeably as standardised inputs in financial models risks introducing systematic and unobservable noise into any analytical framework that relies on them.[13]

1.3 – Limitations of ESG Data

As stated in the previous paragraph, methodological divergence is the main weakness of ESG ratings. The other key criticisms are the following:

- **Lack of standardization**: agencies operate with their individual proprietary methodologies and there is no collective agreement on ESG standards or their relative weights.
- **Black box effect**: rating agencies do not completely disclose their rating formulas, thus creating *"structural opacity"*. As a result, investors are not able to understand the basis for a company's score relative to others, which weakens the capacity for decision-making based on sound information.[14]
- **Fact-checking insufficiency**: the data are largely derived from companies' self-disclosures and there is no systematic external verification. In addition to reporting positive information, companies also omit the negative ones, which are more conducive to greenwashing and misleading narratives.[15]
- **Annual update frequency**: since ratings are generally revised once per year, they miss scandals, controversies and market upheavals that, through the

[13] Berg, F., Kölbel, J. F., & Rigobon, R. (2022). *Aggregate Confusion: The Divergence of ESG Ratings. Review of Finance*, 26(6), 1315–1344. https://doi.org/10.1093/rof/rfac033

[14] SG360. (2024). *ESG rating agency: la sfida di misurare la sostenibilità*. https://www.esg360.it/sustainability-management/esg-rating-agency-la-sfida-di-misurare-la-sostenibilita/

[15] SG360 (2024)

mechanism of high-frequency trading, get immediately reflected in the prices of the financial market.[16]

- **Disclosure ≠ actual results**: scores are more reflective of the quality of reporting and the stated policies ("net-zero 2050") than actual impacts. High-emission companies are rewarded with high ratings simply for the promises they make for the future.[17]

1.4 – ESG as an Informational Variable

The concept of an informational variable originates in the economics of information, where data and signals play a key role in reducing uncertainty and supporting decision-making in financial markets. As highlighted by G. Akerlof, markets are characterized by information asymmetry, meaning that different agents have access to different levels of information[18]. In this context, informational variables act as signals that help investors form expectations and interpret available information under uncertainty.

ESG can be interpreted as a non-financial informational signal that complements traditional financial indicators. Unlike accounting-based measures, ESG information is not fully standardized, but is constructed from a combination of heterogeneous sources, such as corporate disclosures, sustainability reports, media content and third-party evaluations.[19] For this reason, ESG should not be seen as a purely objective measure, but rather as a processed form of information that reflects both the underlying data and

[16] Kotsantonis, S., & Serafeim, G. (2019). *Four things no one will tell you about ESG data. Journal of Applied Corporate Finance*, 31(2), 50–58.

[17] Kotsantonis & Serafeim (2019)

[18] Akerlof, G. A. (1970). *The Market for Lemons*. Quarterly Journal of Economics, 84(3), 488-500.

[19] CFA Institute (2023). *Guidance for Integrating ESG Information*, 12-15.

the methodologies used to organise it. This has important implications. On the one hand, ESG information can enrich the firm's information environment by providing insights into aspects that are not captured by traditional financial data, such as long-term risks, governance structures and sustainability strategies. Empirical evidence suggests that this additional layer of information may contribute to reducing information asymmetry between firms and investors, thereby improving transparency and supporting more informed decision-making.[20] On the other hand, the informational value of ESG depends on how this information is produced, aggregated and interpreted. In this respect, ESG operates within a broader information environment, defined as the availability, quality and dissemination of firm-related information.[21] Overall, ESG should be understood as a dynamic informational construct whose relevance depends on both content and production method.

1.5 – ESG and Financial Markets

Having established ESG as an informational variable, a natural question follows: *what is the relationship between ESG performance and financial markets?* The empirical literature does not offer a single answer. Friede et al. (2015), in the most comprehensive meta-analysis on the topic, covering approximately 2,200 studies, find that roughly 90% of the evidence reports a stable, non-negative relationship between ESG and long-term

[20] Kim, J. W. & Park, C. K. (2023). *ESG Performance and Asymmetry*. Applied Economics, 55(26), 2993-3007.

[21] Bahadır, O., & Akarsu, S. (2024). *Does company information environment affect ESG–financial performance relationship? Evidence from European markets*. Sustainability, 16(7), 2701. https://doi.org/10.3390/su16072701

financial performance.[22] However, the authors highlight two important points: the results may be affected by survivorship bias and correlation does not imply causality.

The picture becomes less clear when more recent and geographically specific evidence is considered. Escobar-Saldívar et al. (2025) find that on the U.S. market, high ESG scores are generally associated with lower long-term stock returns, while increases in ESG ratings tend to produce immediate positive returns and reduced risk, a distinction between ESG level and ESG momentum that points to a more dynamic than aggregate meta-analyses can capture.[23] Gavrilakis and Floros (2023), examining a panel of six European countries over 2010–2020, document a significantly negative relationship between ESG scores and stock performance for Italian firms, a result further confirmed by analysis on the Euronext 100 index.[24]

This divergence is not random. It is, to a significant extent, a direct consequence of the structural limitations discussed in Par. 1.3.: when rating agencies measure different things, apply different methodologies and update at different frequencies, studies built on their output will inevitably produce inconsistent results. *The disagreement in the literature reflects, before anything else, the disparity in the data.*

1.6 – From ESG Metrics to ESG Sentiment

Traditional ESG ratings are, at their core, a static snapshot: they measure what a

---

[22] Friede, G., Busch, T., & Bassen, A. (2015). *ESG and financial performance: Aggregated evidence from more than 2000 empirical studies*. Journal of Sustainable Finance & Investment, 5(4), 210–233.

[23] Escobar-Saldívar, L. J., Villarreal-Samaniego, D., & Santillán-Salgado, R. J. (2025). *The effects of ESG scores and ESG momentum on stock returns and volatility: Evidence from U.S. markets*. Journal of Risk and Financial Management, 18(7), 367. https://doi.org/10.3390/jrfm18070367

[24] Gavrilakis, N., & Floros, C. (2023). *ESG performance, herding behavior and stock market returns: evidence from Europe*. Operational Research, 23(1). https://doi.org/10.1007/s12351-023-00745-1

company reports doing, not how markets perceive it in real time. The informational gap this creates is not trivial. The correlation between ESG sentiment and ESG ratings stands at just 0.158, a result that makes the conceptual separation empirically concrete, the two variables simply do not move together, because they measure different things.[25] Markets, as established in the behavioural finance literature, react to perceptual flows rather than declared fundamentals which is precisely why a signal that updates at the frequency of news may carry more informational content than a rating revised once a year.[26] It is now fundamental to provide a clear definition of what is the core element of this thesis: the ESG sentiment. This indicator is extracted from unstructured text, such as news coverage, earnings calls, social media, sustainability disclosures and the methodological tools used to generate it have evolved substantially across three distinct generations.

The first generation relied on *dictionary-based approaches*. Early sentiment analysis in finance was dominated by lexicon-based tools, of which VADER (Valence Aware Dictionary and sEntiment Reasoner) is the most widely cited example. VADER assigns sentiment scores to individual words using a pre-built valence dictionary, supplemented by a set of grammatical rules designed to account for modifiers such as negation and capitalisation. While computationally efficient and interpretable, its limitations in the ESG domain are well-documented: it ignores contextual meaning, systematically mishandles domain-specific terminology, words such as *"emissions"* or *"risk"* carry

[25] Yu, H., Liang, C., Liu, Z., & Wang, H. (2023). *News-based ESG sentiment and stock price crash risk*. International Review of Financial Analysis, 88, 102646.

[26] Baker, M., & Wurgler, J. (2006). *Investor sentiment and the cross-section of stock returns*. Journal of Finance, 61(4), 1645–1680.

different valences in sustainability discourse than in everyday language and performs at near-chance level on ESG-specific content.[27] The second generation brought *encoder-only* transformer models. The introduction of BERT (Bidirectional Encoder Representations from Transformers) marked a qualitative shift in NLP by processing text bidirectionally, allowing models to consider the full sentential context rather than relying on isolated word scores.[28] Fine-tuned on a large corpus of financial documents, including regulatory filings, earnings call transcripts and corporate announcements, FinBERT represents the leading domain-specific adaptation of this architecture for financial sentiment analysis.[29] In the ESG domain, FinBERT-ESG extended this approach by incorporating E, S, and G label categories, achieving accuracy scores of 0.876 on the Environmental subdomain, 0.758 on the Social subdomain and 0.804 on the Governance subdomain; yet its single-model architecture, forced to differentiate simultaneously between four labels (E, S, G, and none), limits its recall performance and produces systematically noisy inferences whenever a text spans multiple ESG pillars.[30] More broadly, encoder-only models operate within a rigid classification paradigm: they assign discrete labels from a fixed set, which makes them poorly suited to capturing the continuous and contextually layered nature of ESG sentiment in free-form text such as news articles or social media. The third generation is defined by *decoder-only* Large Language Models. In contrast to their predecessors, these

---

[27] Hutto, C., & Gilbert, E. (2014). *VADER: A parsimonious rule-based model for sentiment analysis of social media text*. Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM).

[28] Araci, D. (2019). *FinBERT: Financial sentiment analysis with pre-trained language models*. arXiv preprint, arXiv:1908.10063

[29] Huang, A.H., Wang, H., & Yang, Y. (2023). *FinBERT: A large language model for extracting information from financial text. Contemporary Accounting Research*, 40(2), 806–841.

[30] Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., & Leippold, M. (2024). *Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication*. Finance Research Letters, 61, 104979.

architectures generate responses autoregressively, which allows them to reason about sentiment in a manner closer to human interpretation, rather than simply assigning a label from a fixed taxonomy. The practical implication is substantial: when tested against a human-established gold standard on ESG sentiment classification, GPT-4 achieves 95% accuracy, compared to 48% for both VADER and FinBERT.[31] This performance gap reflects a structural advantage, not merely a quantitative one. Decoder-only models can process ambiguity, domain-specific nuance, and multi-pillar ESG content in ways that lexicon-based and encoder-only architectures cannot. The specific models employed in this research belong to this generation, and their characteristics are examined in detail in Chapter 2. What matters here is the evidential basis for their use: better measurement produces a more reliable signal, one more likely to detect the effects that rating-based approaches miss. Those effects appear to be real, higher ESG sentiment has been shown to reduce the risk of stock price crashes, operating through two concrete channels, namely reduced information asymmetry and lower agency costs.[32] Despite this evidence, ESG sentiment has not yet been systematically adopted as a primary variable in empirical ESG finance. The obstacle is not conceptual but infrastructural: available datasets remain noisy, geographically unequal and methodologically incomparable across studies.[33]

This thesis aims to address the gap in the literature and contribute to it, by building a

---

[31] Derrick, K. (2024). *ESG sentiment analysis: Comparing human and language model performance including GPT*. arXiv preprint, arXiv:2402.16650.

[32] Yu et al., op.cit.

[33] Sultana, N. (2026). *Global ESG sentiment, policy commitments, and sustainability uncertainty: A cross-country analysis*. Sustainable Futures, 11, 101636.

dataset of synthetic ESG sentiment generated by the various LLMs that will be presented in Chapter 3, on a daily basis, across a panel of European countries. In fact, the research seeks to understand whether it is possible to generate a reliable and stable indicator and if there is a relationship between ESG sentiment and financial markets.

CHAPTER 2 – THE ROLE OF LARGE LANGUAGE MODELS

2.1 – Big Data and Alternative Data in Finance

We are living in an era where computers and the internet are everywhere and the amount of data generated has grown accordingly. As stated by Statista [34] the total volume of information collected, copied and used around the world is expected to be 173.4 zettabytes in 2025 and will probably increase to 527.5 zettabytes by 2029. That is nearly three times more in just four years. The growth of data accelerated sharply in 2020, when the COVID-19 pandemic transformed work, education and entertainment habits within a very short period. IBM also reports that 90% of all information made by companies is unstructured including text, image, audio and sensor data that is not in a specific format, pictures, sound and signals from sensors that are hard to analyze using numbers.[35] This suggests that the challenge is not only the sheer quantity of data, but also the difficulty of extracting meaning from increasingly complex data formats.

The main driver of this exponential growth is what is commonly referred to as *Big Data.* At present, numerous definitions coexist in the literature that vary significantly based on the industry in which they are applied. From an engineering and computer science perspective, Big Data is characterised by five dimensions, Volume, Velocity, Variety, Veracity and Value, that distinguish it from traditional datasets.[36] In a financial context, Goldstein et al. (2021) offer a more operational characterisation, based on three structural properties. The first is *large size*, which refers both to the absolute scale of a

[34] Statista (2025). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2029. https://www.statista.com/statistics/871513/worldwide-data-created/

[35] IBM (2023). What is big data? https://www.ibm.com/topics/big-data

[36] IBM (2023)

dataset such as the millisecond-stamped records generated by high-frequency trading markets and to its magnitude relative to datasets historically employed in financial research; access to more complete data helps mitigate sample selection bias and enables a more accurate representation of complex economic phenomena. The second is *high dimensionality*, where the number of variables is large relative to the number of observations. In these settings, machine learning methodologies become particularly relevant, especially when relationships between variables are non-linear, when the problem involves many explanatory variables or when the primary objective is predictive accuracy rather than statistical inference, a condition that is especially central in automated decision-making contexts such as algorithmic trading. Finally, the third and most relevant for this research is *complex structure* that frequently includes unstructured or semi-structured formats that do not conform to traditional tabular arrangements and require advanced feature extraction techniques, often relying on natural language processing (NLP) and computer vision.[37]

It is exactly this latter category that opens the door to a broader and more consequential distinction: that between traditional and alternative data. Traditional or structured financial data (earnings reports, balance sheets, regulatory filings) are authoritative and widely used, but they carry well-known limitations: they are often incomplete due to proprietary constraints, subject to manipulation and published at low frequency, making them insufficiently timely for real-time decision-making. Alternative data, by contrast, address these gaps specifically because they come from non-traditional sources (free

[37] Goldstein, I., Spatt, C. S., & Ye, M. (2021). Big data in finance. The Review of Financial Studies, 34(7), 3213–3225.

text, satellite imagery, social media posts, product reviews, search trends) and can be captured and processed in real time, giving investors and researchers a more dynamic and comprehensive view of market conditions. Sun et al. (2024) identify five advantages that characterise alternative data relative to traditional sources.

1. **Multiple sources**: alternative data ranges from textual content in online news articles and reviews to visual data captured through satellite imagery and unmanned aerial vehicles, enabling more robust conclusions from different angles;
2. **Heterogeneity**: while traditional data are usually structured, organised and formatted, alternative data include semi-structured and unstructured formats including pictures, sentences, videos and audio. This heterogeneity makes manipulation more difficult and allows alternative data to reflect underlying economic conditions more faithfully;
3. **Flexibility**: alternative data can be updated and retrieved within seconds, freeing investors and researchers from dependence on periodic annual or quarterly reports and enabling real-time access to consumer sentiment, transactional data and geospatial information.
4. **Objectiveness**: because alternative data are publicly generated, unfiltered and continuously updated from diverse sources, they are less tied to any individual firm's disclosure strategy or incentives.
5. **Constant evolution**: the universe of alternative data expands as technology advances, with new data sources emerging over time and progressively enriching

existing analytical approaches. [38]

For the purposes of this thesis, however, it is necessary to draw a methodological distinction that is central to the study. Alternative data are *passively* extracted from external sources. In this work, ESG sentiment scores are *actively* generated through Large Language Models: they do not pre-exist in any source but are synthetically produced through a generative approach, making them *"alternative"* not in terms of origin, but in terms of their mode of production, since they bypasses the standard data collection pipelines of vendors such as MSCI or Refinitiv. It is precisely the heterogeneity of alternative data and the dominance of unstructured text within it that makes this approach both necessary and consequential: extracting systematic, comparable signals from corporate disclosures, sustainability reports and policy documents requires models capable of understanding natural language at scale. Understanding how Large Language Models process and generate information is therefore a prerequisite for evaluating the reliability and the economic meaning of the ESG scores they produce, as discussed in the next section.

2.2 – What is a Large Language Model?

Before defining what, a Large Language Model is, it is important to outline the market that is expanding exponentially and has become highly competitive. Until 2020, the number of publicly released large-scale models was still limited: GPT-3 with its 175

[38] Sun, Y., Liu, L., Xu, Y., Zeng, X., Shi, Y., Hu, H., ... & Abraham, A. (2024). *Alternative data in finance and business: emerging applications and theory analysis (review). Financial Innovation*, 10(1), 127.

billion parameters, was considered as an exception.[39] A turning point came in November 2022 with the launch of ChatGPT, based on OpenAI's GPT series. A remarkable indication of this shift is that the daily average of papers published on arXiv containing the term *"Large Language Model"* goes from 0.40 to 8.58 publications per day, a growth of over 2000% within a few months. As shown in Figure 7, from 2023 onwards, the production of new models increased cumulatively, with new versions released every year by an increasing number of providers.[40]

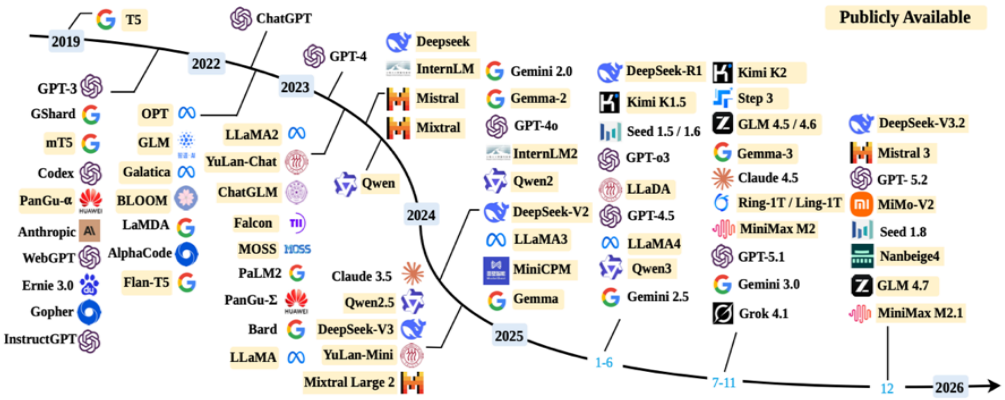


Figure 7 - Timeline of the main Large Language Models released from 2019 to the end of 2025. Models with publicly available checkpoints are highlighted in yellow

Source: Zhao et al. (2023, v13), "A Survey of Large Language Models", Fig. 3.

The figure clearly shows three relevant dynamics of this recent phenomenon. Firstly, growth is structurally exponential: from a few models in 2019–2020 to dozens of simultaneous releases as early as 2024–2025, a sign that the barrier to entry has lowered and innovation has spread across multiple global players. Secondly, a clear distinction emerges between open-source models, whose weights are publicly downloadable (highlighted in yellow, such as Meta's LLaMA) and closed-source models, accessible exclusively via APIs, which include the five models used in this thesis: GPT-5.2

[39] Brown et al. (2020), *Language Models are Few-Shot Learners*, arXiv:2005.14165, p. 3

[40] Zhao et al. (2023), *A Survey of Large Language Models*, arXiv:2303.18223v13, p. 2

(OpenAI), Claude Sonnet 4.5 (Anthropic), Grok 4.1 (xAI), Sonar (Perplexity AI) and Gemini 3 Pro (Google DeepMind).[41] Finally, the timeline highlights how all five belong to the final branch of 2025, representing the most recent and technically advanced generation currently available: models trained on millions of tokens, with extended context windows, multimodal capabilities and sophisticated procedures for aligning with human values.[42] In the next paragraph, the models mentioned above will be presented, to better understand their alignment strategies and architectures.

After presenting the market, it is now possible to define what is meant by a Large Language Model. At a high level, an LLM can be defined as a neural network of exceptional scale, trained on extremely large text corpora with the goal of predicting the next token in a sequence given everything that came before. Zhao et al. define them as *"transformer language models that contain hundreds of billions or more of parameters, trained on massive text data, that exhibit strong capacities to understand natural language and solve complex tasks via text generation".* This does not represent a completely new architecture compared with previous models; rather, it results from the systematic and massive application of the same principles, therefore more parameters, data and computing power. It is this quantitative continuity that generates, at a certain threshold, a qualitative discontinuity.[43] To understand why, it is useful to recall the predecessors of the LLMs. The dominant models until 2017 were Recurrent Neural Networks (RNNs) and their most sophisticated variants, Long Short-Term Memory

[41] Zhao et al. (2023)

[42] Brown et al., op. cit.

[43] Zhao et al., op. cit.

Networks (LSTMs). These systems processed the text sequentially, one token at a time, accumulating in a fixed-size vector the memory of everything that had been read up to that point.[44] The problem here is structural: sequential computation cannot be parallelized, which makes training slow and expensive and the hidden vector tends to *"forget"* information from distant tokens, which makes it difficult to capture long-range relationships, this is the so-called *vanishing gradient problem*.[45] These two limitations placed a practical ceiling on how large and capable RNN-based models could become.

The break with this paradigm came in 2017, when Google's Vaswani et al. published *"Attention Is All You Need",* introducing the *Transformer* architecture. The central mechanism is *"self-attention"*: instead of processing the text token by token, the model simultaneously calculates the relationship of each element of the sequence with all the others, assigning each pair a weight that reflects their mutual relevance. This mechanism is fully parallelizable as there is no sequential dependency, which has made it possible to train models on distributed hardware at a speed previously unthinkable. In the original architecture proposed by Vaswani et al., the Transformer is composed of two distinct blocks, an *"encoder"* and a *"decoder"* (see Figure 8). The encoder receives the input sequence and transforms it into a dense internal representation, a series of vectors that capture the contextual meaning of each token in relation to all the others. The decoder, in turn, receives this representation and generates the sequence of output token by token, but with a fundamental constraint; at each step, it can only wait for tokens already

[44] Zhao et al. (2023)

[45] Vaswani et al. (2017), *Attention Is All You Need*, arXiv:1706.03762, p. 2.

generated previously, not future ones, a mechanism called *"masked self-attention"* that ensures that the generation remains causal.[46] The most recent models, including all five of those used in this thesis, adopt a *"decoder-only"* architecture: they eliminate the encoder and train a single self-regressive block that learns to compress into its parameters a representation of the world rich enough to respond to any type of natural language query.[47]

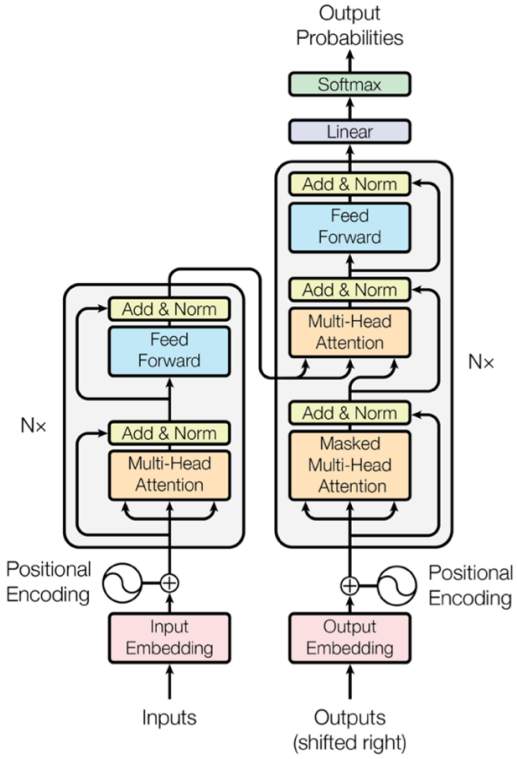


Figure 8 - Graphical representation of the Transformer model architecture, consisting of an encoder (left) and a decoder (right), each organized in N overlapping layers

Source: Vaswani et al. (2017). Attention Is All You Need. arXiv:1706.03762v7, p. 3, Figure 1.

In this context, scaling laws acquire their most precise meaning. Kaplan et al. (2020) showed that the cross-entropy loss (L) of a language model scales according to a power

[46] Vaswani et al. (2017)

[47] Zhao et al., op. cit

law with respect to the number of parameters (N), the size of the dataset (D) and the total compute (C):[48]

$$L(N) \approx \left(\frac{N_c}{N}\right)^{\alpha_N}$$

where N_c is the critical parameter scale and $\alpha_N > 0$ is an empirically estimated exponent. In practical terms, this means that doubling the size of the model, while also scaling the data proportionately, produces a predictable and consistent improvement in the quality of language comprehension and generation, across different tasks, languages and domains, without any specific redesign.[49] It is exactly this law that explains the extraordinary leap from GPT-2 (1.5 billion parameters, 2019) to GPT-3 (175 billion, 2020): not a different architecture nor a new algorithmic idea but the scale applied in a systematic way.[50]

2.2.1 – Differences between LLMs models

The theoretical framework introduced above acquires concrete meaning when applied to the five models at the core of this research: GPT-5.2, Claude Sonnet 4.5, Grok 4.1, Sonar and Gemini 3 Pro. All five share the same foundational architecture, the decoder-only Transformer, yet diverge along dimensions that directly affect the quality and comparability of their ESG outputs. One of the most consequential is *alignment strategy*, which warrants a brief explanation before the models are introduced individually. A raw LLM, newly trained on the pre-training corpus, is optimized to predict the next token,

[48] Zhao et al. (2023)

[49] Zhao et al. (2023)

[50] Brown et al., op.cit.

not to be useful, honest, or secure. It can generate content such as fake answers safely also called *"hallucinations"* or simply output that is not relevant to the question. Strategic alignment is the process by which the model is *"instructed"* to behave in a desirable way after pre-training.[51] The most common technique is Reinforcement Learning from Human Feedback (RLHF) where human evaluators compare pairs of model responses and indicate the preferable one. These preferences are used to train a reward model that then guides the main model towards outputs that are increasingly valued by humans.[52]

GPT-5.2 is the most directly relevant model for this research as it represents the current state of the art in the GPT series, the family that has set the benchmark for the industry since 2020. OpenAI employs RLHF combined with supervised instruction tuning, an approach formalised with InstructGPT in 2022 and progressively refined in subsequent releases.[53] The decision to include GPT-5.2 is motivated by its broad thematic coverage during pre-training, its documented ability to follow structured instructions and its widespread adoption in the emerging literature on LLMs applied to financial tasks which allows for direct comparison with prior empirical findings

Claude Sonnet 4.5 is distinguished by a structurally different alignment philosophy. Anthropic developed Constitutional AI (CAI): rather than relying exclusively on human annotators, the model is trained to critique and revise its own responses according to an

[51] Zhao et al., op.cit.

[52] Zhao et al. (2023)

[53] Ouyang, L., et al. (2022). *Training language models to follow instructions with human feedback (InstructGPT).* arXiv:2203.02155, p. 1.

explicit set of ethical principles, a so-called "constitutional document" that includes directives such as *"be honest"*, *"avoid causing harm"* and *"respect user autonomy"*.[54] This reduces dependence on direct human feedback in the later stages of fine-tuning, with the goal of producing a more consistent model, less susceptible to variation in annotator judgment. For the purposes of this thesis, Claude Sonnet 4.5 represents a methodologically valuable control case: any systematic divergence in its ESG scores from those of the other models could be partially attributed to its distinct alignment architecture, a hypothesis testable in the empirical section.

Grok 4.1 was developed by xAI, the company founded by Elon Musk in 2023 with the stated objective of building an AI system that is "*maximally curious and broadly uncensored"*. This positioning carries a specific implication for ESG analysis: on topics involving political governance, human rights, and institutional transparency, a model with less stringent alignment constraints may produce more polarised or explicit assessments of countries with authoritarian governments or controversial records. Grok 4.1 operates with instruction tuning but xAI has not published detailed technical papers on its alignment procedure, a transparency limitation that will be discussed in the methodology section.

Gemini 3 Pro is the flagship model of Google DeepMind, developed with a particular emphasis on multimodal capabilities and an exceptionally extended context window, documented up to one million tokens in version 1.5, with further expansions in the 2.x

[54] Bai, Y., et al. (2022). *Constitutional AI: Harmlessness from AI Feedback*. arXiv:2212.08073, p. 1.

series.[55] Its relevance to this thesis is twofold. First, Google DeepMind has historically invested in pre-training datasets with strong coverage of institutional sources, scientific papers, and government reports. Second, Gemini 3 Pro is the only model in the sample developed by a company that operates directly in the institutional information services market, which could be reflected in a greater sensitivity to the sources and standards used by international organizations in their ESG reports.[56]

Sonar (Perplexity AI) is the most structurally diverse model of the sample. Contrasting the other four, Sonar integrates a Retrieval-Augmented Generation (RAG) mechanism: before generating a response, the system queries a search engine in real time and incorporates the retrieved documents in the context of the generation.[57] In practice, Sonar not only relies on the knowledge compressed in the pre-training but also draws on updated information at the time of the query. For the production of ESG Sentiment scores, this has a double implication: on the one hand, Sonar could be more up-to-date on recent events such as environmental crises, governance scandals, new regulations that models with a more distant training cut-off may not have incorporated. On the other hand, the output of Sonar depends on the quality and selection of sources retrieved in real time, introducing a systematic difference that other models do not have. This structural difference will be treated as an explicit limitation in the discussion of the results. To conclude, Table 1 provides a summary of the main characteristics of the models included in the sample.

---

[55] Reid et al. (2024), *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*, arXiv:2403.05530, p. 1.

[56] Team Gemini et al. (2023), *Gemini: A Family of Highly Capable Multimodal Models*, arXiv:2312.11805, p. 1–3.

[57] Lewis, P., et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. arXiv:2005.11401, p. 1.

MODEL	PROVIDER	RELEASE DATE	CONTEXT WINDOW	ALIGNMENT STRATEGY
GPT-5.2	OpenAI	December 11,2025	400,000 tokens	RLHF + Instruction Tuning
Claude Sonnet 4.5	Anthropic	September 25 ,2025	200,000 tokens	Constitutional AI (CAI)
Grok 4.1	xAI	November 19,2025	256,000 tokens	Instruction Tuning
Gemini 3 Pro	Google DeepMind	November 18,2025	≥1,000,000 tokens	RLHF + Multimodal Fine-tuning
Sonar	Perplexity AI	January 21,2025	128,000 tokens[58]	RAG + Instruction Tuning

Table 1 - Overview of the LLMs included in the study sample

Source: author's own elaboration based on official provider documentation [59]

2.3 – Prompt Design

A Large Language Model (LLM) as discussed before, is essentially an autoregressive prediction system: given a textual context, it estimates the probability distribution on the next token and by iterating this process, generates text sequences consistent with the input received.[60] The mechanism through which the user interacts with this system is the *prompt*, that is, any textual sequence provided to the model before or during the generation of the response. The prompt is the only operating lever available to the user to orient the behaviour of a model whose parametric weights remain fixed, i.e., in the

[58] Sonar is included as a retrieval-augmented system, so its practical effective context may vary depending on retrieval depth and the documents injected into the prompt. The value reported here refers to the model's stated context window in the cited documentation.

[59] Release dates and context windows are based on official provider documentation and release posts. GPT-5.2: OpenAI, Introducing GPT-5.2; context window 400,000 tokens. Claude Sonnet 4.5: Anthropic, Introducing Claude Sonnet 4.5; context window 200,000 tokens. Grok 4.1: xAI, official Grok 4.1 release post; context window as reported in xAI documentation/release materials. Gemini 3 Pro: Google Cloud, Gemini 3 Pro; context window 1,000,000 tokens. Sonar: MindStudio, Sonar — AI Model; context window 128,000 tokens.

[60] Zhao et al., op.cit.,

absence of fine-tuning, after pre-training. This finding is the logical premise of the entire discipline of prompt engineering: since the model cannot be modified, it is the prompt that must be optimized. Zhao et al. define *prompt engineering* as the iterative process of designing textual inputs that induce the model to produce accurate, relevant, and replicable outputs, acting on variables such as the syntactic structure of the text, the tone, the information context provided, the presence or absence of demonstrative examples and the role assigned to the model.[61] Empirical research shows that large LLMs are sensitive to the exact formulation of the input as small lexical variations in the order of examples or in the structure of the role assignment can produce systematically different outputs, making the standardization of the prompt a necessary condition for any comparative analysis.[62]

To understand why the choice of prompt type is a methodologically relevant decision, it is first necessary to understand the phenomenon of LLMs *emergent abilities*. As explained above, scaling laws describe a power relationship between model size, training corpus size and language loss: as computational resources increase, performance improves continuously and predictably.[63] However, Zhao et al. distinguish this continuous trend from a second qualitatively different phenomenon: emergent abilities, formally defined as capabilities that are not present in small-scale models but emerge in large-scale models, typically beyond a parametric threshold in the order of

[61] Zhao et al. (2023)

[62] Lu, Y., et al. (2022). *Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity*. arXiv:2104.08786

[63] Zhao et al. (2023)

tens of billions of parameters.[64] The distinguishing feature of these skills is their discontinuity: performance remains essentially flat until a critical threshold is reached, after which there is a sudden and marked jump in performance.[65]

Three emergent abilities are of particular relevance for the present research. The first is *in-context learning* (ICL), formally introduced by Brown et al. (2020) with GPT-3, is the ability of a model to adapt to a new task simply by conditioning itself on the text of the prompt that includes instructions and, possibly, even demonstration examples without any update of the parametric weights, operating exclusively through the *forward pass*.[66] Brown et al. show that GPT-3, with its 175 billion parameters, exhibits a strong ICL capability, whereas smaller models of the same family such as GPT-1 and GPT-2 do not possess it to a comparable extent. Crucially, in-context learning curves are markedly steeper for larger models: the same amount of contextual information provided in the prompt produces increasing performance gains as the scale of the model increases.[67] The second is *instruction following*: models trained with instruction tuning on mixed datasets show the ability to perform tasks described in natural language without explicit examples, but this ability emerges significantly only beyond 68 billion parameters. The third and last is *step-by-step reasoning*: the ability to break down complex problems into intermediate steps, which manifests itself robustly only in

64 Zhao et al. (2023)

65 Wei, J., et al. (2022). *Emergent Abilities of Large Language Models*. arXiv:2206.07682

66 Brown et al., op.cit.

67 Brown et al. (2020)

models that exceed the threshold of 60-100 billion parameters, and which can be elicited but not created from scratch through special prompting techniques.[68]

The implications of this framework for comparability between models are straightforward: models of different scales react qualitatively differently to the same prompt.[69] The five models used in this research are all above the thresholds where the emergent abilities described are documented, implying that each of them is in principle capable of exploiting the content of the prompt in a sophisticated way. However, sensitivity to variations in the structure of the prompt i.e., the presence or absence of examples, or the formulation of the chain-of-thought differs between architectures and model families, introducing a source of systematic variance that can confuse model comparison if the prompt is not rigorously standardized.[70]

The literature distinguishes a consolidated taxonomy of prompting techniques, ordered by increasing complexity and by the amount of contextual information provided to the model.[71]

1. **Zero-shot prompting**: the most basic technique in which the model is only provided with a description of the task in natural language such as a question, an instruction, an indication of the expected format without the indication of any type of input-output demonstration example. The model relies entirely on the

[68] Wei, J., et al. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv:2201.11903

[69] Zhao et al., op.cit.

[70] Lu et al., op. cit.

[71] Zhao et al.,op.cit.

parametric knowledge gained during pre-training to interpret the task and generate the response. Brown et al. show that, under zero-shot conditions, GPT-3 already achieves competitive results on numerous benchmarks, but with lower performance on average than conditions that include contextual examples, highlighting the incremental contribution of demonstrative information. In the zero-shot prompting, the model has maximum interpretive autonomy: there are no examples that constrain the form or content of the response, which makes the results more dependent on the alignment between the prompt formulation and the model training distribution.[72]

2. **Few-shot prompting**: it is an extension of the zero-shot by prefixing the real prompt with a small number of input-output demonstration pairs, typically from 2 to 100, depending on the length of the model's context window. Brown et al. (2020) demonstrate that this technique substantially improves performance on specific tasks, since the examples provide the model with explicit information on the expected format, the desired semantic mapping, and the level of detail of the response.[73] However, the technique introduces a critical dependency on three factors: the choice of examples, their order of presentation, and their distribution across the task domain. Variations in these factors can systematically steer the model toward certain response patterns, making it difficult to isolate the pattern effect from the prompt effect. This phenomenon, known as prompt sensitivity, is extensively documented in the literature and is one of the main reasons why

[72] Brown et al., op.cit.

[73] Brown et al. (2020)

research aiming at inter-model comparability avoids the few-shot in the absence of a rigorous example selection protocol.[74]

3. **Chain-of-Thought (CoT) prompting**: this technique explicitly asks the model to externalize the intermediate reasoning steps, the so-called *"chain of thought"* before formulating the final answer. In practice, the prompt includes examples where the solution is preceded by a step-by-step explanation of the procedure followed. Wei et al. (2022) show that this technique significantly improves performance on arithmetic, symbolic and common-sense reasoning tasks, but only in models that exceed the critical threshold of about 100 billion parameters while in smaller models, CoT can even worsen performance compared to standard prompting. CoT introduces variability in inferential paths: the model can follow different reasoning to reach the same conclusion, or seemingly correct reasoning that leads to incorrect conclusions, making the output less deterministic than the zero-shot.[75]
4. **Chain-of-Thought zero-shot**: A particularly interesting variant is the CoT zero-shot, documented by Kojima et al. (2022). Adding just the *"Let's think step by step"* statement at the end of the zero-shot prompt is enough to elicit the chain-of-thought reasoning without providing demonstrative examples.[76] This result shows that the ability to reason by steps is already latent in the parameters of large-scale models and can be activated with minimal textual indication. However, even in this variant, the variability of the generated reasoning path

[74] Brown et al. (2020)

[75] Wei, J., et al., op.cit.

[76] Kojima, T., et al. (2022). *Large Language Models are Zero-Shot Reasoners*. arXiv:2205.11916

remains a source of non-determinism in the output.[77]

5. **Instruction prompting and system prompting**: Modern LLMs APIs support a separate input layer, the system prompt or instruction prompt, which precedes the user's query and defines the role, tone, constraints, and operational context of the model. The role prompting *"You are an ESG analyst..."* belongs to this category. Empirical research shows that explicit definition of a role improves thematic coherence and domain specificity in responses, particularly for tasks that require specialized knowledge.[78]

The review of the main prompting techniques presented in this Chapter does not have an exclusively descriptive function but constitutes the theoretical foundation on which the entire empirical framework of the research rests. The choice of interaction strategy with the language models is not a secondary operational detail, it is a substantive methodological decision that directly influences the validity and comparability of the results. As demonstrated by Brown et al., the difference between zero-shot and few-shot is not a simple gradation of complexity but radically different configurations of the inferential process of the model, each with distinct implications in terms of generalization and dependence on the context provided.[79]

In this research, the choice to operate in the zero-shot regime responds to a precise methodological rationale: eliminating any conditioning induced by in-context examples

[77] Wei, J., et al., op.cit.
[78] Zhao et al., op.cit.
[79] Brown et al. (2020)

allows the researcher to evaluate the intrinsic ability of the models to interpret instructions expressed in natural language, making the results comparable between different architectures without introducing distortions related to the selection or quality of the examples.[80] At the same time, setting the temperature to zero ensures that the model's output is deterministic and reproducible. In the absence of stochastic sampling, the model systematically selects the maximum probability token, eliminating the variance introduced by the sampling parameters and making each experiment replicable under identical conditions.[81]

It is from these choices that Chapter 3 develops the complete methodological architecture of the research. After defining the research questions that guide the analysis, the chapter describes the pre-training phase of the models involved, the ESG sentiment generation process and the data sources used and finally illustrates the construction of the final dataset and provides an overview of the experimental process. Every operational decision will be anchored in the theoretical principles set out here, making explicit the path that leads from theory to practice.

[80] Bianchi, M., et al. (2024). *Optimizing Large Language Models for ESG Activity Detection in Financial Texts*. Section 4.3.1.

[81] Boonstra, L. (2024). *Prompt Engineering*. Google, p. 9.

CHAPTER 3 – RESEARCH DESIGN & METHODOLOGY

3.1 – Research Questions

After building the foundations in Chapters 1 and 2, Chapter 3 introduces the core methodological stage of the research. This part presents the research questions and the full methodological framework used to build the empirical basis for the analysis developed in the following chapter. The study is organized around four main research questions, each of which is framed analytically and then linked to a specific conceptual objective.

> ***RQ-1**: To what extent is LLM-generated ESG sentiment associated with financial market performance across the selected European countries, after controlling for broader market movements through the STOXXEurope600?*

This first research question investigates the empirical correlation between synthetic ESG sentiment and country-level financial returns with the *STOXXEurope600* included as a control variable to isolate broader market effects. The analysis is designed to verify whether the generated ESG signal contains informational content relevant for financial performance beyond general regional market movements.

> ***RQ-2:** How does LLM-generated ESG sentiment evolve over time and across the selected European countries in 2024?*

This second question focuses on the temporal and cross-country dynamics of the daily ESG series in 2024, allowing the identification of country-specific patterns, volatility

and recurring shifts in sentiment. The descriptive analysis is carried out through Excel pivot tables and graphical inspection, which make it possible to compare countries and LLMs outputs at a granular level.

***RQ-3**: Does LLM-generated ESG sentiment exhibit sufficient internal coherence to be interpreted as a stable informational construct, or does it mainly reflect model-specific dispersion and noise?*

This third research question examines whether the generated ESG signal behaves as a common latent factor or whether it is mainly fragmented across models and observations. Principal component analysis (PCA) is used as a diagnostic tool, with the first principal component (PC1) serving as the main criterion for assessing whether the synthetic series displays sufficient internal coherence to be treated as a stable informational construct.

***RQ-4**: How do prompt design, model architecture, and source calibration affect the construction of synthetic ESG sentiment data?*

This fourth research question addresses the epistemic role of LLMs in the construction of the dataset, with attention to prompt design, internal knowledge bases, event-based calibration and the limitations of traceability. It is intended to clarify not only how the data were generated but also what kind of methodological object an LLM-generated ESG series actually is.

Following the presentation of the research questions, Chapter 3 presents the methodological framework adopted to operationalize the study. The chapter first explains how the selected LLMs are interrogated through a standardized prompt structure, before moving to the algorithmic construction of a daily ESG sentiment panel for eight European countries in 2024. It then describes how the synthetic outputs are assembled into the final dataset, integrated with the financial variables used in the empirical tests and prepared for the final statistical analysis. Figure 9 gives a clear overview and serves as a roadmap for the methodological part explained in the next sections. In this way, the chapter moves progressively from model interaction to data construction and finally to empirical validation.

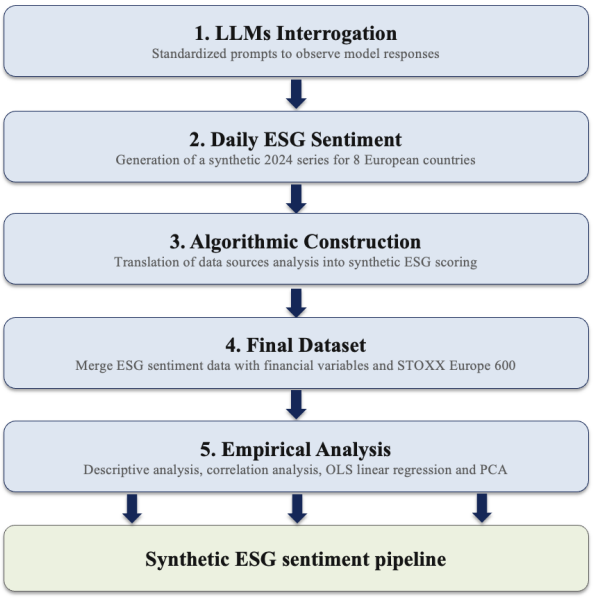


Figure 9 - Methodological workflow of the study (author's own construction)

3.2 – Preliminary Testing Phase

Before constructing the final ESG sentiment dataset, a preliminary testing phase was conducted to see how the selected LLMs would understand the same task when prompted in the same way. In this thesis, *preliminary testing phase*, does not refer to the

original large-scale pre-training that the model vendors do but a preliminary testing phase before the final data-generation process. It was aimed to assess whether the five selected models could transform a common instruction related to ESG into outputs that would be sufficiently structured, interpretable and in a format that would be comparable for follow-on empirical work.

This phase is directly grounded in the theoretical framework developed in Chapter 2. As discussed there, all five systems belong to the decoder-only Transformer family, but they differ in alignment strategy, context management and in the specific case of Sonar, retrieval configuration. These differences are expected to influence how each model interprets an instruction, prioritizes information, calibrates sentiment and balances synthesis against explanation. For this reason, prompt design is treated as a methodological variable rather than a neutral interface and the pilot comparison in Chapter 3 serves to observe how the structural characteristics outlined in Chapter 2 become operational when the same ESG prompt is applied to the same country case.

Italy was selected as the pilot country for this calibration stage because its 2024 context was shaped by climate-related shocks, territorial asymmetries, governance concerns and an intense public debate on adaptation and energy transition, all of which made it especially suitable for testing ESG sentiment generation. Using Italy as an initial benchmark therefore made it possible to test not only whether the models could identify ESG themes but also whether they could preserve coherence when asked to synthesize a complex national case into a structured sentiment output.

Analyze public sentiment in Italy regarding climate change and ESG (Environmental, Social, Governance) topics during Jan 2024 - Dec 2024, using news articles, social media posts, corporate reports, etc.
Identify key themes under each ESG dimension, assign a sentiment score from –1 (negative) to +1 (positive) for Environmental, Social, and Governance aspects and briefly explain the reasoning behind each score.
Describe how sentiment evolved during the period, highlighting major events or discussions that shaped public perception.
Provide a concise summary of overall national sentiment (positive, neutral, or negative), the main topics or concerns driving it, and any observed trends or shifts in tone.
Conclude with a short strategic insight on how this sentiment might influence policy decisions, corporate ESG strategies, or public communication on climate issues.

Figure 10 - Preliminary Testing Phase First Prompt (author's own construction)

Figure 10 reports the first standardized zero shot prompt used as a benchmark. This prompt was deliberately broad and multi-layered, because its objective was not simply to generate a summary but to test whether the models could maintain analytical coherence across classification, scoring, temporal interpretation and strategic synthesis within a single answer. The pilot responses showed that the same prompt generated materially different ESG narratives:

- **Claude Sonnet 4.5** produced the most cautious and well-structured answer of the five. It leans toward a mixed or at times moderately negative, reading of Italian ESG sentiment in 2024, with particular attention to political tensions, governance concerns and economic friction. In this sense, the output does not frame the year as broadly optimistic but rather as one marked by caution and balance. This is consistent with the Chapter 2 profile of Claude Sonnet 4.5 as a more moderated and normatively careful model.
- **Grok 4.1** gave the most explicitly negative response. Its answer places strong emphasis on institutional weakness, insufficient government action, and frustration with policy inaction, making the overall tone sharper and more critical than the others. What stands out here is not only the negativity itself, but the way

it connects climate sentiment to governance failure and execution problems. This fits the Chapter 2 discussion of Grok 4.1 as a model that tends to produce more direct and less filtered judgments.

- **Gemini 3 Pro** adopted a broader and more synthetic style. It interpreted Italian public sentiment as predominantly positive, although still clearly marked by concern and gave substantial space to support for climate action, adaptation, and sustainability transition. At the same time, it did not ignore skepticism toward institutions and firms, which kept the response analytically balanced rather than celebratory. This matches the Chapter 2 description of Gemini 3 Pro as a model with strong contextual aggregation and wide informational coverage.
- **Sonar** also leaned toward a generally positive but frustrated interpretation. The answer stressed climate awareness, adaptation and policy urgency, but it did so with stronger attention to concrete events and evidence, which gives the output a more grounded and event-sensitive tone. This makes the response feel more source-anchored than the others, which is coherent with Sonar's retrieval-augmented structure. For this reason, Sonar appears to be methodologically distinct from the purely parametric models discussed in Chapter 2.
- **GPT-5.2** was the most intermediate and balanced of the five. It described Italian sentiment as slightly positive to neutral, avoiding both an overly optimistic and an overly pessimistic framing. The answer combines support for adaptation and renewable transition with skepticism toward uneven implementation and corporate ESG claims, which gives it a measured and stable tone. This aligns

with the Chapter 2 characterization of GPT-5.2 as a model with strong instruction-following and reliable task completion

This comparison is methodologically useful because it confirms two things at once. First, the benchmark prompt is sufficiently broad to test whether the models can handle the complete ESG workflow, from thematic extraction to sentiment scoring and strategic synthesis. Second, the output differences indicate that prompt design alone does not guarantee comparability, since each model filters the same instruction through its own alignment style, context handling, and reasoning preferences. Therefore, the first prompt functions as the interpretive anchor of the entire preliminary testing phase, it establishes the baseline against which all subsequent, more specific prompts can be read.

The remaining prompts were then grouped into analytical families in order to test whether the same differences persisted when the task became more specific. The first group concerns territorial differentiation and includes the prompts:

- *How did public sentiment differ between northern and southern Italy in 2024?*
- *Which regions in northern Italy showed the largest sentiment shift in 2024?*

These prompts test whether the models can move from a national narrative to a sub-national one without losing coherence and whether they can preserve internal segmentation when the ESG issue is framed geographically.

The second group is formed by event-correlation prompts, namely:

- *Correlate regional sentiment shifts with extreme weather events in 2024*
- *Quantify sentiment change before and after major 2024 floods per region*

- *Calculate the correlation coefficient between tropical nights and social sentiment shifts, including health impacts*
- *What is the monthly average of extreme events and the delta in awareness pre/post-seasonal peaks, with focus on north?*

These are the most demanding prompts in the series because they require the models to link climate shocks, temporal variation, and sentiment dynamics within a single answer, thus testing the stability of their event-based reasoning.

A third group includes the socio-economic prompt:

- *What is the percentage variation in ESG scores by demographic category in 2024?*
- *How does it correlate with energy poverty?*

This prompt is particularly useful because it pushes the models to connect ESG discourse with inequality and vulnerability, rather than treating the Social dimension as a generic category. Its value lies in checking whether the models can move beyond abstract fairness language and engage with distributional differences that may shape how ESG sentiment is experienced across population groups.

The final group includes governance and corporate prompts, specifically:

- *Quantify the total economic damages from extreme events and their impact on regional governance scores*
- *Estimate the variance in ESG scores between companies with renewables and without, correlated to greenwashing litigation*

These prompts extend the analysis into the relationship between environmental shocks, governance quality and corporate ESG positioning. They are especially useful because they reveal whether the models remain descriptively grounded or begin to produce numerically plausible but methodologically fragile claims when asked to infer correlations, damages, and score variances.

Overall, the structure of the prompt battery moves from a broad interpretive benchmark to increasingly specific and demanding analytical tasks. This progression is intentional: it allows the chapter to show that the selected LLMs can produce structured ESG outputs, but that their reliability varies as the prompt becomes more granular, more temporal, and more inferential. The preliminary testing phase therefore serves as a methodological bridge between the theoretical discussion of model behaviour and the empirical construction of the final dataset, while also making clear that some prompt types are better suited to synthesis than to quasi-quantitative reasoning.

3.3 – ESG Sentiment Generation

This section presents the standardized zero-shot prompt used to generate the daily series of ESG sentiment. As shown in the Figure 11, the prompt is not constructed as a generic

request but as an ordered sequence of instructions, in which each component serves a specific methodological function and contributes to the construction of the final output.

Generate a daily sentiment time series for [country] regarding climate change and ESG (Environmental, Social, Governance) topics over the period Jan 2024 – Dec 2024, using news articles, social media posts, corporate reports and public documents.
For each day, identify ESG-related themes and assign:
E sentiment score (–1 to +1)
S sentiment score (–1 to +1)
G sentiment score (–1 to +1)
Daily aggregated sentiment (mean of E/S/G)
Add a short note for any key event influencing that day's sentiment (e.g., extreme weather, policy announcements, environmental scandals, corporate ESG disclosures).
Output a table with one row per day: Date | E score | S score | G score | Aggregated score | Key themes/events
Conclude with a brief 150-word summary highlighting: major peaks or drops in sentiment, correlations with notable climate/ESG events, differences in sentiment between early, mid and late 2024

Figure 11 - Color-coded structure of the standardized zero-shot prompt used for ESG Sentiment Generation (author's own construction)

The color-coding adopted does not have a purely graphic function but has been designed to reflect the logical progression of the entire process: yellow identifies the delimitation of the perimeter of the analysis, green the universe of information sources, cyan the classification and attribution of ESG scores, magenta highlights the short qualitative note on the key event affecting each day, red indicates the required tabular output and grey refers to the final summary interpretation. In this way, the visual decomposition of the prompt makes its functional articulation immediately readable and allows it to be interpreted not as a single compact instruction but as an analytical pipeline composed of distinct and consecutive phases.

The first section of the prompt, highlighted in yellow, defines the perimeter of the analysis, explicitly delimiting the reference country, the time frame considered and the object of observation. This initial phase is fundamental because it reduces interpretative ambiguity and ensures that all models operate within the same geographical, temporal and thematic framework, making the answers more comparable with each other. Immediately after, the green section specifies the universe of sources to be used,

including press articles, social media posts, company reports and public documents. This element also serves a clear methodological function, as it orients the model towards a representation of ESG sentiment built on heterogeneous and complementary signals, in line with what has been discussed in previous chapters regarding the informational, composite and non-standardized nature of ESG data. This dimension will be analyzed in more detail later where, through a dedicated prompt, each model has been explicitly questioned on the sources used, so as to make the relationship between the generated output, the underlying knowledge base and the processing logic more transparent. The central core of the prompt is represented by the cyan section, in which the model is asked to identify, for each day, the topics attributable to the three dimensions Environmental, Social and Governance and to assign each of them a score between -1 and +1, accompanied by an aggregate value calculated as the average of the three scores. It is at this stage that the structured logic of the prompt emerges most clearly: the model is not simply asked to describe events or content, but to make a thematic classification, formulate a synthetic evaluation and convert it into a measure comparable over time. It is, therefore, the step in which the text is transformed into data, making the prompt similar to a real coding device. The magenta section introduces an additional layer of qualitative contextualization, as it requires each day to be associated with a brief note relating to any key event that influenced sentiment, such as extreme weather events, regulatory announcements, environmental scandals or ESG disclosures by companies. This component is particularly relevant because it links the numerical value to a concrete explanatory reference, preventing the daily series from being presented as a simple abstract succession of scores and at the same time strengthening the interpretive

readability of the dataset. The red section, on the other hand, performs a structural formalization function, as it specifies that the entire output must be returned in tabular form, with a row for each day and with predefined columns relating to the date, the three ESG scores, the aggregate score and key themes or events. In this way, the prompt not only requests content, but also imposes a standardized format, an essential condition for making the results comparable between models and subsequently transferable to Excel for the construction of the final dataset. Finally, the grey section concludes the prompt with a brief interpretative summary, limited to about 150 words, in which the model is called upon to highlight the main peaks and troughs in sentiment, correlations with relevant climate or ESG events, and the differences between the beginning, middle and end of 2024. This closure does not coincide with a simple descriptive summary but represents a final phase of narrative re-composition and coherence control, through which the model rereads in synthetic form what was previously structured at a daily and tabular level. In Figure 12, it is possible to observe the final output after manipulation in Excel, offering a concrete demonstration of the Grok 4.1 output after the data processing phase.

Date	E_score	S_score	G_score	Aggregated_score	Key_themes_events
2024-01-01	0,15	0,04	0,06	0,08	Routine climate/ESG discussion
2024-01-02	0,12	0,03	-0,02	0,09	Routine climate/ESG discussion; minor event
2024-01-03	0,13	0,15	-0,06	0,07	Routine climate/ESG discussion
2024-01-04	0,05	-0,01	-0,09	-0,02	Routine climate/ESG discussion
2024-01-05	-0,16	0,15	0,08	0,02	Routine climate/ESG discussion
2024-01-06	-0,05	0,01	-0,07	-0,04	Routine climate/ESG discussion
2024-01-07	0,03	-0,16	-0,06	-0,07	Routine climate/ESG discussion
2024-01-08	0,02	0,31	0,04	0,17	Routine climate/ESG discussion; minor event
2024-01-09	-0,02	0,09	-0,06	0	Routine climate/ESG discussion
2024-01-10	0,25	0,19	0,29	0,28	EU Green Deal discussion; optimism; minor event

Figure 12 - Concrete demonstration of Grok 4.1 output after manipulation in Excel (author's own construction)

Overall, the prompt can therefore be read as a miniature methodological pipeline, articulated in six consecutive steps. These steps directly reflect the principles of prompt

design discussed in Chapter 2. The choice to operate in the zero-shot regime responds to the need to observe how each model interprets the same instruction, while the rigorous standardization of the prompt is the necessary condition to ensure inter-model comparability within the research design.

3.3.1 – Data Sources

This section goes into detail about the topic of data sources, an element common to all models: the data have not been directly extracted from official ESG providers (see Chapter 1) rather the models describe the data as the result of a synthetic construction, in which the sources play a role of support, justification or calibration of the scores attributed to individual indicators and countries. This means that the notion of *source* does not coincide, in these documents, with a database observed and downloaded automatically but with the set of information materials that the models claim to have recalled to construct or justify the final data. In this perspective, the relationship between the model's internal knowledge base and the synthetic algorithm becomes the real methodological point to be clarified. The sources do not directly generate the number but contribute to orienting the logic with which that number is constructed.

Claude Sonnet 4.5 is the clearest model in terms of the distinction between real data and synthetic data, because its response opens by explicitly stating that *no external ESG data sources were used* and that the values of the file are not calculated from rating agencies or ESG databases. In fact, the document specifies that the dataset was produced through an algorithmic simulation that applies an *event-driven* logic to verified news events of 2024, converting qualitative judgments into numerical values on a scale of -1 to +1. The

most interesting part, in this case, is that Claude Sonnet 4.5 does not limit itself to denying the use of external sources but also builds a systematic comparison with real ESG methodologies, recalling MSCI, Sustainalytics, BNP Paribas AM, ECPI and IMF to show the distance between the file under analysis and the corporate rating models actually used on the market. In other words, sources have above all a *critical and comparative* function: they serve less to explain how the dataset was concretely constructed and more to clarify what the dataset is not.

GPT-5.2 adopts a different position, closer to the logic of punctual signal justification. The paper states that the values are not the result of a single *sentiment engine* or a reproducible NLP pipeline but of *expert-coded proxy scores* assigned based on documented events from 2024 and then distributed over daily time windows. Here the sources are organized into very clear categories: documentation on extreme climate events to motivate negative variations in E and S, surveys and public opinion materials to support positive shifts in E and sometimes in G, and regulatory or governance documentation to justify changes in the G dimension. GPT-5.2 also provides concrete examples: for Italy it refers to the EIB Climate Survey and its press release as the basis for a positive window of adaptation climate; for Portugal, it cites the IPMA bulletins on hot and dry 2024; for Spain, it refers to the documentation on the DANA floods in Valencia; for Germany, it cites sources on floods in Bavaria and Baden-Württemberg; for Norway, it refers to Climate Action Tracker and CCPI to justify a slight weakening of governance at the end of the year. Precisely because of this structure, GPT-5.2 is particularly useful in the subsection on sources, because it shows well that they act as

an *evidence base for scoring windows*, not as a direct origin of each individual daily value.

Gemini 3 Pro moves in a different direction, more interpretive and closer to the language of contextualization. The text clarifies that the numbers were generated by a Python algorithm and were not downloaded from terminals such as Bloomberg or Refinitiv but immediately adds that the simulation was designed to reflect really geopolitical and climate dynamics of 2024. Gemini 3 Pro builds a real map of references to be used for each country and for each ESG pillar. For Norway, for example, it refers to NBIM and the Longship project; for Germany and Poland, it cites BNetzA, Eurostat and the German Ministry of Economy; for France it refers to RTE and the Ministry of the Interior; for the United Kingdom, to the Climate Change Committee and YouGov/Ipsos polls; finally, for southern Europe, it identifies Copernicus and civil protection agencies as the main references to explain physical vulnerability to extreme events. In addition, Gemini 3 Pro proposes an explicit methodological distinction between proxy sources for E, S and G: IPCC, IEA and Copernicus for the environment; Eurobarometer, Lancet Countdown and local social media; World Bank Governance Indicators, EU Rule of Law Reports and MSCI/Sustainalytics as methodological references for governance. For this reason, Gemini 3 Pro is very useful for clarifying the role of the knowledge base as an interpretive framework, although it remains less precise in translating this framework into a data generation formula.

Grok 4.1 takes a step forward in terms of quantitative formalization, because it makes it clear that the values are synthetically generated according to a methodology calibrated to the climate and political events of the European context 2024. The most relevant point of the document is the breakdown of the data into three components with declared weights:

- seasonal baselines for 60%
- major events for 30%
- daily noise for 10%.

The sources, in this case, enter into the construction of the data in a more technical way: Copernicus and the EEA reports are used to define seasonal climate baselines; real events by country, such as the floods in Emilia-Romagna, DANA in Valencia or the postponement of the coal phase-out in Poland, are used to calibrate negative or positive peaks in scores. It follows that, compared to GPT-5.2 and Gemini 3 Pro, Grok 4.1 attributes a less narrative and more *parametric* role to the sources, because they help to fix the amplitude and direction of the variations within the series. The presence of daily noise, explicitly declared, is also important, because it makes it transparent that a share of the data does not depend on documentary sources, but on an artificial component inserted to increase the realism of the series.

Finally, Sonar proves to be the most transparent. In its response to the prompt, it directly states that *no external sources were used for the provided values* and that the data are entirely synthetic and programmatically generated. It, then, makes explicit the formula for constructing each daily value:

$$Daily\ Score_{c,t} = Country\ Base_c + Random\ Variation_{c,t} + Event\ Peaks_{c,t}$$

with simple mean of E, S and G for the aggregate score. Here the role of sources is described very clearly: they are not used to produce values directly, but possibly to suggest real alternatives or to inspire event peaks, while the structure of the data remains explicitly synthetic. In methodological terms, Sonar is therefore the model that best separates three distinct levels: the general knowledge base, the algorithmic logic of generation and the possible comparison with real external sources such as EEA, Deutsche Bank or World Bank ESG. Precisely because of this clarity, Sonar can be placed at the top of the hierarchy of methodological transparency, followed by Grok 4.1 for the percentage breakdown of the data, then by GPT-5.2 for the richness in the association between sources and events, by Gemini 3 Pro for the ability to contextualize and finally by Claude Sonnet 4.5, which is very strong in comparative criticism but less detailed in formalizing the generative process.

In general, the comparative analysis therefore makes it clear that data sources did not play the same role in all models. In some cases, such as Claude Sonnet 4.5 and Sonar, they serve above all to delimit by contrast the synthetic nature of the dataset; in others, such as GPT-5.2 and Gemini 3 Pro, they constitute the documentary framework that justifies the variations in scores; in Grok 4.1, on the other hand, they also become a quantitative calibration mechanism for the components of the series. This passage is also important in relation to Chapter 2, because it shows that the different behaviour of the models reflects not only different response styles but also different modes of interaction between internal knowledge, context recovery and formalization of the synthetic data.

3.4 – Final Dataset Construction

From a methodological point of view, the construction of the dataset was divided into two distinct but closely related phases. In a first phase, the sample of countries to be analysed was defined, selecting France, Germany, Italy, Norway, Poland, Portugal, Spain and the United Kingdom, with the aim of obtaining a sufficiently heterogeneous set from a geographical, economic and financial point of view, so as to include both large European economies and smaller but still comparatively relevant markets. Five different LLMs were then queried on each of these eight countries, in order to generate the required daily ESG sentiment data. The results thus obtained were initially organized in Excel files separated by model and country, so as to keep the information production of each provider separate and preserve the full traceability of the observations generated. Subsequently, these files were recomposed in a more orderly structure through an aggregation by provider, obtaining five main sheets, one for each LLM, containing the data relating only to the selected countries. This step constituted the original dataset of the research (see Figure 13 and 14), useful for preserving the identity of the individual models and making possible a direct comparison between the different synthetic outputs.

Date	Country	Provider	E	S	G	AggregateScore
2024-01-01	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-02	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-03	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-04	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-05	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-06	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-07	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-08	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-09	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-10	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-11	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-12	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-13	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-14	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-15	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-16	France	ChatGPT 5.2	0.14	0.10	0.16	0.13
2024-01-17	France	ChatGPT 5.2	0.14	0.10	0.16	0.13

Figure 13 - Original Aggregated Dataset in R (author's own construction)

```
> str(esg)
tibble [14,640 × 8] (S3: tbl_df/tbl/data.frame)
 $ Date             : Date[1:14640], format: "2024-01-01" "2024-01-02" "2024-01-03" "2024-01-04" ...
 $ Country          : chr [1:14640] "France" "France" "France" "France" ...
 $ Provider         : chr [1:14640] "ChatGPT 5.2" "ChatGPT 5.2" "ChatGPT 5.2" "ChatGPT 5.2" ...
 $ E                : num [1:14640] 0.14 0.14 0.14 0.14 0.14 0.14 0.14 0.14 0.14 0.14 ...
 $ S                : num [1:14640] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ G                : num [1:14640] 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.16 ...
 $ Aggregate_Score  : num [1:14640] 0.13 0.13 0.13 0.13 0.13 0.13 0.13 0.13 0.13 0.13 ...
 $ Key_themes_events: chr [1:14640] "ESG reporting ramp-up; CSRD transposed into French law (FY starting 1 Jan 2024); governance and disclo
sure focus" "ESG reporting ramp-up; CSRD transposed into French law (FY starting 1 Jan 2024); governance and disclosure focus" "ESG reporti
ng ramp-up; CSRD transposed into French law (FY starting 1 Jan 2024); governance and disclosure focus" "ESG reporting ramp-up; CSRD transpo
sed into French law (FY starting 1 Jan 2024); governance and disclosure focus" ...
```

Figure 14 - Dataset Structure Obtained with the str () function in R (author's own construction)

In a second phase, this initial archive was reworked to build the dataset currently used in quantitative analyses. Since the empirical objective of the work is to correlate synthetic ESG scores to the performance of European financial markets, it was necessary to align the observations with the actual financial trading calendar. For this reason, non-trading days, in particular weekends, were excluded from the final dataset and each country was associated with its corresponding national benchmark stock index: CAC 40 for France, DAX for Germany, FTSE MIB for Italy, OBX for Norway, WIG20 for Poland, PSI for Portugal, IBEX 35 for Spain and FTSE 100 for the UK (see Figure 15).

Index	Country
DAX	Germany
FTSE_MIB	Italy
CAC_40	France
FTSE_100	United Kingdom
PSI	Portugal
OBX	Norway
WIG20	Poland
IBEX_35	Spain

Figure 15 - Index Returns Map created in R (author's own construction)

In this way, each ESG observation was made consistent with a specific market day, improving the temporal comparability between the variables under study. The daily return of the indices was then calculated and included in the final panel together with the ESG scores and the related provider, so as to obtain a database suitable for both correlation analyses and subsequent linear regressions. To complete the empirical structure, a control variable represented by the *STOXXEurope600* was also introduced in the final dataset. The STOXX methodological documentation places the European indices within a broader family of regional benchmarks constructed according to standardized criteria and based on free float-adjusted market capitalization, with the aim of offering a broad representation of the European stock market. The inclusion of the *STOXXEurope600* in the dataset therefore responded to the need to control the systemic component common to European markets, to distinguish the effects related to the general trend of the area from those more properly attributable to individual national markets.[82] It follows that the final dataset (see Figure 16 and Figure 17) does not represent a simple collection of synthetic scores produced by the LLMs, but a structured panel consistent with the econometric objectives of the research in which the selection

[82] STOXX Ltd. (2025). STOXX® Index Methodology Guide (Portfolio Based Indices). STOXX. https://www.stoxx.com

of countries, the aggregation of files, the temporal cleaning of observations and the integration of financial variables contribute to defining a solid and methodologically justified analytical framework.

Date	Country	Provider	E	S	G	AggregateScore	Index	IndexReturn	STOXX600Return
2024-01-03	France	Sonar	0.568	0.633	0.305	0.502	CAC_40	-0.0159278252	-0.008626262
2024-01-03	France	Gemini Pro 3	0.200	0.050	-0.160	0.030	CAC_40	-0.0159278252	-0.008626262
2024-01-03	France	ChatGPT 5.2	0.140	0.100	0.160	0.130	CAC_40	-0.0159278252	-0.008626262
2024-01-03	France	Grok 4.1	0.250	0.060	0.020	0.110	CAC_40	-0.0159278252	-0.008626262
2024-01-03	France	Claude Sonnet 4.5	-0.061	-0.045	0.003	-0.035	CAC_40	-0.0159278252	-0.008626262
2024-01-03	Germany	ChatGPT 5.2	0.160	0.100	0.140	0.130	DAX	-0.0138690666	-0.008626262
2024-01-03	Germany	Sonar	0.390	0.379	0.176	0.315	DAX	-0.0138690666	-0.008626262
2024-01-03	Germany	Gemini Pro 3	-0.070	-0.230	-0.490	-0.260	DAX	-0.0138690666	-0.008626262
2024-01-03	Germany	Claude Sonnet 4.5	-0.051	-0.035	0.013	-0.025	DAX	-0.0138690666	-0.008626262
2024-01-03	Germany	Grok 4.1	0.200	0.190	0.020	0.140	DAX	-0.0138690666	-0.008626262
2024-01-03	Italy	Sonar	0.100	0.050	-0.050	0.030	FTSE_MIB	-0.0139808204	-0.008626262
2024-01-03	Italy	Claude Sonnet 4.5	-0.100	-0.050	0.000	-0.050	FTSE_MIB	-0.0139808204	-0.008626262
2024-01-03	Italy	ChatGPT 5.2	0.150	0.100	0.050	0.100	FTSE_MIB	-0.0139808204	-0.008626262
2024-01-03	Italy	Grok 4.1	0.130	0.150	-0.060	0.070	FTSE_MIB	-0.0139808204	-0.008626262
2024-01-03	Italy	Gemini Pro 3	0.210	0.110	-0.240	0.030	FTSE_MIB	-0.0139808204	-0.008626262
2024-01-03	Norway	Grok 4.1	0.180	0.100	0.040	0.110	OBX	0.0011097252	-0.008626262
2024-01-03	Norway	ChatGPT 5.2	0.180	0.120	0.140	0.150	OBX	0.0011097252	-0.008626262

83

Figure 16 - Final Dataset after financial manipulation in R (author's own construction)

```
> str(final_db2)
'data.frame':   9950 obs. of  10 variables:
 $ Date          : Date, format: "2024-01-03" "2024-01-03" "2024-01-03" "2024-01-03" ...
 $ Country       : chr  "France" "France" "France" "France" ...
 $ Provider      : chr  "Sonar" "Gemini Pro 3" "ChatGPT 5.2" "Grok 4.1" ...
 $ E             : num  0.568 0.2 0.14 0.25 -0.061 0.16 0.39 -0.07 -0.051 0.2 ...
 $ S             : num  0.633 0.05 0.1 0.06 -0.045 0.1 0.379 -0.23 -0.035 0.19 ...
 $ G             : num  0.305 -0.16 0.16 0.02 0.003 0.14 0.176 -0.49 0.013 0.02 ...
 $ AggregateScore: num  0.502 0.03 0.13 0.11 -0.035 0.13 0.315 -0.26 -0.025 0.14 ...
 $ Index         : chr  "CAC_40" "CAC_40" "CAC_40" "CAC_40" ...
 $ IndexReturn   : num  -0.0159 -0.0159 -0.0159 -0.0159 -0.0159 ...
 $ STOXX600Return: num  -0.00863 -0.00863 -0.00863 -0.00863 -0.00863 ...
```

Figure 17 - Final dataset structure obtained with the str() function in R (author's own construction)

83 The index return columns were calculated as log returns, $r_t = \log(P_t) - \log(P_t - 1)$ by applying the *diff(log(x))* function in R

CHAPTER 4 – DATA ANALYSIS

4.1 – Descriptive Analysis

Chapter 4 moves from dataset construction to empirical examination. Consistently with the framework developed in Chapters 2 and 3, the analysis begins with a descriptive investigation of the daily ESG sentiment scores generated by the different LLMs across the selected European countries, with the aim of assessing the basic statistical structure of the synthetic panel before proceeding to the following analysis. The first step is carried out in Excel through pivot tables and graphical representations, in order to obtain an initial overview of the temporal dynamics, cross-country heterogeneity and model-specific dispersion underlying the generated ESG signal. This preliminary phase is crucial to address the core requirements of Research Question 2 (RQ-2), which focuses on how LLM-generated ESG sentiment evolves over time and across different geographic contexts. By analyzing the statistical behaviour of the data before applying inferential regressions, this section shows how the architectural constraints, informational knowledge bases and alignment strategies outlined in Chapter 2 manifest as distinct empirical properties within the original dataset.

The analyzed sample, as presented in Chapter 3, includes a total of 14.640 daily observations, distributed over eight European countries (France, Germany, Italy, Norway, Poland, Portugal, Spain and the United Kingdom) for the entire calendar year 2024. A combined visual inspection of the daily time-series profiles reveals a remarkable lack of uniform convergence across models, confirming that the algorithmic mode of production heavily shapes the data properties.

The empirical analysis of the individual architectures opens with an examination of the series generated by GPT 5.2, as shown in Figure 18, whose graph develops on y-axis (Avg Aggregate Score) included in the range [−0.4; +0.3]. The most immediately visible feature when observing the series is its purely discontinuous pattern, which can be defined as "*stepped*". During large blocks of the year, particularly in the months from January to April and then from June to August, the model completely eliminates the daily information noise, drawing perfectly horizontal lines. In this scenario, the variations do not occur in a fluid way but manifest themselves exclusively through discrete vertical jumps that suddenly reposition the metrics on new levels of equilibrium.

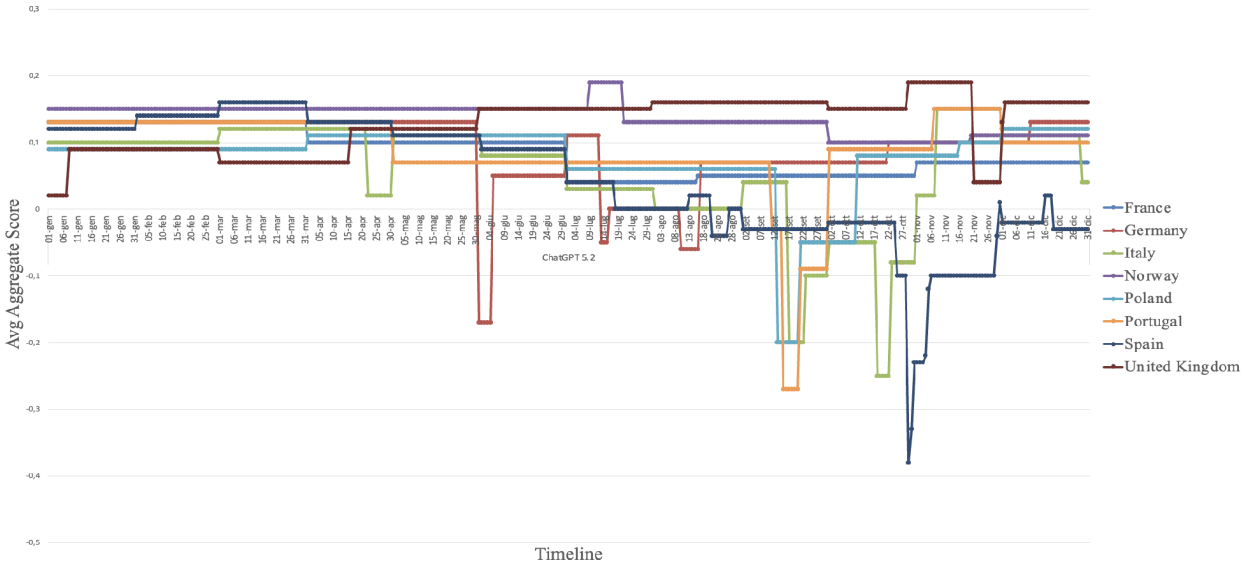


Figure 18 - GPT-5.2 Daily ESG Sentiment Aggregate Score Across European Countries (author's own construction)

In particular, a temporary decline in the German series was observed in the summer months and a contraction in September for the profiles of Italy and Portugal. The most marked cross-country deviation, however, concerns Spain, whose series undergoes a sudden vertical decline at the end of October, touching the lowest point of the entire graph before stabilizing on a new plateau in the final months of 2024. As highlighted in the previous chapter, this specific morphology confirms the nature of GPT-5.2's

information signal, which does not respond to a continuous flow of news but distributes fixed assessments (proxy scores) associated with time windows and linked to macro-events documented during the year.

Claude Sonnet 4.5 's daily series, presented in Figure 19, shows a clearly different structure, projected on y-axis in the range [−0.5; +0.2]. The most evident feature is the strong symmetrical compression of the lines around the neutrality axis. For most of the year, the trajectories of almost all countries remain tightly clustered near zero, indicating a quasi-stationary pattern and a very limited propensity to register daily fluctuations or assign polarized valences to ordinary background events. From a cross-country perspective, geographical differentiation remains largely subdued throughout most of the observed horizon: France, Germany, Italy, Norway, Poland, Portugal, and the United Kingdom move as a single flattened block, with limited dispersion.

A clear divergence emerges only in the last quarter of the year, when Spain separates from the broader European cluster and follows a marked downward trend before partially recovering in December. This pattern is consistent with the role of the alignment filters and Constitutional AI described in the Chapter 2. By promoting caution and ethical moderation, Claude Sonnet 4.5 compresses ordinary variability and shifts the series toward neutrality, while deviations appear mainly in response to exceptional and well-documented territorial shocks. In this sense, the model represents the clearest case of methodological control within the research design.

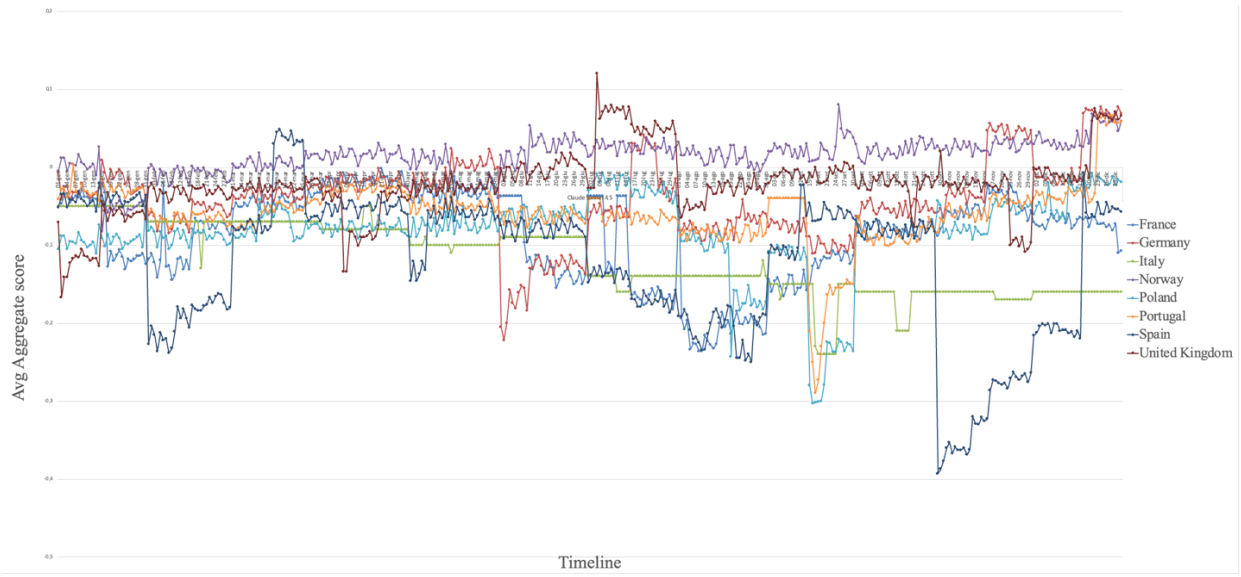


Figure 19 - Claude Sonnet 4.5 Daily ESG Sentiment Aggregate score Across European Countries (author's own construction)

A contrasting scenario emerges from the examination of the series generated by Grok 4.1, whose graph shows the most dynamic and oscillatory structure of the sample, extending on a y-axis between [−0,5; +0.2]. The series shows no clear plateau or compression but instead follows a high-frequency pattern marked by continuous daily variation. At the aggregate level, the European panel traces a broad saddle-shaped trajectory over the year: scores begin at moderately positive levels in January, decline steadily during the summer months, and then recover collectively in the final two months of the year.

Despite this synchronized macro-trend, cross-country differences remain evident. Norway and the United Kingdom display more attenuated and resilient trajectories, whereas Germany, Italy, and Spain show stronger volatility, with repeated negative spikes that temporarily push the scores toward the lower end of the scale. This pattern is consistent with the parametric structure described in Chapter 3: the seasonal trend reflects the transition across seasonal baselines, the continuous jaggedness reflects the

injection of daily stochastic noise, and the intensity of the spikes mirrors the model's tendency to produce more explicit and less censored evaluations in response to critical events.

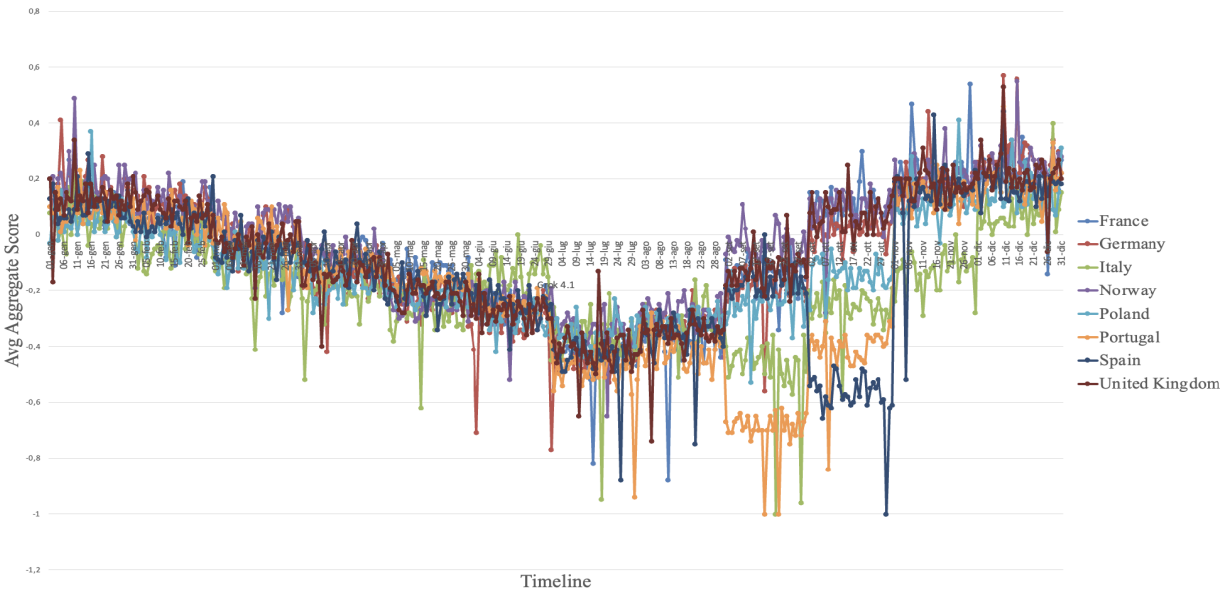


Figure 20 - Grok 4.1 Daily ESG Sentiment Aggregate score Across European Countries (author's own construction)

The analytical framework of the individual providers continues with the examination of the profile generated by Sonar (see Figure 21), whose daily graph remains entirely in the positive or weakly neutral quadrant, within an ordinate range between [0; +0.9]. From a geometric perspective, the series displays a more optimistic average tone than the other models, but it also reveals a marked morphological split that isolates one country from the pattern observed in the rest of the sample.

From a cross-country perspective, the graph highlights two contrasting statistical regimes. On one side, France, Germany, Spain, the United Kingdom, Norway, Poland, and Portugal form a hyper-volatile cluster, with trajectories that are closely intertwined and subject to continuous high-frequency oscillations. Their lines intersect repeatedly across most of the central y-axis, moving between 0.1 and 0.8 and making it difficult to identify a stable long-term trend. On the other side, Italy stands out as an exception: its

trajectory is isolated, follows a step-like pattern, and shows a broadly upward movement in the first part of the year. After starting near zero in early January, it increases at almost monthly intervals, stabilizes throughout the third quarter, rises again in autumn, and then declines slightly in December.

This morphological divergence is consistent with the structure of Sonar's native Retrieval-Augmented Generation architecture discussed in Chapter 3. The volatility observed in most countries reflects the system's real-time web retrieval, which incorporates the instability of live news into the daily score through random variation and event peaks. By contrast, the exceptional regularity of the Italian line suggests a form of context saturation in the pilot case. Since Italy was used for prompt calibration, the zero-shot and zero-temperature setup appears to have induced a more rigid output, limited daily stochastic variation and produced a deterministic upward trend.

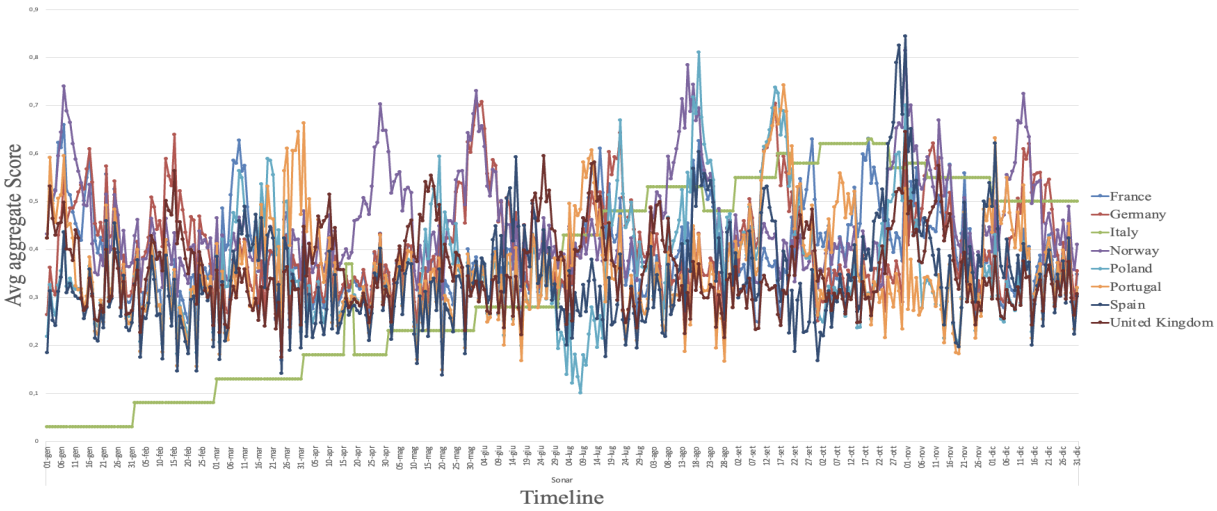


Figure 21 - Sonar Daily ESG Sentiment Aggregate score Across European Countries (author's own construction)

Finally, the examination of the last provider in the sample, Gemini 3 Pro, highlights a mode of time-series decomposition rooted in macro-regional dynamics (see Figure 22). The graph develops within a y-axis range between [−0.8; +0.6]. During the first two

quarters of the year, the country curves remain relatively dispersed but stable: Norway and Portugal fluctuate mainly in positive territory, while Germany stays at lower levels, close to the weakly negative range.

The main structural break appears in September, when the panel clearly separates into geopolitical blocks. France, Norway, and the United Kingdom preserve a coherent trajectory, with a gradual move toward positive values in the second half of the year. By contrast, the Southern and Eastern European bloc, led by Spain and Italy, experiences a sharp decline in early autumn and remains at depressed levels between −0.6 and −0.8 for several months before showing a partial recovery in December. This pattern suggests that Google DeepMind's simulation gives greater weight to structural data and scientific reports from European agencies such as Copernicus and Eurostat, and therefore tends to map territorial vulnerabilities and physical shocks at a macro-regional level rather than isolating single daily events

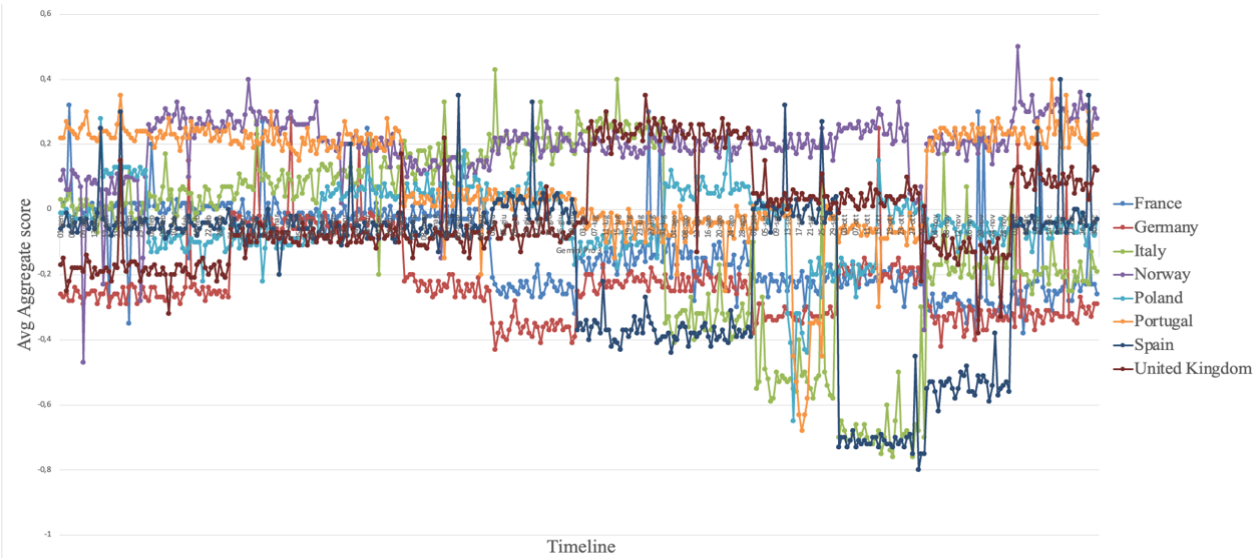


Figure 22 - Gemini 3 Pro Daily ESG Sentiment Aggregate score Across European Countries (author's own construction)

4.1.1 – Variance Analysis

To obtain a more complete picture of the first part of the analysis, variance was also examined as a measure of the dispersion of daily sentiment scores around their monthly arithmetic mean. Monthly aggregation was preferred over daily aggregation in order to avoid the systematic #DIV/0! error produced by Excel. This issue arises when variance is computed on a single observation, since $n = 1$ and the denominator of the sample variance formula become zero ($n - 1 = 0$), making the calculation undefined. By contrast, monthly grouping increases the number of observations to the set of trading days in each month, thereby allowing a finite measure of internal volatility to be obtained. For this reason, the monthly matrices reported below provide a quantitative confirmation of the fluctuations, plateaus, and peaks already observed in the daily series of each LLM.

Provider	ChatGPT 5.2								
Var of Aggregate_Score	Countries								
Months	France	Germany	Italy	Norway	Poland	Portugal	Spain	United Kingdom	Grand Total
Jan	2,96059E-17	2,96059E-17	0	0	0	2,96059E-17	0	0,000885161	0,000664501
Feb	3,37032E-17	3,37032E-17	0	3,96508E-18	0	3,37032E-17	0	0	0,000477056
Mar	2,96059E-17	2,96059E-17	0	0	0	2,96059E-17	0	0	0,000771862
Apr	0	3,44552E-17	0,002172414	0	5,74253E-18	3,44552E-17	3,44552E-17	0,000646552	0,000737055
May	0	0,002903226	5,55112E-18	0	5,55112E-18	0	5,55112E-18	2,90323E-05	0,000796786
Jun	0	0,005785747	0	0	5,74253E-18	0	0	0	0,002294449
Jul	0	0,003618065	0	0,000713978	0	0	0,000409462	3,22581E-06	0,003112257
Aug	2,58065E-05	0,002863226	0	2,96059E-17	0	0	0,000419785	0	0,003452904
Sep	0	0	0,009791264	3,44552E-17	0,011197241	0,015363678	0	0	0,011048703
Oct	0	0,000203226	0,005202366	0	0,003816129	0	0,00829828	0,000144516	0,008097107
Nov	0	0	0,003127471	2,4023E-05	0,000103448	0,000595862	0,002651034	0,005172414	0,007259721
Dec	0	0,000203226	0,000442581	5,55112E-18	0	0	0,000203226	2,90323E-05	0,002567531
Grand Total	0,001111829	0,003040117	0,006231785	0,000460316	0,003117395	0,004095313	0,008416085	0,00169483	0,004336401

Figure 23 - Monthly ESG Sentiment Variance – ChatGPT 5.2 (author's own construction)

As shown in Figure 23, for GPT-5.2 the monthly matrix indicates that variance remains equal to zero, or close to zero, for almost all countries between January and August. This provides quantitative evidence of the rigidity of the daily series and the absence of meaningful short-term variation. Variance increases only in correspondence with discrete model updates triggered by major macro-events. This is reflected in the September rise for Portugal, Poland, and Italy, and in the October peak for Spain, which

is consistent with the sharp downward movement associated with the DANA event in Valencia.

Provider	Claude Sonnet 4.5								
Var of Aggregate_Score	Countries								
Months	France	Germany	Italy	Norway	Poland	Portugal	Spain	United Kingdom	Grand Total
Jan	0,001727462	0,000177665	0	0,000939961	0,000146398	7,82323E-05	6,01892E-05	0,001184247	0,001227061
Feb	0,000951261	0,000153618	0,000124138	6,60271E-05	6,60271E-05	5,16798E-05	0,000540239	4,9E-05	0,00292081
Mar	5,99032E-05	5,78065E-05	0	5,78065E-05	0,000248258	0,000178206	0,00292234	3,42624E-05	0,001032416
Apr	4,85621E-05	5,22862E-05	8,37931E-05	5,14437E-05	5,22862E-05	3,61103E-05	5,28276E-05	0,001061926	0,001129183
May	8,94473E-05	0,000247467	6,66667E-06	0,000400757	0,000107723	6,0628E-05	0,001003032	6,13957E-05	0,001249206
Jun	0,001940552	0,000869275	0	0,000208602	8,34954E-05	6,45575E-05	8,81483E-05	0,000107959	0,003031199
Jul	0,003556832	0,001447058	5,5914E-05	5,3929E-05	0,000110561	0,0001136	0,000353026	0,000967662	0,006022081
Aug	0,001199832	0,000159252	1,29032E-05	0,000108383	0,00195869	5,74925E-05	0,000513316	0,000211557	0,005844559
Sep	0,000367352	0,000233357	0,001577126	0,000244478	0,006023689	0,006654395	0,000773197	6,09299E-05	0,00683589
Oct	7,07118E-05	5,97118E-05	0,000349462	8,47398E-05	0,000130912	0,000120265	0,008536366	0,000273189	0,004012999
Nov	0,000193978	0,001778489	2,54023E-05	4,97931E-05	0,00030349	0,000108685	0,001607085	0,001241476	0,011040108
Dec	0,000125628	0,001867996	0	0,000248946	0,000398359	0,002084346	0,006336858	0,001368983	0,006557297
Grand Total	0,003279921	0,002527401	0,002057089	0,000541951	0,002405597	0,001568189	0,008100769	0,001738607	0,004678896

Figure 24 - Monthly ESG Sentiment Variance – Claude Sonnet 4.5 (author's own construction)

In Figure 24, Claude Sonnet 4.5 displays an exceptionally compressed variance profile, with values remaining close to zero across nearly the entire sample. The only meaningful exception is Spain in October, where variance reaches 0.00853. This isolated increase indicates that only a major exogenous shock, such as the DANA event in Valencia, was sufficient to generate a measurable dispersion in an otherwise highly stable time series

Provider	Grok 4.1								
Var of Aggregate_Score	Countries								
Months	France	Germany	Italy	Norway	Poland	Portugal	Spain	United Kingdom	Grand Total
Jan	0,002594624	0,005624731	0,006047957	0,005651828	0,005553118	0,001904946	0,002895699	0,005965161	0,006059565
Feb	0,00408202	0,002023399	0,00436133	0,002214286	0,001797291	0,002703448	0,002036207	0,002111823	0,00539432
Mar	0,003385161	0,002520645	0,007127312	0,003716129	0,003119785	0,004755699	0,002196989	0,002950323	0,006518846
Apr	0,001674023	0,005102989	0,005706207	0,001778851	0,003154138	0,00162023	0,002639195	0,004322299	0,004906771
May	0,002212258	0,002333978	0,008273118	0,001066452	0,001284946	0,000878925	0,002632903	0,00190129	0,004050044
Jun	0,001582644	0,015104023	0,006524713	0,004561954	0,002148161	0,001570575	0,002661609	0,002087931	0,006740976
Jul	0,005771613	0,002106667	0,015032473	0,004805161	0,003223226	0,008844731	0,007658925	0,00662	0,008809664
Aug	0,010529032	0,001598495	0,004145591	0,002511183	0,001892473	0,003410323	0,007147312	0,006389032	0,006174218
Sep	0,00363954	0,006254713	0,02114954	0,002814828	0,005879195	0,007651034	0,002600575	0,003735747	0,049239454
Oct	0,00292	0,001918495	0,005889892	0,006775914	0,003204516	0,008352903	0,008512473	0,003884516	0,062608985
Nov	0,007666782	0,004301264	0,004705862	0,003306782	0,005915402	0,002313103	0,019818851	0,002397816	0,017167739
Dec	0,006536129	0,007871613	0,006202796	0,004684731	0,004225161	0,005229892	0,003775699	0,007513978	0,008677465
Grand Total	0,048092322	0,053113617	0,035198706	0,044596152	0,031012957	0,077682018	0,05886522	0,046280323	0,051982754

Figure 25 - Monthly ESG Sentiment Variance – Grok 4.1

As presented in Figure 25, Grok 4.1 exhibits persistent volatility throughout the sample, with variance never fully converging to zero and remaining within a relatively narrow baseline range even in months without major events. This background variability is consistent with the stochastic noise introduced by the model. Above this floor, the matrix shows distinct peaks in Germany in June, France in August, Italy in September, and

Spain in November, confirming the presence of event-driven dispersion and explaining the jagged structure already observed in the daily series.

Provider	Gemini Pro 3								
Var of Aggregate_Score	Countries								
Months	France	Germany	Italy	Norway	Poland	Portugal	Spain	United Kingdom	Grand Total
Jan	0,010705591	0,000373118	0,001451183	0,018189892	0,00675957	0,000772903	0,006784516	0,004114624	0,0261465
Feb	0,002335714	0,005966256	0,001709606	0,001604187	0,000889901	0,000468966	0,000364286	0,000933744	0,031259649
Mar	0,003170323	0,005653978	0,002098065	0,000993118	0,000953333	0,000847312	0,001223656	0,000513118	0,019116062
Apr	0,002771954	0,002342644	0,004486897	0,001606437	0,000621724	0,000686897	0,002897586	0,002476897	0,013104909
May	0,001016989	0,000590323	0,002437849	0,002916989	0,001189892	0,003109462	0,006000645	0,003716129	0,0162012
Jun	0,000751264	0,000825402	0,003101609	0,000733448	0,003882299	0,000225862	0,004054713	0,001121379	0,037257976
Jul	0,012198495	0,000573118	0,001396129	0,000525161	0,000805806	0,001376989	0,001738925	0,010362796	0,047662595
Aug	0,002264516	0,000699785	0,003978925	0,000698495	0,001386237	0,002810323	0,000631613	0,004762796	0,048823874
Sep	0,000758161	0,000355057	0,027317241	0,000534368	0,031361954	0,060308506	0,006392644	0,000874023	0,056409608
Oct	0,000612258	0,006704516	0,007217849	0,018046452	0,011539785	0,002224516	0,00286	0,002698065	0,097568785
Nov	0,021847816	0,001053448	0,007108161	0,000624023	0,000705862	0,000658506	0,001676897	0,004085057	0,063295056
Dec	0,008772903	0,01399828	0,00357828	0,001884946	0,00071828	0,002139785	0,013365591	0,001045161	0,047602271
Grand Total	0,01645613	0,013894486	0,082358551	0,00848571	0,011230734	0,028006539	0,061070853	0,020479853	0,047559926

Figure 26 - Monthly ESG Sentiment Variance – Gemini 3 Pro (author's own construction)

Figure 26 displays the Gemini 3 Pro matrix that follows an intermediate pattern, with relatively low variance in the spring and a marked structural break in September. The highest volatility is concentrated in Portugal, followed by Poland and Italy, suggesting a stronger responsiveness to macro-regional information flows and institutional sources. The persistence of elevated values in the following months for Italy and France points to a form of medium-term contextual memory.

Provider	Sonar								
Var of Aggregate_Score	Countries								
Months	France	Germany	Italy	Norway	Poland	Portugal	Spain	United Kingdom	Grand Total
Jan	0,009671473	0,009019525	0	0,01200977	0,003491959	0,011712458	0,004687966	0,006746118	0,025497982
Feb	0,003455739	0,006345187	0	0,002202456	0,003548648	0,005518616	0,004835537	0,005870677	0,014337671
Mar	0,012504013	0,003757591	2,96059E-17	0,002844499	0,010289103	0,015407166	0,009420157	0,002365258	0,014871689
Apr	0,002133651	0,001785344	0,003361034	0,010159926	0,00221162	0,008322671	0,002951034	0,005304547	0,009061213
May	0,002969178	0,01077054	0	0,009420452	0,008179323	0,004654746	0,004119757	0,006727895	0,011090167
Jun	0,002859237	0,015389789	0	0,009994395	0,003259099	0,00455634	0,00670074	0,009592299	0,010212084
Jul	0,008359649	0,007752103	0,000811828	0,001319391	0,022422757	0,012995065	0,00288597	0,009399858	0,011266417
Aug	0,008770981	0,003242998	0,000591398	0,013159673	0,017740123	0,005054746	0,012027226	0,005884095	0,014244253
Sep	0,00885419	0,011703099	0,000393678	0,001804441	0,016236879	0,016812041	0,009579747	0,004598575	0,014627406
Oct	0,00711248	0,002982413	0,000422366	0,01419257	0,01445569	0,009968796	0,032735591	0,009207437	0,017772957
Nov	0,006184179	0,009065895	0,000166552	0,009517206	0,006281683	0,007046516	0,012913126	0,00697743	0,012996414
Dec	0,002814237	0,0095004	0	0,009683213	0,002864673	0,008917925	0,006618323	0,001423946	0,010027011
Grand Total	0,008420615	0,009893906	0,040844174	0,010278094	0,013400932	0,012143884	0,011626065	0,006842736	0,015500205

Figure 27 - Monthly ESG Sentiment Variance – Sonar (author's own construction)

Ultimately, in Figure 27, Sonar's pivot table shows a sharp asymmetry between countries: Spain records a clear October spike linked to DANA, while Italy remains almost flat for the first half of the year, with variance effectively at zero. This suggests

that, in the calibration case, the zero-shot and zero-temperature setup largely suppressed stochastic dispersion, producing a highly deterministic Italian trajectory consistent with the stepped pattern observed in the daily series.

4.2 – Correlation Analysis

To examine how the variables move together, this section evaluates the pairwise correlations between financial returns and the ESG scores generated by the different LLMs. The main objective is to assess whether a linear association emerges between market performance and daily synthetic ESG sentiment. All correlation matrices were computed in R and the complete results for the eight countries are reported in Appendix A in order to preserve readability and allow the main text to focus on the most relevant and distinctive patterns. The analysis is structured along two complementary dimensions. First, it examines whether LLM-generated ESG scores display a contemporaneous association with national market returns and with the broader *STOXXEurope600* benchmark. Second, it assesses the internal consistency of the ESG scores by analysing the correlation among providers within the same country, in order to identify whether the different models converge on a similar representation of ESG sentiment or instead produce systematically divergent outputs.

4.2.1 - Financial Market Interaction Analysis

The dominant pattern across all countries and matrices is a null or extremely weak correlation between daily stock returns and daily ESG sentiment scores. As reported in the complete Figures in Appendix A, almost all correlation coefficients between stock returns and LLM-generated ESG sentiment remain confined to the range between -0.12

and +0.10. From a statistical perspective, coefficients close to zero indicate the absence of any meaningful linear association. Substantively, this means that a positive or negative ESG score generated by an LLM model on a given day does not coincide with same-day stock price movements. In other words, daily market returns and daily LLM-generated ESG sentiment appear to follow largely separate paths.

Although this null pattern is the rule, a few country-specific deviations stand out. The United Kingdom is the clearest weak negative outlier, with correlations coefficients reaching -0.11 under GPT-5.2 and -0.12 under Sonar's Social pillar. In practical terms, this suggests that, in the London market, stronger LLM-generated sustainability sentiment tends to coincide slightly with days of marginal market declines. Spain, by contrast, displays a weak but consistent positive association under GPT-5.2, with a coefficient of +0.10 for the aggregate score. In this case, positive sustainability sentiment appears to align modestly with days of market growth.

Figure 28 presents the Italian correlation matrix, with a focus on Claude Sonnet 4.5 which is particularly informative because it departs from the more regular structure observed in the rest of the sample. While the other models generally show strong positive association among the ESG pillars, Claude Sonnet 4.5 reports a pronounced negative correlation between Governance and the other dimensions, with coefficients of -0.63 with Environment and -0.56 with the *AggregateScore*. Italy is therefore useful not because it reveals a market-driven association, but because it exposes a model-specific asymmetry in the internal structure of ESG.

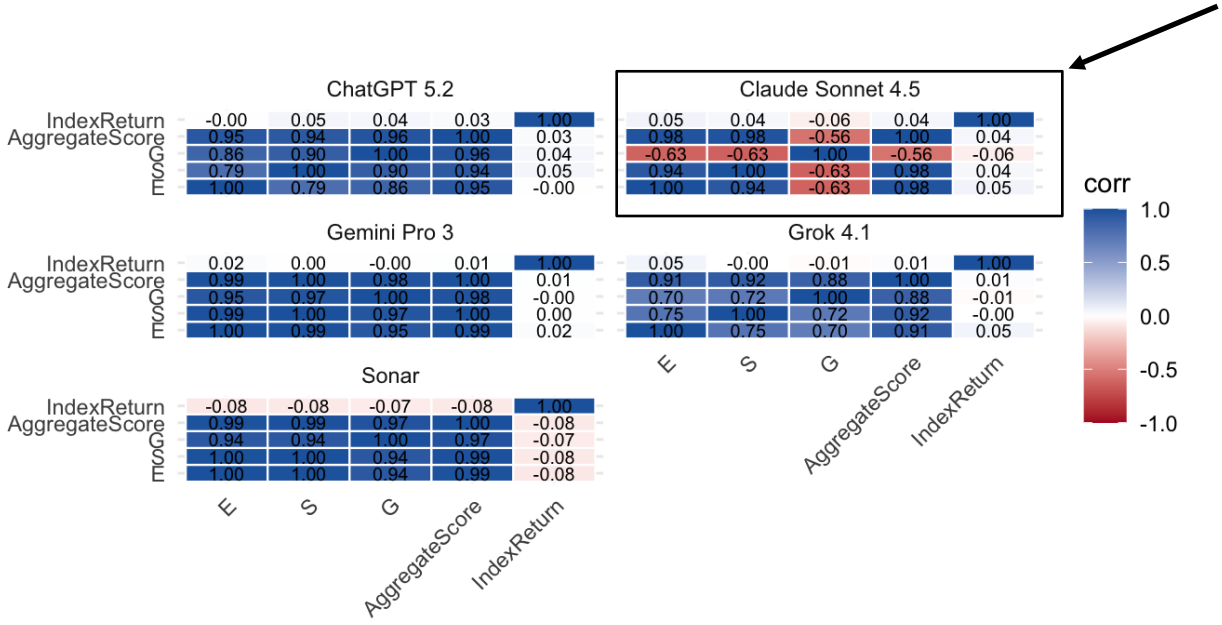


Figure 28 - Evidence from the Italian Case study - Claude Sonnet 4.5 (author's own construction)

The second financial variable considered is the pan-European benchmark, *STOXX600Return*. In methodological terms, its role is very similar to that of the national index returns, since it allows the analysis to capture the broader European market component alongside the domestic one. When *STOXX600Return* is correlated with the LLM-generated ESG scores, the same null or extremely weak pattern observed for the national indexes emerges, with coefficients generally remaining between -0.12 and +0.10. This suggests that broad European market movements do not appear to influence the daily generation of AI-based ESG sentiment

The only minor exception is Norway, in Figure 29, where Claude Sonnet 4.5 records a weak negative association between *STOXX600Return* and the Governance pillar, which falls to -0.13 during phases of downward pressure in the broader European market. However, this deviation remains limited in magnitude. Overall, *STOXX600Return* performs effectively as a control variable, confirming that daily financial movements and daily LLM-generated ESG scores remain statistically independent in the sample.

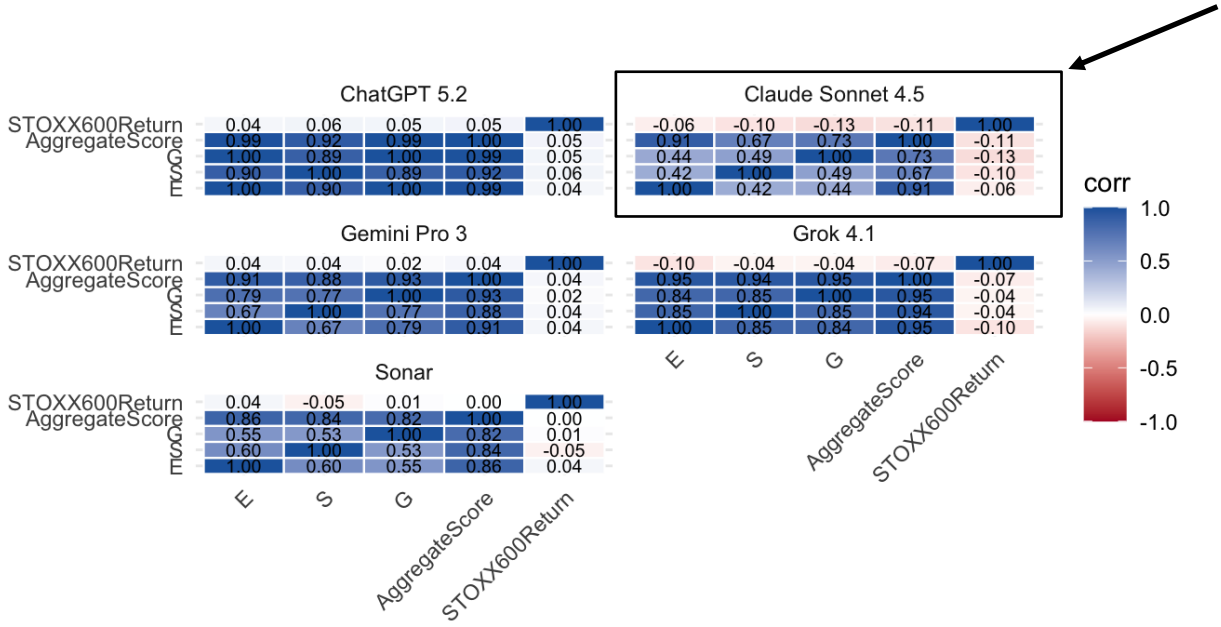


Figure 29 - Evidence from the Norway Case Study - Claude Sonnet 4.5 (author's own construction)

Once the financial controls have been considered, the analysis moves one step deeper into the synthetic data itself, through a cross-provider consistency check. Unlike the previous matrices, the following section does not examine the association between ESG sentiment and market variables, but rather the degree to which different LLMs converge when exposed to the same informational context. In this sense, Figure 30 functions as a pilot case for the broader provider-by-provider comparison, since it allows us to observe whether the generated ESG signal is shared across models or remains strongly model-specific.

4.2.2 - Cross-Provider Correlation Analysis

The finding of a widespread null or extremely weak correlation between the macro-financial variables and the daily sentiment scores requires a further step within the dataset itself. This second level of analysis is directly connected to the previous one: if market variables do not show a strong linear association with LLM-generated ESG sentiment, it becomes necessary to verify whether the synthetic signal is actually stable and shared across providers or whether the absence of an immediate financial response

reflects an internal fragmentation in the way the models interpret the same daily information. In theoretical terms, if the algorithms produce divergent judgments in response to identical events, the lack of a strong correlation with stock market returns can be explained by the non-uniform nature of the signal itself, which is therefore not absorbed by markets as a single coherent input.

To examine this issue, the analysis turns to the cross-provider correlation matrices computed on the aggregate scores (*AggregateScore*) for each country. The full set of eight national matrices is reported in Appendix A to preserve the readability of the main text. As anticipated above, the Italian matrix is presented as the pilot case in Figure 30 because it offers a particularly clear visual contrast between zones of consensus and zones of divergence across providers.

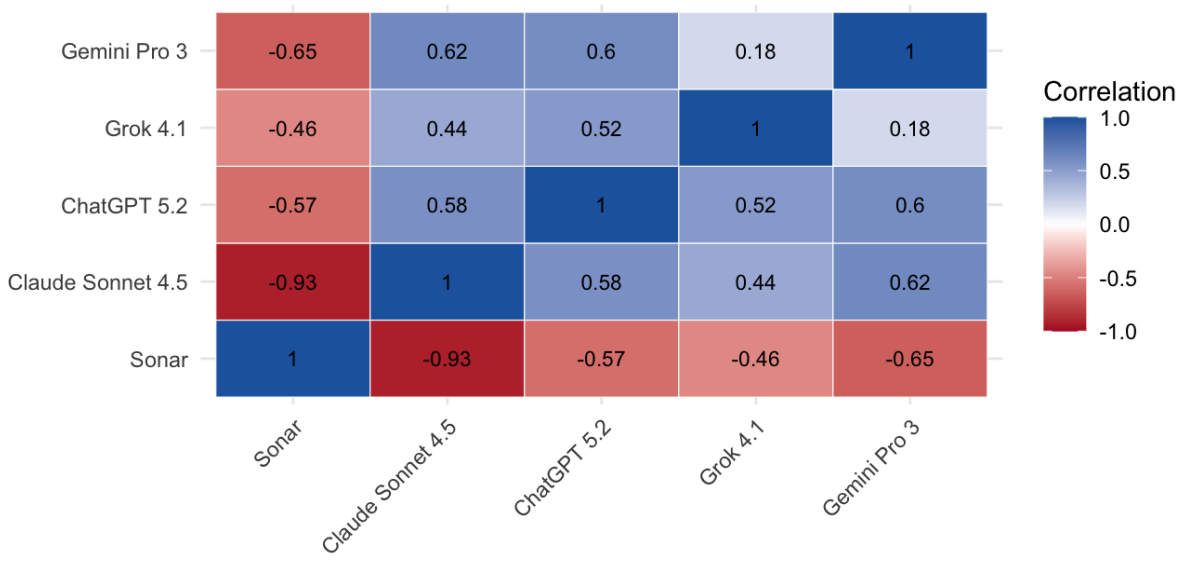


Figure 30 - Correlation of ESG Aggregate Scores Across Providers – Italy (author's own construction)

Across the European panel, two main statistical patterns emerge. The first is the relatively stable consensus among the core commercial models, namely GPT-5.2, Claude Sonnet 4.5 and Gemini 3 Pro. These providers consistently show moderate to

strong positive correlations, especially in continental European countries. Portugal represents the clearest case of convergence, where Gemini Pro 3 displays a strong positive correlation of 0.82 with both GPT-5.2 and Grok 4.1, while Claude Sonnet 4.5 shows a correlation of 0.76 with GPT-5.2. Similar structures appear in Spain, France and Poland, where Claude Sonnet 4.5 shows a correlation of 0.66 with GPT-5.2 and 0.63 with Gemini 3 Pro. This recurring pattern suggests that these models share a common interpretive core, shaped by comparable pre-training weights and similar institutional alignment mechanisms, which tend to produce synchronized readings of daily news and ESG-related information. The second common pattern is the systematic separation of retrieval-augmented architectures, especially Sonar. In almost all matrices, Sonar shows weak, near-zero or negative correlations with the core models. This is not an isolated anomaly but a structural feature of the dataset, and it highlights a fundamental divide between static models and systems driven by live search retrieval. Sonar therefore appears to process the same informational environment through a different pipeline, one that is more exposed to immediacy, volatility and local framing effects.

Beyond these general tendencies, the data also reveal some uncommon patterns that expose the interpretive limits of specific architectures. The most striking case is the Italian matrix, where Sonar shifts from weak independence to a strong negative correlation with all core models, reaching -0.93 with Claude Sonnet 4.5, -0.65 with Gemini 3 Pro and -0.57 with GPT-5.2. This inversion is methodologically important because it shows that Sonar does not merely differ from the other systems in degree; in some contexts, it behaves in the opposite direction. The most plausible explanation lies

in its Retrieval-Augmented Generation (RAG) architecture, which draws directly on the volatility and emotional tone of local press coverage, whereas the core models rely on more stable long-term moderation patterns.

A second uncommon pattern appears in Norway and Germany, where the usual core consensus weakens considerably. Norway is the clearest case of fragmentation, with GPT-5.2 showing negative correlations with both Grok 4.1 (-0.45) and Claude Sonnet 4.5 (-0.37). Germany also shows a general flattening of coefficients, with Gemini Pro 3 dissociating from Sonar (-0.35) and maintaining only very weak links with the rest of the panel. These patterns suggest that in some national contexts the models are unable to converge on a shared interpretive frame.

A further model-specific pattern concerns Grok 4.1, whose behaviour is more volatile and context-dependent than that of the other providers. Across the country matrices, Grok 4.1 does not settle consistently within either the core cluster formed by GPT-5.2, Claude, and Gemini, or the retrieval-augmented logic of Sonar; instead, it alternates between convergence and dissociation depending on the national setting. This makes Grok 4.1 the most idiosyncratic model in the panel. In some Southern European cases, Grok 4.1 remains relatively close to the core consensus. This is especially visible in Portugal, where it records strong positive correlations with Gemini Pro 3 and ChatGPT 5.2, and in Italy, where it still moves in a broadly similar direction to the other commercial models, albeit with lower intensity. In these cases, Grok 4.1 contributes to the formation of a coherent synthetic signal. The picture changes in more fragmented or

polarised contexts. In the United Kingdom, Grok 4.1 breaks away from the core group and shows moderate negative correlations with Gemini and Claude, while in Norway it also diverges from GPT-5.2. This suggests that, in certain national settings, Grok 4.1 responds more sharply to local informational inputs, producing outputs that are less aligned with the broader provider consensus.

Taken together, these findings support a cautious but analytically useful conclusion. The cross-provider analysis does not reveal a single, perfectly homogeneous sentiment engine operating behind all synthetic ESG scores. What emerges instead is a layered structure composed of a relatively stable commercial core, a retrieval-sensitive source of divergence, and an intermediate zone of provider-specific volatility exemplified by Grok 4.1. This does not weaken the value of the dataset but on the contrary, it clarifies its nature. The synthetic ESG series are not random outputs, but neither are they interchangeable across models. Their differences are systematic and interpretable, which is why the cross-provider matrices are essential for the rest of the empirical analysis. This evidence also helps explain why the near-zero correlations observed between financial markets and ESG sentiment should be read as signal dispersion across heterogeneous model architectures rather than as the absence of a signal.

4.3 – OLS regression

To complement the correlation analysis presented in the previous section and to address the research question more directly, it is necessary to examine more closely the relationship between ESG sentiment and financial markets. For this purpose, the study employs an Ordinary Least Squares (OLS) linear regression, a standard econometric

technique used to estimate the relationship between a dependent variable and one or more explanatory variables by fitting the line that minimizes the sum of squared residuals. In this setting, the aim is to assess whether an association exists between *AggregateScore* and *IndexReturn*, and whether it remains relevant once broader market dynamics are considered. The empirical analysis is structured around four main specifications, each designed to capture a different aspect of the association under investigation. The first model is a baseline specification, in which *IndexReturn* is explained solely by *AggregateScore*, in order to test whether the aggregated ESG signal has any standalone explanatory power. The second model decomposes the aggregate measure into its three ESG pillars allowing for an examination of whether the individual components behave differently from the composite score. The third model introduces *STOXX600Return* as the main control variable, with the purpose of isolating the systematic component common to European markets. Finally, the fourth model combines the ESG pillars with the same market control, serving as a robustness specification to verify whether any ESG-related relationship persists once the general market trend is accounted for. The results of these four regressions are presented and discussed in the following section.

```
> m1_baseline <- lm(IndexReturn ~ AggregateScore,data = ols_db)
> summary(m1_baseline)

Call:
lm(formula = IndexReturn ~ AggregateScore, data = ols_db)

Residuals:
      Min        1Q    Median        3Q       Max
-0.034131 -0.004780  0.000376  0.005288  0.034246

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.663e-04  8.598e-05   1.934   0.0531 .
AggregateScore 1.206e-04  3.530e-04   0.342   0.7326
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.00841 on 9943 degrees of freedom
Multiple R-squared:  1.174e-05, Adjusted R-squared:  -8.883e-05
F-statistic: 0.1167 on 1 and 9943 DF,  p-value: 0.7326
```

Figure 31 - OLS Model 1 Summary (author's own construction)

In the baseline specification, *IndexReturn* is regressed on *AggregateScore*, the synthetic ESG sentiment indicator constructed as the average of the three ESG pillars, with the aim of assessing whether the aggregated signal has any standalone explanatory power for daily stock index returns (see Figure 31). The estimated coefficient on *AggregateScore* is positive but extremely small in magnitude (1.206e-04) and statistically insignificant (p-value = 0.7326), indicating that changes in the composite ESG sentiment are not associated with systematic changes in index returns within this simple linear framework. The intercept is only marginally significant at the 10% level, and the overall fit of the model is essentially null, as shown by an R^2 of 1.174e-05 and a negative adjusted R^2, which together imply that the regression explains virtually none of the cross-sectional and time-series variation in daily returns. From an economic perspective, these results suggest that, at a daily frequency, the LLM-generated ESG sentiment score does not behave as an autonomous driver of market performance: the information contained in the aggregate indicator either does not translate into contemporaneous price movements or is too weak and noisy to be detected once

embedded in overall market dynamics. For this reason, the baseline model is best interpreted as a benchmark specification, useful to frame the subsequent analysis: the lack of significance at the aggregate level motivates both the decomposition of ESG sentiment into its environmental, social and governance components and the introduction of a broad market control, in order to investigate whether more granular ESG dimensions or the isolation of the systematic market factor reveal patterns that the simple aggregate regression is unable to capture.

```
> m2_esg_components <- lm(IndexReturn ~ E + S + G,data = ols_db)
> summary(m2_esg_components)

Call:
lm(formula = IndexReturn ~ E + S + G, data = ols_db)

Residuals:
      Min        1Q    Median        3Q       Max
-0.034376 -0.004741  0.000354  0.005324  0.034210

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.539e-04  8.707e-05   1.768   0.0771 .
E            1.548e-03  6.728e-04   2.301   0.0214 *
S           -2.147e-03  9.266e-04  -2.317   0.0205 *
G            5.035e-04  5.724e-04   0.880   0.3791
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.008408 on 9941 degrees of freedom
Multiple R-squared:  0.0006095, Adjusted R-squared:  0.0003079
F-statistic: 2.021 on 3 and 9941 DF,  p-value: 0.1087
```

Figure 32 - OLS Model 2 Summary (author's own construction)

When the aggregate ESG score is decomposed into its three pillars, the results reveal a more nuanced pattern (see Figure 32). The environmental pillar enters the regression with a positive and statistically significant coefficient, suggesting that higher environmental sentiment is associated with slightly higher index returns. By contrast, the social pillar exhibits a negative and statistically significant coefficient, indicating an inverse association with returns, while the governance pillar is not statistically significant. This heterogeneity helps explaining why the aggregate score was not

significant in the baseline model, since the three dimensions do not move uniformly and partly offset one another when combined into a single composite measure. Nevertheless, the model's overall explanatory power remains very limited, as reflected in the extremely low R^2, which indicates that ESG pillars account for only a negligible share of return variation.

```
> m3_main <- lm(IndexReturn ~ AggregateScore + STOXX600Return,data = ols_db)
> summary(m3_main)

Call:
lm(formula = IndexReturn ~ AggregateScore + STOXX600Return, data = ols_db)

Residuals:
      Min        1Q    Median        3Q       Max
-0.031906 -0.003243  0.000140  0.003475  0.033790

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.469e-06  6.312e-05  -0.118    0.906
AggregateScore  2.742e-05  2.591e-04   0.106    0.916
STOXX600Return  8.772e-01  9.501e-03  92.325   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.006171 on 9942 degrees of freedom
Multiple R-squared:  0.4616,    Adjusted R-squared:  0.4615
F-statistic:  4262 on 2 and 9942 DF,  p-value: < 2.2e-16
```

Figure 33 - OLS Model 3 Summary (author's own construction)

After introducing *STOXX600Return* as a control variable the coefficient on *AggregateScore* becomes virtually zero and fully insignificant, whereas the market control is highly significant and strongly associated with national index returns (see Figure 33). The large and precisely estimated coefficient on *STOXX600Return* indicates that the common European market trend explains most of the variation in the dependent variable, leaving no independent explanatory content for the aggregated ESG score. The substantial increase in R^2 confirms that the model fits the data far better than the earlier specifications, but this improvement is driven almost entirely by the control variable rather than by ESG sentiment itself. Accordingly, the baseline association suggested by the simpler model does not survive once the systematic market component is considered.

```
> m4_robust <- lm(IndexReturn ~ E + S + G + STOXX600Return,data = ols_db)
> summary(m4_robust)

Call:
lm(formula = IndexReturn ~ E + S + G + STOXX600Return, data = ols_db)

Residuals:
      Min        1Q    Median        3Q       Max
-0.032030 -0.003238  0.000133  0.003465  0.033935

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.336e-05  6.393e-05  -0.209    0.835
E               5.633e-04  4.939e-04   1.140    0.254
S              -7.309e-04  6.802e-04  -1.074    0.283
G               1.138e-04  4.202e-04   0.271    0.786
STOXX600Return  8.770e-01  9.504e-03  92.269   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.006171 on 9940 degrees of freedom
Multiple R-squared:  0.4617,    Adjusted R-squared:  0.4615
F-statistic:  2131 on 4 and 9940 DF,  p-value: < 2.2e-16
```

Figure 34 - OLS Model 4 Summary (author's own construction)

The full specification confirms the dominant role of *STOXX600Return*, while none of the ESG pillars remains statistically significant once the common European market trend is considered (see Figure 34). This indicates that the pillar-level associations observed in the previous model are not robust to the inclusion of a market control, suggesting that the ESG-return relationship is largely absorbed by broader market dynamics. The model’s explanatory power is virtually unchanged relative to the previous specification, confirming that the ESG variables do not add meaningful incremental information once the systematic market component is controlled for. Accordingly, the evidence points to a weak and non-robust ESG association at the daily frequency considered in this study.

Overall, this analysis provides a coherent but cautious picture of the association between LLM-generated ESG sentiment and daily equity returns. While the baseline model finds no standalone explanatory power for the aggregate ESG score, the disaggregated specification reveals heterogeneous pillar-level association that are subsequently not

robust to the inclusion of a broad market control. Once *STOXX600Return* is introduced, the systematic European market component becomes the dominant driver of returns, and the ESG variables lose statistical significance. The full specification therefore indicates that the apparent ESG–return association observed in simpler models is largely absorbed by broader market dynamics. Therefore, within the daily-frequency framework adopted in this study, ESG sentiment appears to be a weak and specification-sensitive signal rather than an independent determinant of market performance.

4.4 – Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised dimension-reduction method that transforms a set of potentially correlated variables into a smaller number of uncorrelated linear combinations, called principal components. The first principal component (PC1) is the linear combination that captures the greatest possible variance in the data, while each successive component explains the largest remaining share of variance subject to being orthogonal to the previous ones.[84] In this sense, PCA is not primarily a predictive technique, but an exploratory tool designed to summarize the main structure of a dataset, reduce noise and reveal whether the information contained in the original variables can be concentrated in a few dominant directions.

In the framework of this thesis, PCA is used to assess whether the ESG sentiment series generated by each LLM exhibit a common latent structure across countries or whether they remain highly dispersed and model specific. This makes this analysis particularly

[84] J. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R* and Python, 2nd ed., Springer, 2023. ISLRv2_corrected_June_2023.pdf

suitable for the present study, because the objective is to evaluate the internal coherence of the synthetic ESG signal rather than to forecast an external outcome. The explanatory power of the first principal component (PC1) therefore becomes a useful diagnostic indicator: when the first principal component (PC1) captures a large share of total variance, the dataset displays stronger commonality; when it captures little variance, the signal is more fragmented and less stable.

For this part of the analysis, the original dataset constructed in the initial phase of the research was preferred to the later version used for correlation analysis and OLS regression, in which weekend observations had been removed to align ESG scores with actual market trading days. That adjustment was necessary for financial econometric analyses, where temporal correspondence with stock market returns is essential, but it was not required for PCA, whose purpose is instead to study the internal statistical structure of the ESG sentiment series themselves. Retaining the original daily dataset therefore made it possible to preserve the full temporal continuity of the synthetic signal and to analyse the latent cross-country structure of the scores without introducing a filter motivated only by market-calendar constraints. The dataset was then organized separately for each LLM and, within each model, into four distinct matrices corresponding to *AggregateScore*, Environmental, Social and Governance. In each table, the columns represent the countries included in the sample, while the rows correspond to daily observations across 2024, thus producing a coherent country-by-time structure suitable for principal component extraction. This organization is necessary because PCA requires a homogeneous matrix in which all variables are

measured according to the same logic and can therefore be meaningfully summarized through orthogonal components that capture the main directions of variation in the data.

The analysis was then carried out by centering the variables and, where appropriate, standardizing them, since PCA is sensitive to the scale of the original variables and tends otherwise to assign greater importance to those with larger variance. Once the components were extracted, particular attention was devoted to the proportion of variance explained by the first principal component (PC1), as this measure provides an immediate indication of how concentrated or fragmented the information contained in the dataset is. In this thesis, the proportion of variance explained by first principal component (PC1) was used as an operational screening criterion: when the first principal component (PC1) accounted for more than 50% of total variance, the structure was considered sufficiently concentrated to justify a more detailed examination of the subsequent components; when first principal component (PC1) remained below this threshold, the corresponding LLM was interpreted as generating a signal that was too dispersed to support a robust latent structure. This threshold should not be understood as a universal statistical rule, but rather as a pragmatic methodological criterion adopted to distinguish between cases in which the first principal component (PC1) clearly dominates the variance structure and cases in which the information is spread too unevenly across multiple directions.

A central element in the interpretation of PCA results is represented by the component loadings. The loadings indicate the weight with which each original variable contributes

to a given principal component and therefore allow identification of the countries that drive the common structure summarized by the component. In the present setting, where countries are treated as variables and dates as observations, the loadings show which national ESG trajectories contribute most strongly to the pattern captured by first principal component (PC1), second principal component (PC2), and the subsequent components. This is particularly useful because it moves the analysis beyond a purely descriptive account of explained variance and makes it possible to identify whether the latent structure is broadly shared across countries or instead driven by a restricted subset of cases.

At the same time, weak first components require careful interpretation. A low proportion of variance explained by first principal component (PC1) may indicate that no strong common latent dimension exists across the countries under examination, but it may also reflect the presence of excessive dispersion, instability, or noise in the ESG scores produced by a given model. For this reason, a weak first principal component (PC1) should not automatically be interpreted as a substantive empirical finding in itself but rather, it should be treated as evidence that the corresponding LLM generates a less coherent cross-country structure, making comparison across multiple models especially relevant for the broader purposes of the thesis. In this sense, PCA does not simply reduce dimensionality, but also functions as a diagnostic tool for evaluating the internal coherence and interpretability of LLM-generated ESG signals.

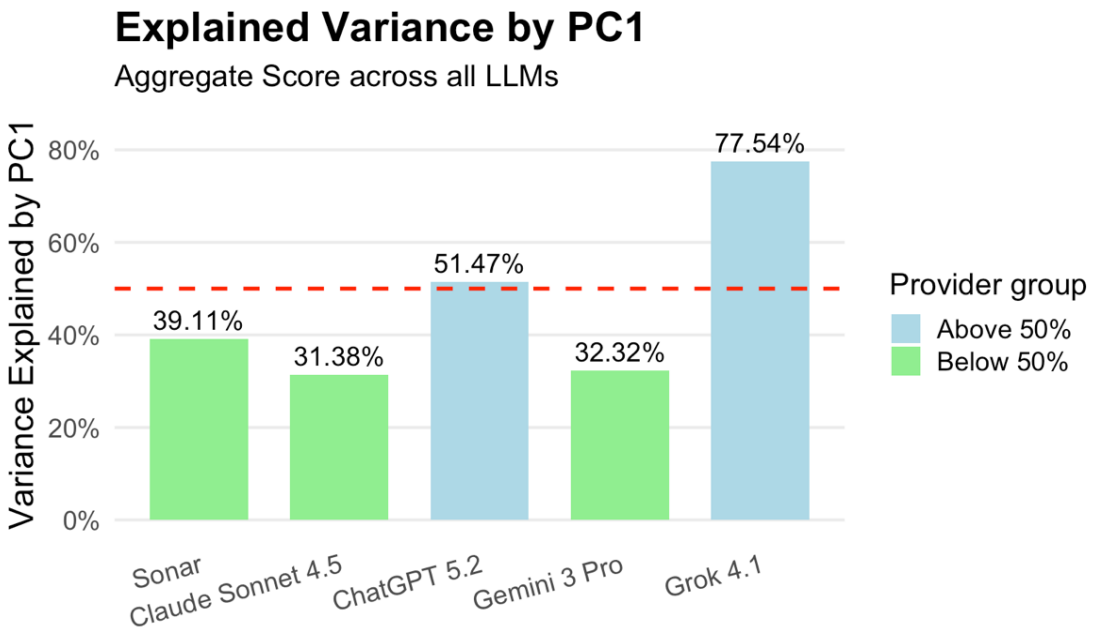


Figure 35 - Explained Variance by First Principal Component (PC1) across all LLMs (author's own construction)

In Figure 35 are presented the PCA results confirming that the models do not generate a homogeneous cross-country ESG structure. Among the five LLMs, only GPT-5.2 and Grok 4.1 exhibit a first principal component (PC1) explaining more than 50% of total variance, which suggests that their *AggregateScore* series are more strongly concentrated along a dominant latent dimension. By contrast, Sonar, Claude Sonnet 4.5 and Gemini 3 Pro remain below the threshold, indicating a more fragmented variance structure and a weaker common factor across countries. This evidence is consistent with the broader pattern emerging from the correlation analysis, where GPT-5.2 and Claude Sonnet 4.5 form part of a relatively stable core cluster, Gemini 3 Pro partially converges with them, Sonar behaves more independently and Grok 4.1 shows the highest volatility and the most idiosyncratic dynamics. From this perspective, PCA does not simply reduce dimensionality, but provides a diagnostic test of how coherent each model's synthetic ESG signal really is.

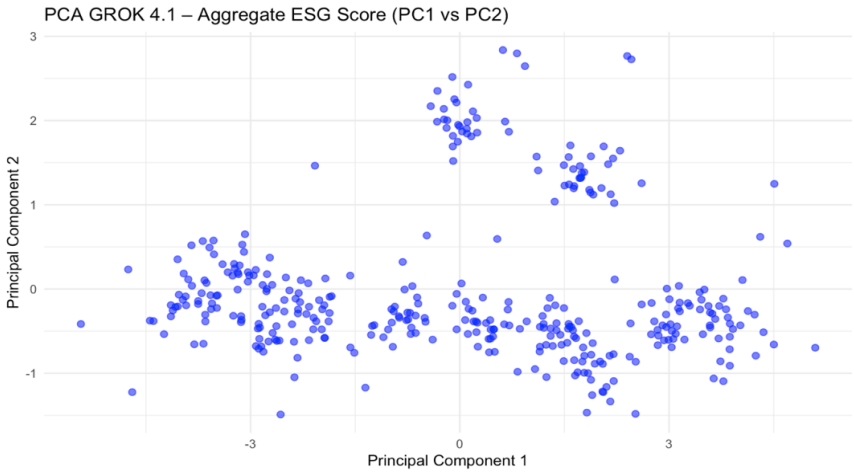


Figure 36 - 2D PCA Representation of the Aggregate ESG Score of Grok 4.1 (author's own construction)

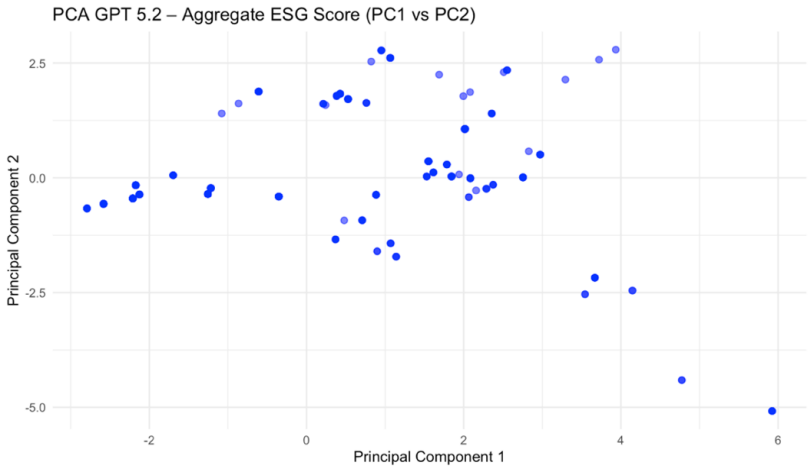


Figure 37 - 2D PCA Representation of the Aggregate ESG Score of GPT 5.2 (author's own construction)

The two-dimensional PCA plots (see Figure 36 and Figure 37) show the position of countries in the reduced space defined by first principal component (PC1) and second principal component (PC2). In the GPT-5.2 graph, the observations are more dispersed, and the cloud is less compact, with some points located far from the centre, especially along the positive side of first principal component (PC1) and the negative side of second principal component (PC2). This suggests a weaker geometric concentration in the first two components and a more fragmented latent structure. By contrast, the Grok

4.1 graph displays a more structured configuration, with denser clusters and clearer groupings of points around specific regions of the plane. This pattern indicates that the first two principal components capture a more coherent organization of the aggregate ESG scores under Grok 4.1.

After the analysis of the *AggregateScore*, the same procedure is applied to the Environmental, Social, and Governance dimensions for the models that already exceeded the 50% threshold at the aggregate level. The results are reported in Table 2, which shows the percentage of variance explained by first principal component (PC1) for each dimension and makes it possible to compare the aggregate structure with its ESG components. At this point, the discussion can move from a purely descriptive reading of the threshold to a more substantive interpretation of whether the latent structure remains stable across E, S, and G or weakens in one of the three dimensions

Provider	Dimension	PC1 (%)
GPT 5.2	E	46.7
GPT 5.2	S	53
GPT 5.2	G	50.1
GROK 4.1	E	74.8
GROK 4.1	S	71.9
GROK 4.1	G	66.1

Table 2 - First Principal Component (PC1) Variance Explained in the ESG Dimensions for GPT 5.2 and Grok 4.1 (author's own construction)

Table 2 shows whether the first principal component (PC1) remains above the threshold in each ESG dimension, and this is the key point for the interpretation. If a model remains above 50% in all dimensions, it suggests a stable and coherent latent structure

across the entire ESG framework. If one dimension falls below the threshold, as happens for GPT-5.2 in Environmental, this indicates that the structure is less concentrated there and that the aggregate result is not fully replicated in every pillar.

Once the explained variance has been compared across the aggregate score and the three ESG dimensions, the analysis can be completed by examining the PCA loadings. The loadings indicate which countries contribute most strongly to the first principal component (PC1) and therefore help identify the underlying pattern behind the values reported in the table. In this way, the interpretation moves from variance capture to substantive meaning.

Country	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
France	-0.361	0.332	0.096	0.304	-0.722	0.175	-0.231	-0.232
Germany	-0.380	0.190	0.061	0.004	0.516	0.083	0.081	-0.732
Italy	-0.310	-0.499	0.632	-0.460	-0.192	-0.042	0.049	-0.054
Norway	-0.366	0.328	-0.044	-0.389	0.252	0.033	-0.629	0.381
Poland	-0.380	0.037	-0.119	0.153	-0.035	-0.887	0.147	0.089
Portugal	-0.337	-0.483	0.064	0.658	0.267	0.210	-0.199	0.248
Spain	-0.316	-0.401	-0.751	-0.296	-0.195	0.177	0.074	-0.113
United Kingdom	-0.371	0.319	0.064	-0.032	0.046	0.313	0.689	0.424

Table 3 - Loadings of AggregateScore for Grok 4.1 (author's own construction)

Table 3 reports the loading matrix for Grok 4.1 across the eight countries. The first point to note is that first principal component (PC1) displays loadings that are all negative and relatively close in magnitude, which suggests that this component captures a broad common structure shared across the entire sample rather than a sharp contrast between a few countries. In this sense, first principal component (PC1) can be interpreted as the

dominant latent dimension of the dataset, and the homogeneity of the coefficients indicates that no single country drives the component in an isolated way. The subsequent components are more selective and therefore more difficult to interpret as general patterns. Second principal component (PC2) separates countries such as Italy, Portugal, and Spain, which have negative loadings, from France, Germany, Norway and the United Kingdom, which have positive loadings, indicating an initial internal contrast within the sample. Third principal component (PC3) is mainly dominated by Italy on the positive side and Spain on the negative side, while fourth principal component PC4 is strongly associated with Portugal; fifth principal component (PC5) highlights a contrast between France and Germany; sixth principal component (PC6) is almost entirely driven by Poland; seventh principal component (PC7) is defined above all by the opposition between the United Kingdom and Norway; and eighth principal component (PC8) is mainly shaped by Germany. These later components therefore seem to capture residual and more idiosyncratic variation rather than a stable cross-country structure. From a critical perspective, this pattern has two implications. First, the fact that first principal component (PC1) is highly coherent supports the idea that the Grok 4.1 output contains a strong common signal across countries, which is useful for dimensionality reduction and substantive interpretation. Second, the increasing specificity of the later components cautions against over-interpreting higher-order PCs, since they appear to reflect narrow country-level contrasts rather than robust latent dimensions. For this reason, the main analytical value of the PCA lies in the interpretation of first principal component (PC1) and, to a lesser extent, the first few secondary components.

In conclusion, the PCA results indicate that the synthetic ESG signal produced by LLMs is not homogeneous but varies significantly in terms of internal consistency and cross-country stability. Among the models analyzed, Grok 4.1 emerges as the most structured case, as it has a dominant main component both in the *AggregateScore* and in the E, S and G dimensions, while GPT-5.2 shows good concentration only at the aggregate level and less robustness in at least one of the three ESG dimensions. Overall, therefore, PCA does not limit itself to reducing the dimensionality of the data but performs an essential diagnostic function: it shows which models generate an interpretable and compact pattern and which instead produce a more fragmented and less stable signal. This evidence is relevant because it suggests that the use of a synthetic ESG score depends substantially on the model that generates it, and that the reliability of the signal must be evaluated not only on a descriptive level, but also on that of its latent structure. The tables of loadings and the PCA graphs shown in the appendix complete this reading, allowing you to check in detail which countries contribute most to the individual components without weighing down the main body of the chapter

4.5 – Main Findings

This final paragraph brings together the different strands of empirical evidence presented in Chapter 4 to provide an integrated assessment of how the five LLMs generate and structure ESG sentiment across European countries. The descriptive statistics and monthly variance profiles show that the synthetic ESG signal is neither uniform nor interchangeable across models: some providers produce comparatively smooth and stable trajectories, while others exhibit higher volatility and marked country-specific fluctuations, indicating that the way ESG-related information is filtered

and aggregated into daily scores depends strongly on the underlying model architecture. The correlation matrices and OLS regressions extend this picture by examining how LLM-based ESG indicators co-move with each other and with financial market indices: while GPT-5.2 and Claude Sonnet 4.5 tend to form part of a relatively stable core cluster, Gemini 3 Pro only partially converges towards this group, Sonar remains more independent and Grok 4.1 displays the most idiosyncratic dynamics, whereas the regression models do not reveal a clear and stable linear relationship between LLM-generated ESG sentiment and countries' index returns once the market-wide control variable is taken into account, with coefficients that are often weak or unstable across specifications and countries. In this sense, the OLS analysis suggests that the incremental explanatory power of the ESG scores over and above general market movements is limited in the present setting, and that any interaction between sentiment and local financial conditions is more heterogeneous and complex than a simple linear association. Taken together, these findings already suggest that the informational content and practical usability of LLM-generated ESG sentiment differ across providers and across countries, even when the underlying news flow is the same and the prompt structure is standardised. The principal component analysis completes this framework by focusing on the internal cross-country structure of the scores: only GPT-5.2 and Grok 4.1 exhibit a first principal component (PC1) that explains more than half of the total variance in the *AggregateScore*, and Grok 4.1 is the only model for which this dominant latent dimension remains strong across the Environmental, Social and Governance pillars, whereas the remaining LLMs generate a more fragmented variance structure and weaker common factors; the loading patterns further reveal that a limited subset of

countries systematically drives the main components, while others contribute more to higher-order contrasts, underscoring that the latent ESG structure is not evenly shared across the European sample.

Overall, the combined evidence from descriptive statistics, variance analysis, correlation and regression models, and PCA indicates that LLM-based ESG sentiment should not be treated as a homogeneous data product: both the association with financial indices and the cross-country patterns of the scores are materially shaped by the specific model used, as documented by the detailed figures and model outputs reported in this chapter and in the appendices.

CONCLUSION

This thesis examined whether Large Language Models can be used as generators of synthetic ESG sentiment and whether the resulting signal can function as a meaningful informational variable in finance. The research stemmed from two converging observations: the growing importance of ESG in academic and financial debate, and the persistent limits of traditional ESG ratings, which often remain fragmented, only partially comparable and insufficiently responsive to the way sustainability-related information evolves over time.

Chapter 1 provided the conceptual foundation for the study. ESG was presented not only as non-financial metric ut also as a form of information whose meaning depends on how it is produced, aggregated, and interpreted. From this perspective, the move from ESG metrics to ESG sentiment was not a simple methodological adjustment. It marked a shift from static, rating-based evaluation to a more dynamic view of how sustainability-related signals circulate within broader informational environments.

Chapter 2 then turned to Large Language Models as analytical tools capable of processing unstructured information at scale, while also emphasizing that different models cannot be assumed to behave in the same way. Rather than treating these models as interchangeable, the chapter emphasized that they differ in meaningful way. Any ESG signal generated through LLMs would be shaped both by the underlying information domain and by the model used to process it.

That insight guided the design of Chapter 3, where a standardized zero-shot prompting framework was developed to construct a daily panel of ESG sentiment for eight European countries over the full calendar year 2024, using five different LLMs. Chapter 4 shows that the resulting signal can be produced in a structured and analytically usable form, but not as a homogeneous output across providers. The descriptive evidence points to substantial variation in the behaviour of the generated series, both across models and across countries, confirming that architecture and alignment logic matter a great deal. Put differently, the same thematic domain does not yield the same empirical series when filtered through different LLMs.

The empirical results also indicate that the relationship between synthetic ESG sentiment and financial market returns is weak. Correlation analysis shows that the association between daily ESG sentiment and country-level stock returns is generally close to zero. Once the *STOXXEurope600* is introduced as a control variable, the explanatory contribution of ESG sentiment becomes negligible, while broader market dynamics dominate the model. At daily frequency, then, the signal does not appear to operate as a robust short-term predictor of market performance.

Even so, the findings should not be read as a rejection of the signal's informational relevance. Although the generated ESG sentiment does not emerge as a strong direct driver of returns, it still appears to capture a distinct informational layer, one that does not fully overlap with conventional financial variables or with traditional ESG rating logic. This becomes particularly evident in the internal coherence analysis. The PCA

results show that the latent structure of the signal varies substantially across models. Grok 4.1 stands out as the most coherent case, with a dominant first principal component (PC1) for both the aggregate score and the individual E, S, and G dimensions. GPT-5.2 displays a more concentrated structure mainly at the aggregate level. The other models show weaker common structures and a more fragmented variance pattern, reinforcing the view that reliability must be assessed separately for each provider.

Taken together, these findings lead to a clear conclusion. Large Language Models can generate synthetic ESG sentiment, and the resulting output can be organized into a structured dataset suitable for empirical analysis. At the same time, the signal should not be understood as neutral, automatic, or interchangeable across models. Its meaning depends on the way it is elicited, on the architecture and alignment strategy of the model, and on the analytical framework through which it is interpreted. The broader implication is methodological as much as substantive. When Generative AI is used to construct financial information, the model itself becomes part of the epistemic structure of the data. This is perhaps the most important contribution of the thesis: not simply to offer a new tool for ESG-related financial analysis, but to show that the act of generating the signal is itself part of the analysis.

FINAL CONTRIBUTION & FUTURE RESEARCH

The final contribution of this thesis lies in proposing and testing a comparative framework for the generation of synthetic ESG sentiment through Large Language Models. Rather than treating ESG sentiment as a variable already available in external databases, this study approached it as a construct that can be actively generated through a standardized interaction with different LLMs and then evaluated through statistical analysis. In this regard, the work contributes not only by producing an original dataset, but also by showing that the process through which the data are generated is itself a central methodological issue.

At both the empirical and methodological level, the research offers a distinct contribution. On the empirical side, it builds a daily panel of ESG sentiment scores for eight European countries over the full calendar year 2024, generated through a common zero-shot prompting framework applied to five different models. On the methodological side, it shows that LLM-generated ESG sentiment cannot be treated as homogeneous across providers, since the resulting series differ in terms of volatility, internal coherence, and latent cross-country organization. What emerges most clearly is that the model is not simply a technical instrument used to retrieve information, but an active component in the construction of the signal itself. Synthetic ESG sentiment, in this sense, is not a neutral output that remains stable regardless of the model used, but a model-dependent informational construct whose characteristics are shaped by architecture, alignment strategy, prompt sensitivity, and, in some cases, retrieval design.

For this reason, the research contributes to the emerging literature on Generative AI in finance and sustainability not by claiming that LLMs already provide a definitive substitute for traditional ESG indicators, but by demonstrating that they open a new methodological space for the production and study of ESG-related information.

At the same time, the results of this work also open several directions for future research. A first natural extension would be to broaden the number of models included in the analysis, incorporating additional frontier systems as well as open-source LLMs, in order to verify whether the patterns observed here remain stable across a wider and more heterogeneous model ecosystem. This would make it possible to distinguish more clearly between provider-specific relationship and broader regularities in the generation of synthetic ESG sentiment. A second possible extension concerns the temporal and geographical scope of the dataset. Applying the same framework to a longer time horizon and to a broader set of countries, including non-European contexts, would help assess whether the weak relationship observed between ESG sentiment and financial market returns is specific to the 2024 European sample or reflects a more general feature of this type of signal. A wider panel would also make it possible to test the behaviour of LLM-generated ESG sentiment across different market regimes, political environments, and sustainability-related event cycles. A further development could also integrate synthetic ESG sentiment with other sources of information. Future studies might compare LLM-generated scores with traditional ESG ratings, news-based sentiment indicators, macro-financial variables, or firm-level data in order to evaluate whether hybrid frameworks can improve interpretability and explanatory power. In the same

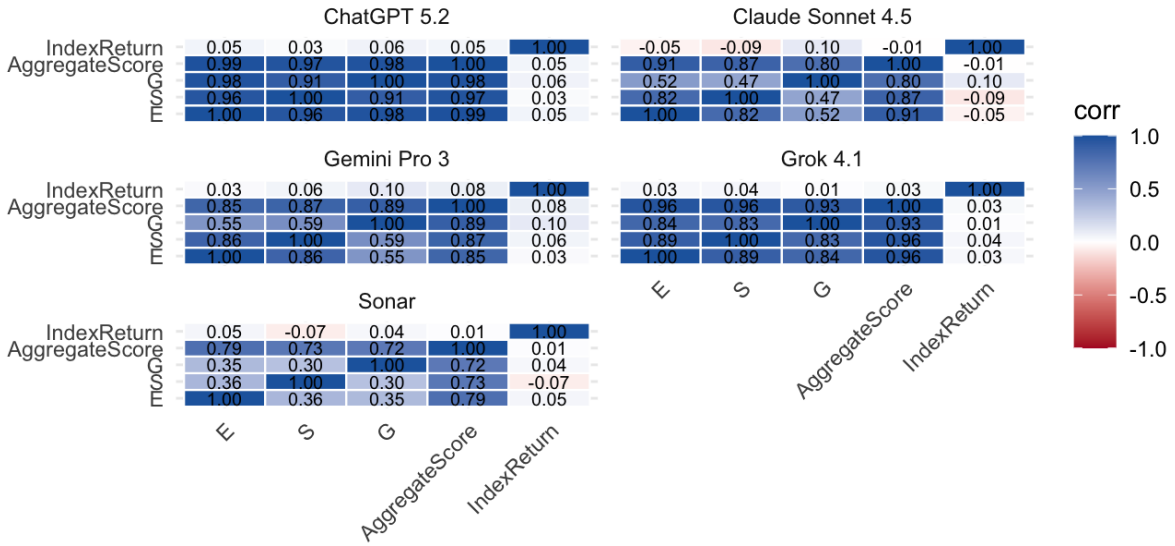
direction, additional robustness checks, alternative specifications, and different lag structures could clarify whether the informational value of ESG sentiment becomes more visible under different empirical settings or at different frequencies of analysis.

To conclude, these findings should be understood as a first step within a broader research agenda. Its main contribution is not limited to the construction of a new dataset but lies above all in showing that LLM-generated ESG sentiment can be treated as a meaningful methodological object in its own right, located at the intersection of ESG measurement, financial analysis, and generative artificial intelligence. Precisely for this reason, the framework developed here can serve as a basis for future work aimed at expanding the number of models, refining the generation process, and better understanding under which conditions synthetic ESG sentiment can become a reliable and analytically useful informational variable.

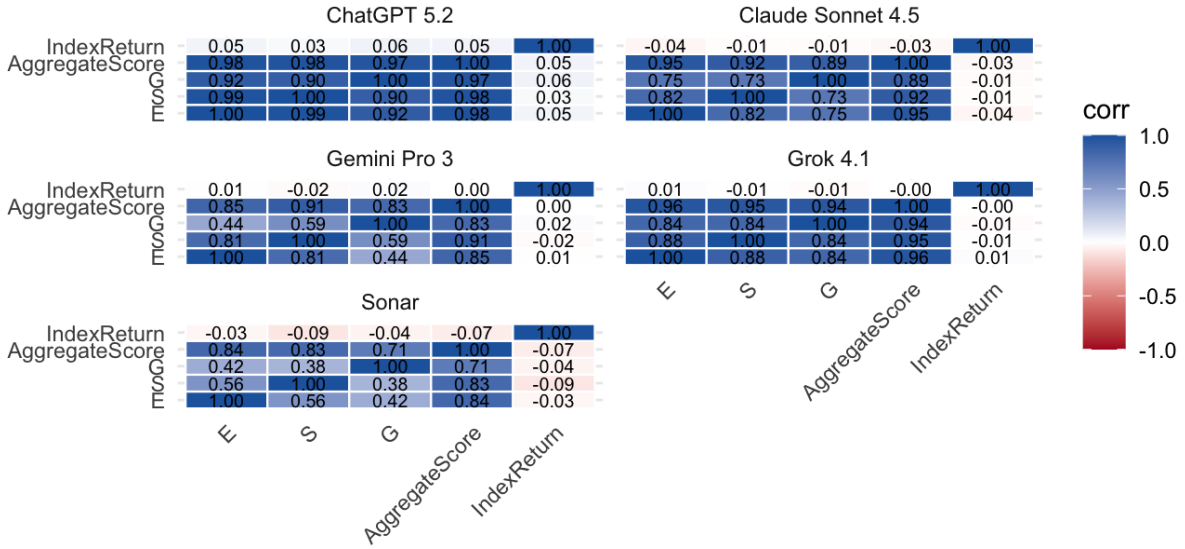
APPENDICES

APPENDIX A – Correlation Matrices

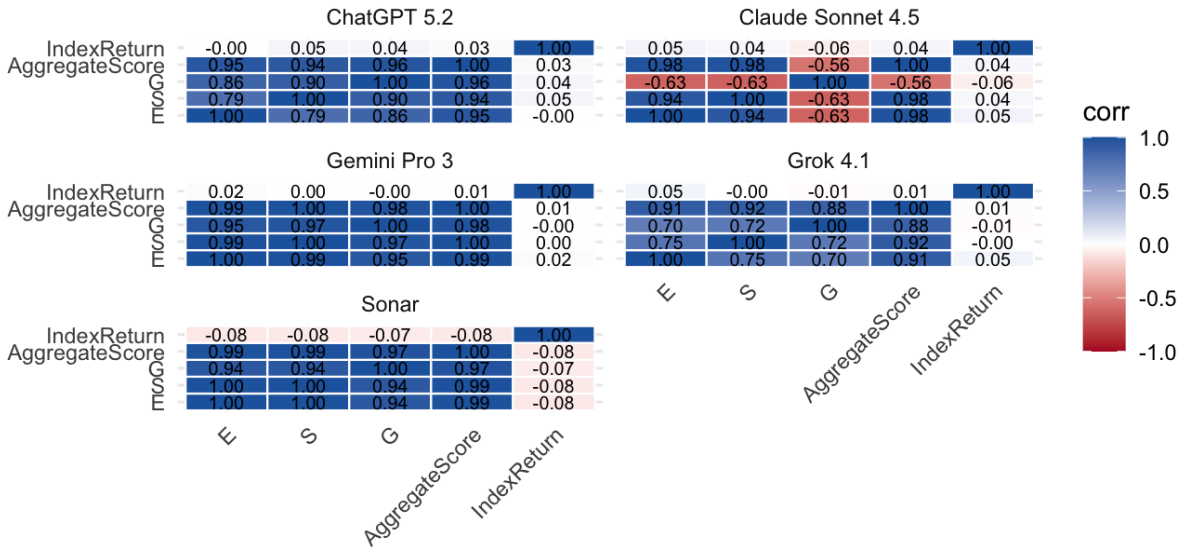
This appendix presents the complete set of correlation matrices relating *IndexReturn* and the *STOXXEurope600* benchmark to the ESG scores generated by the different providers across the full country sample.



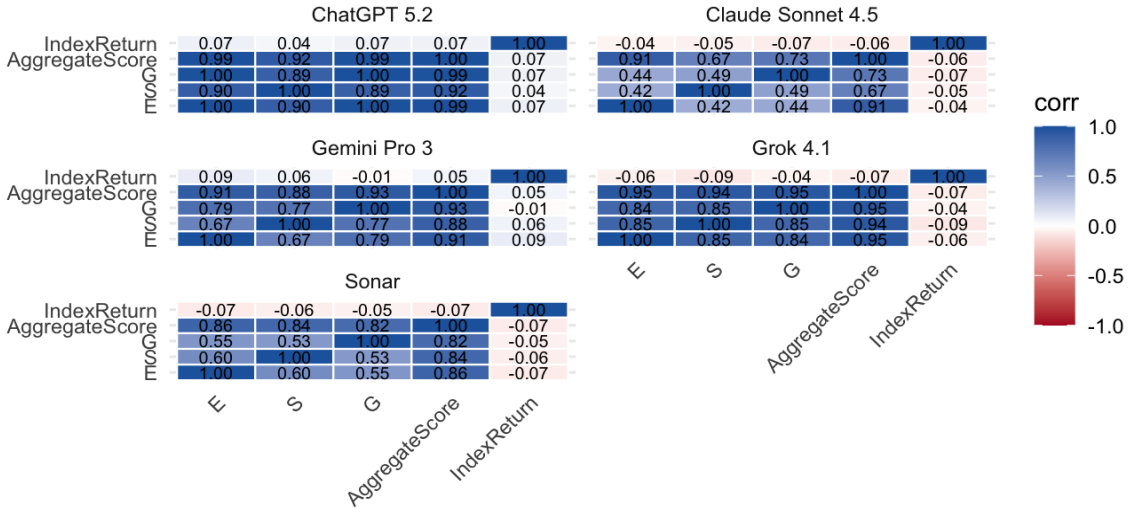
Appendix A. 1 - Correlation Matrices: France – All Providers (IndexReturn) (author's own construction)



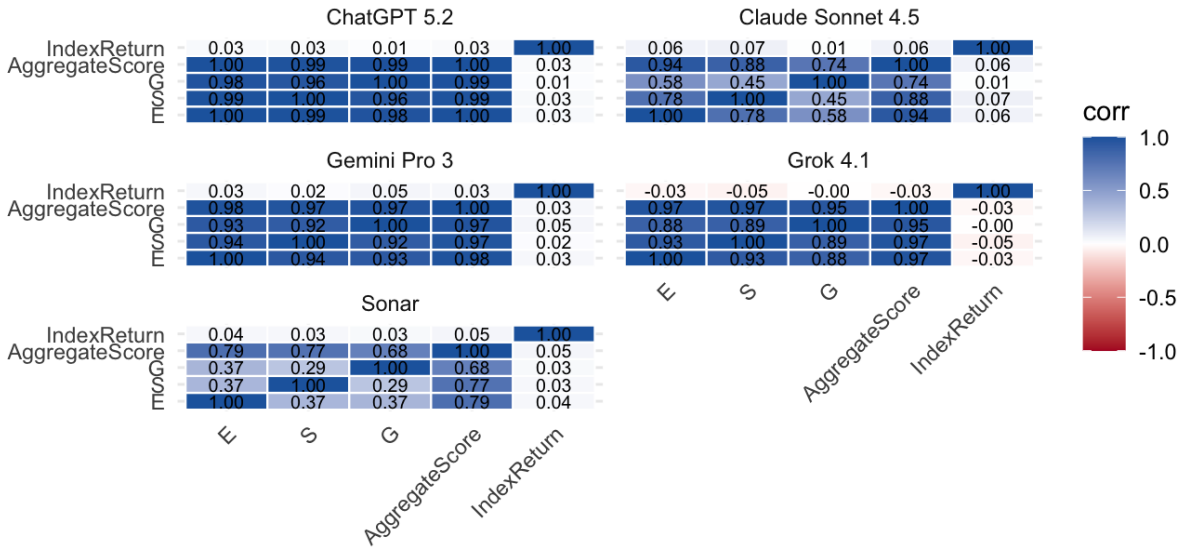
Appendix A. 2 - Correlation Matrices: Germany – All Providers (IndexReturn) (author's own construction)



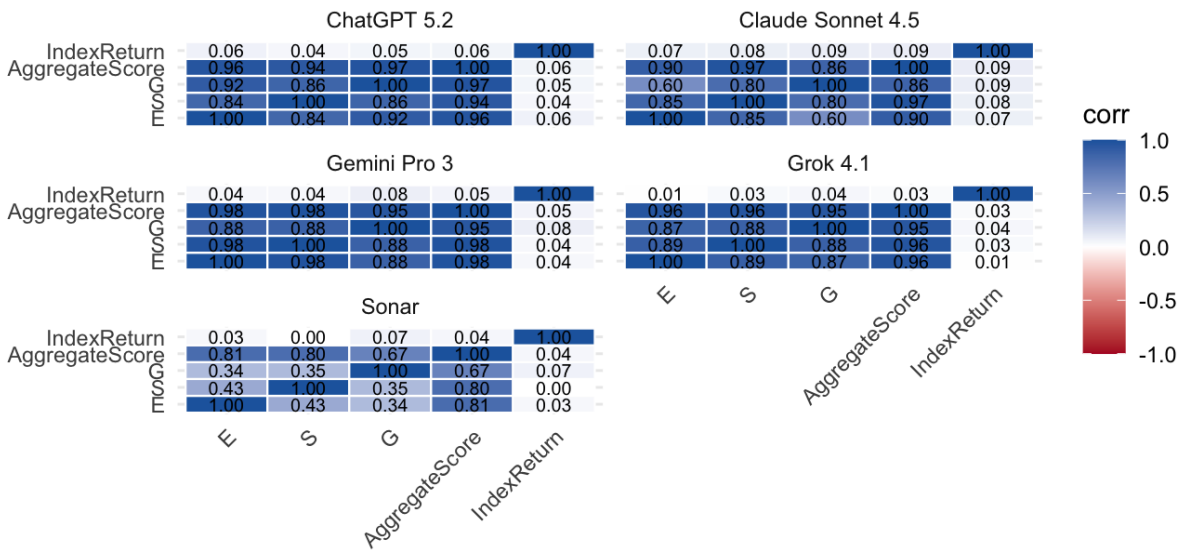
Appendix A. 3 - Correlation Matrices: Italy – All Providers (IndexReturn) (author's own construction)



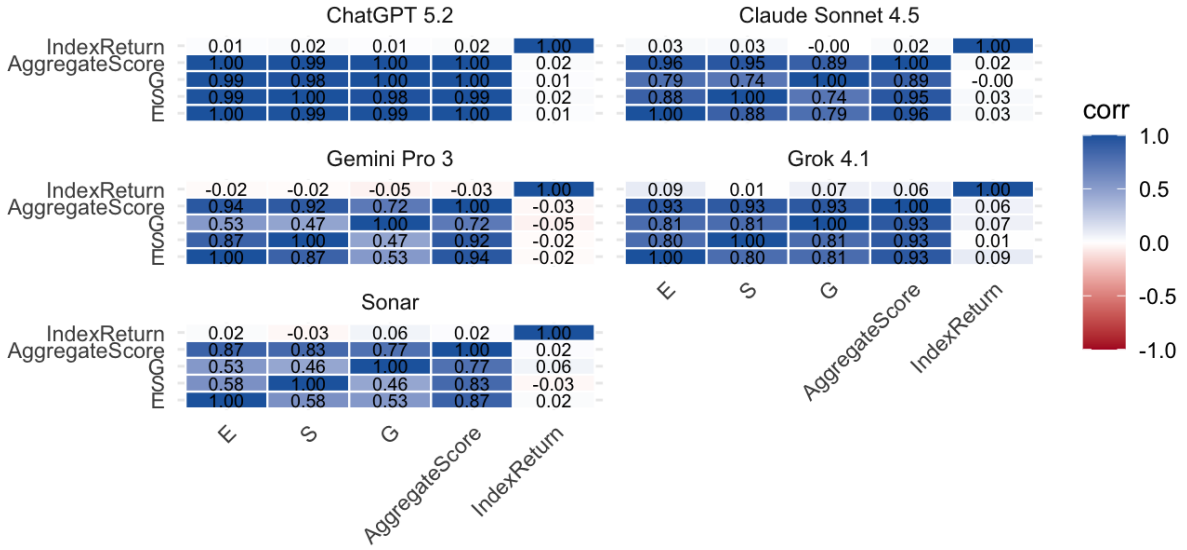
Appendix A. 4 - Correlation Matrices: Norway – All Providers (IndexReturn) (author's own construction)



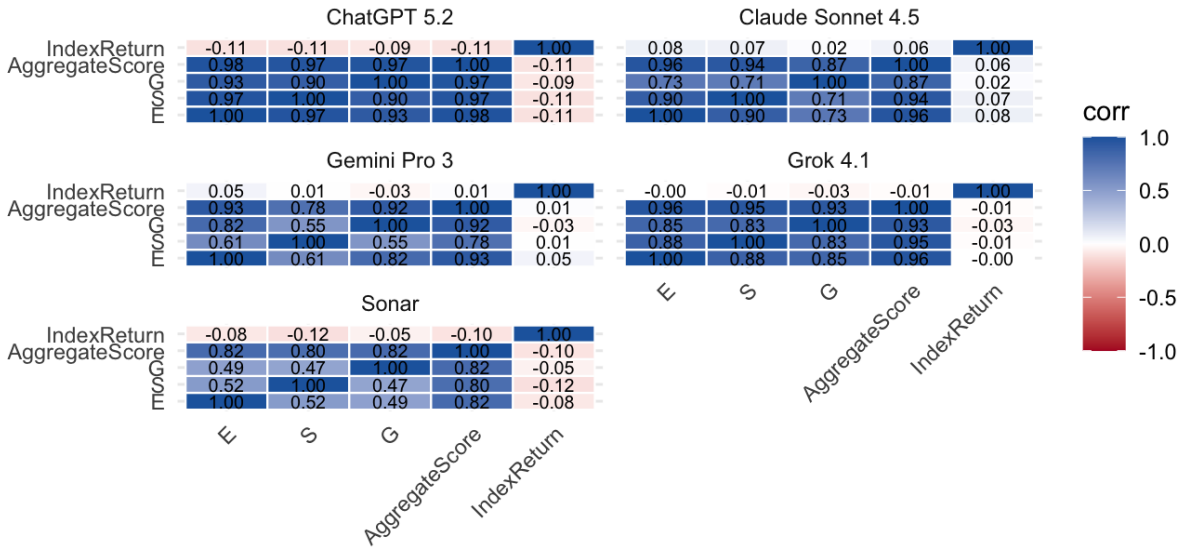
Appendix A. 5 - Correlation Matrices: Portugal – All Providers (IndexReturn) (author's own construction)



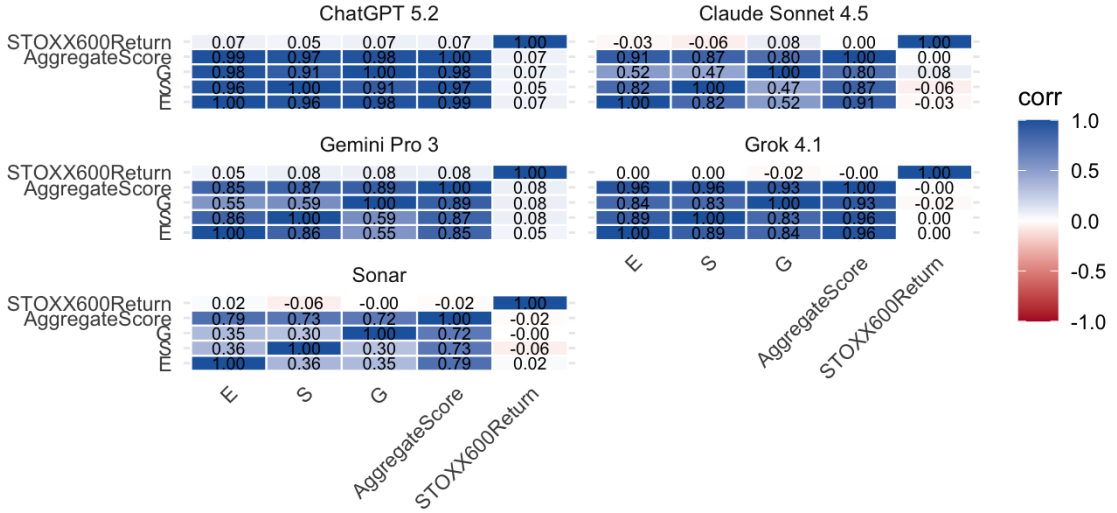
Appendix A. 6 - Correlation Matrices: Spain – All Providers (IndexReturn) (author's own construction)



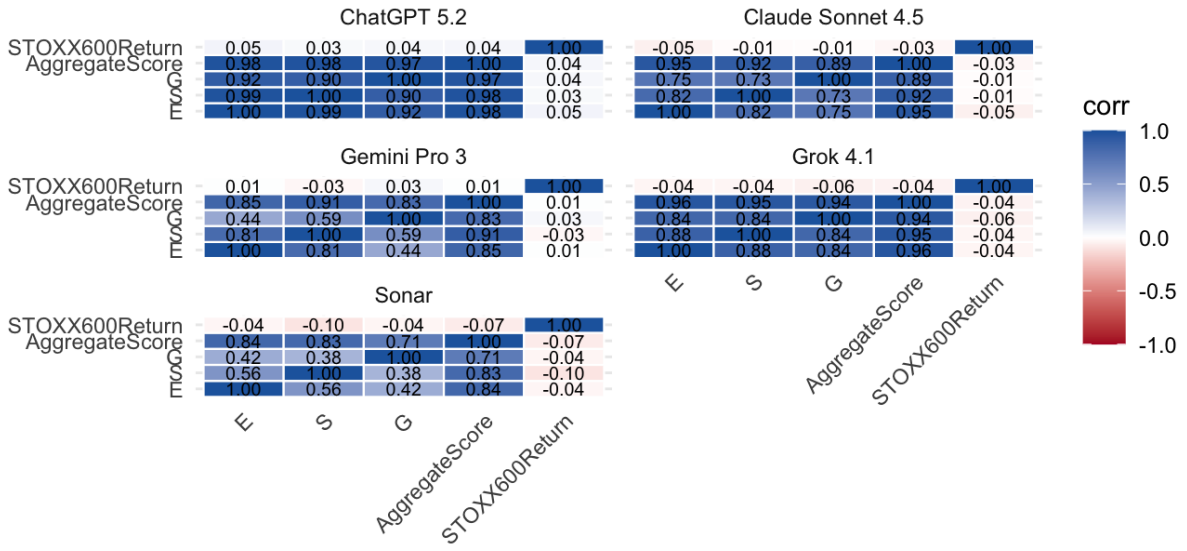
Appendix A. 7 - Correlation Matrices: Poland – All Providers (IndexReturn) (author's own construction)



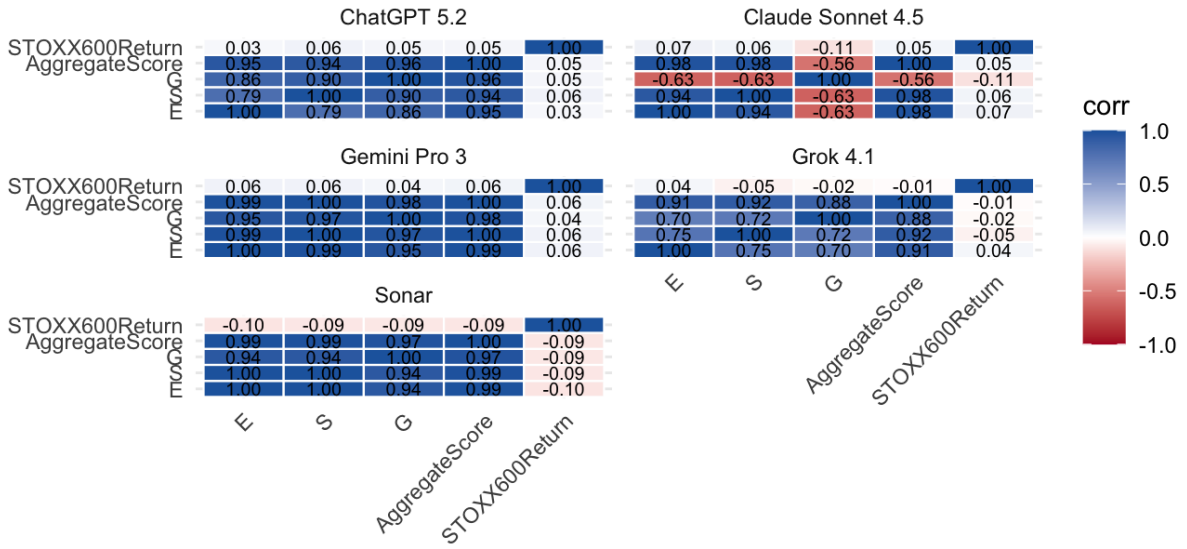
Appendix A. 8 - Correlation Matrices: United Kingdom – All Providers (IndexReturn) (author's own construction)



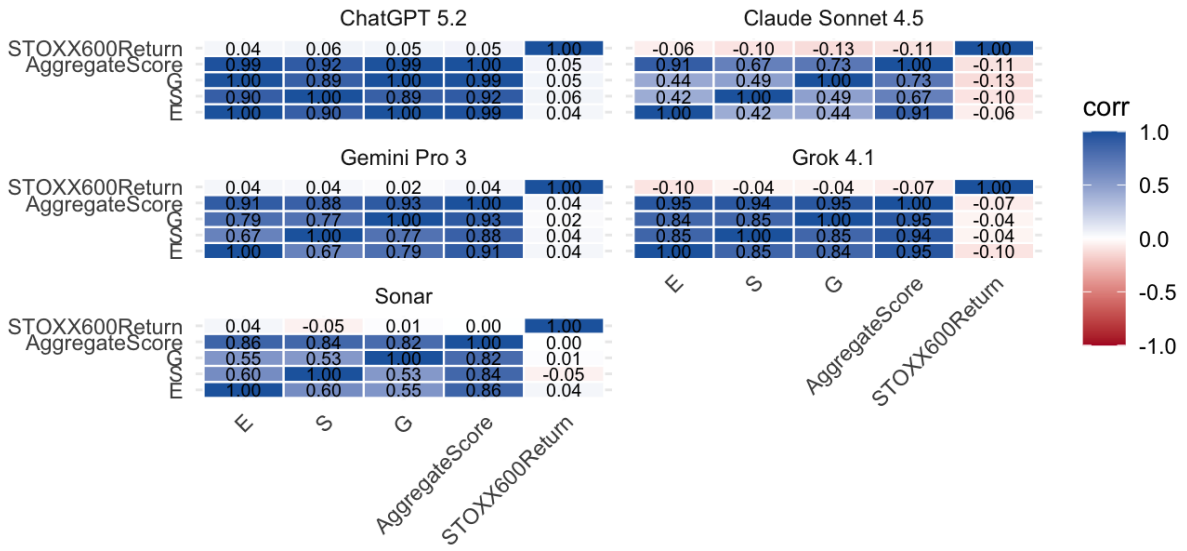
Appendix A. 9 - Correlation Matrices: France – All Providers (STOXX600Return) (author's own construction)



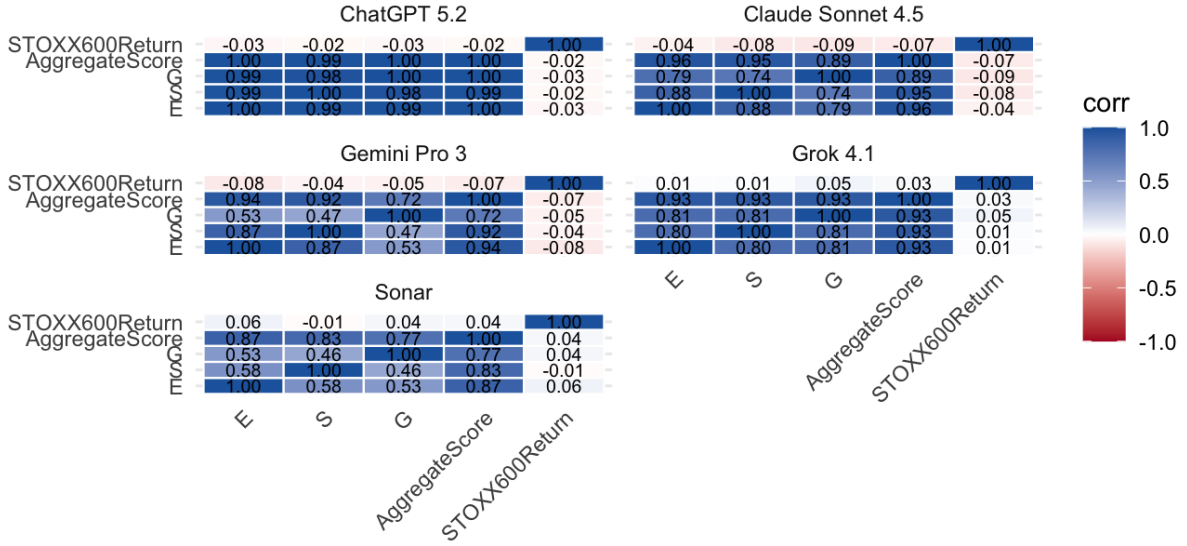
Appendix A. 10 - Correlation Matrices: Germany – All Providers (STOXX600Return) (author's own construction)



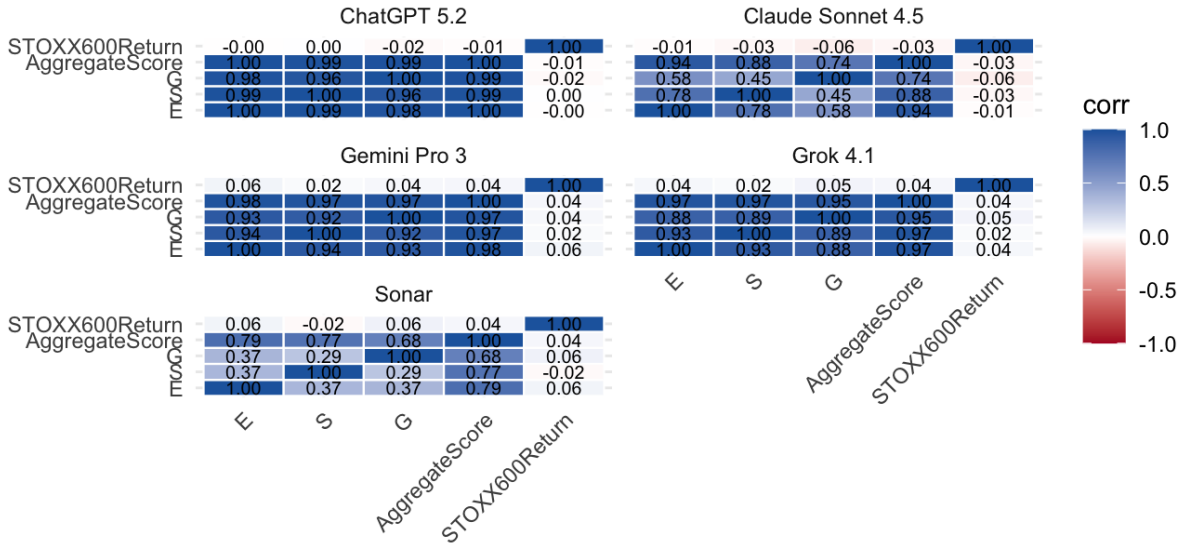
Appendix A. 11 - Correlation Matrices: Italy – All Providers (STOXX600Return) (author's own construction)



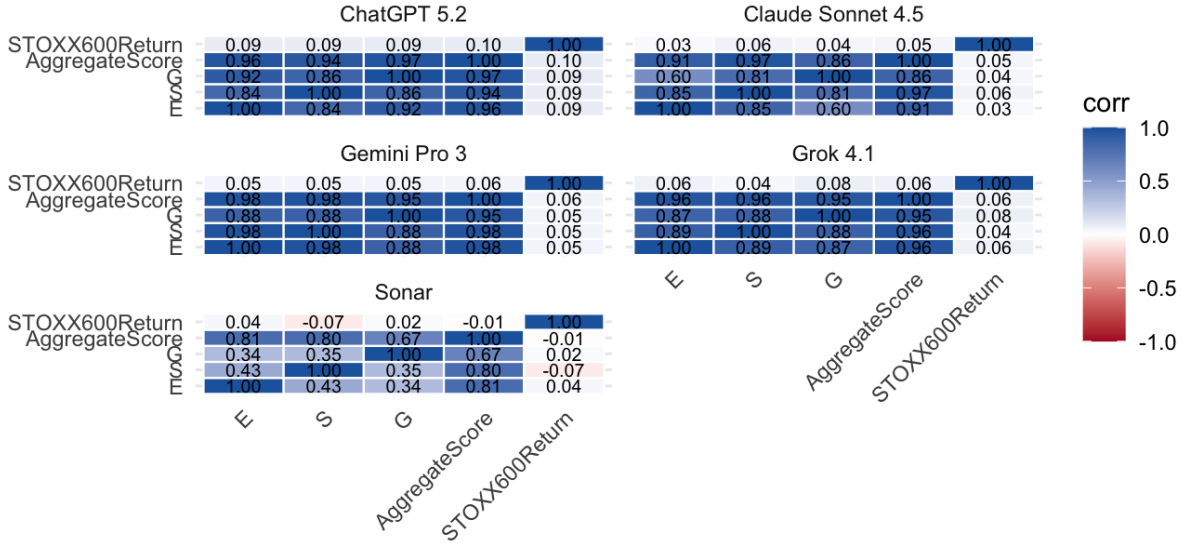
Appendix A. 12 - Correlation Matrices: Norway – All Providers (STOXX600Return) (author's own construction)



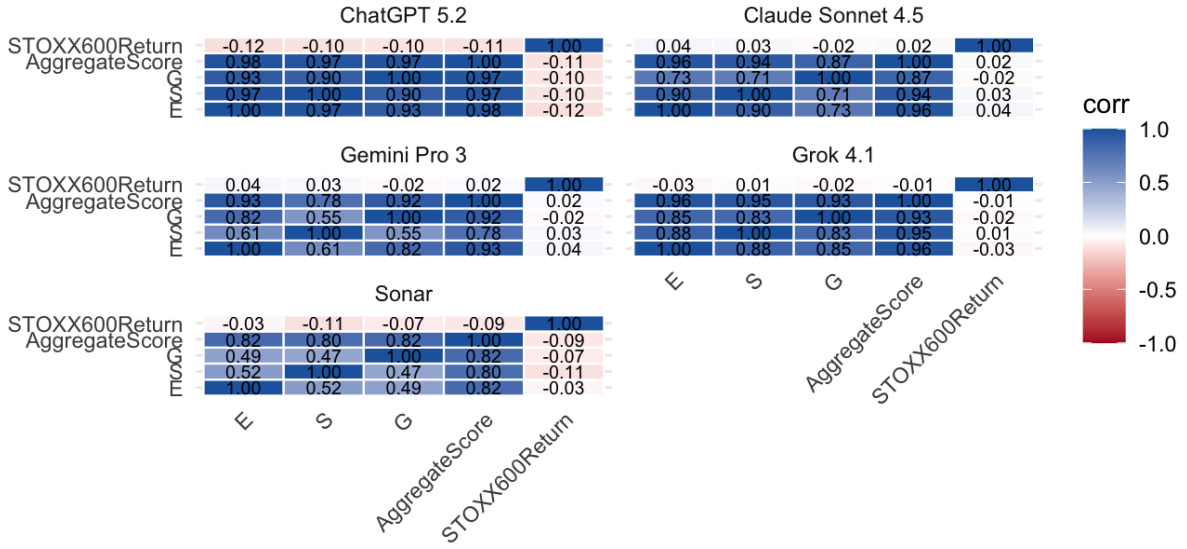
Appendix A. 13 - Correlation Matrices: Poland – All Providers (STOXX600Return) (author's own construction)



Appendix A. 14 - Correlation Matrices: Portugal – All Providers (STOXX600Return) (author's own construction)



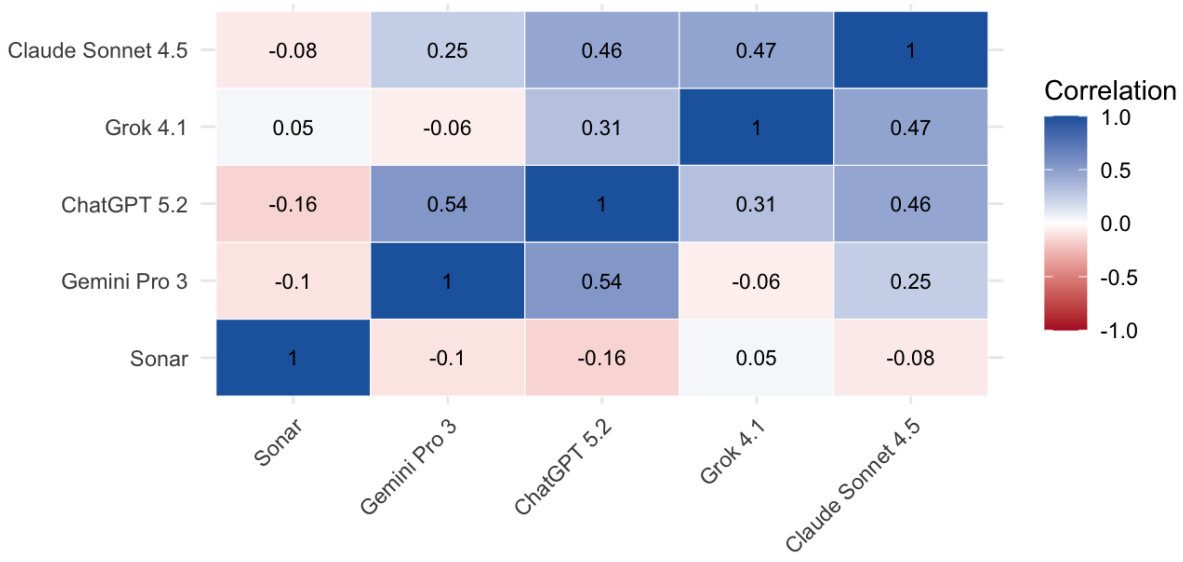
Appendix A. 15 - Correlation Matrices: Spain – All Providers (STOXX600Return) (author's own construction)



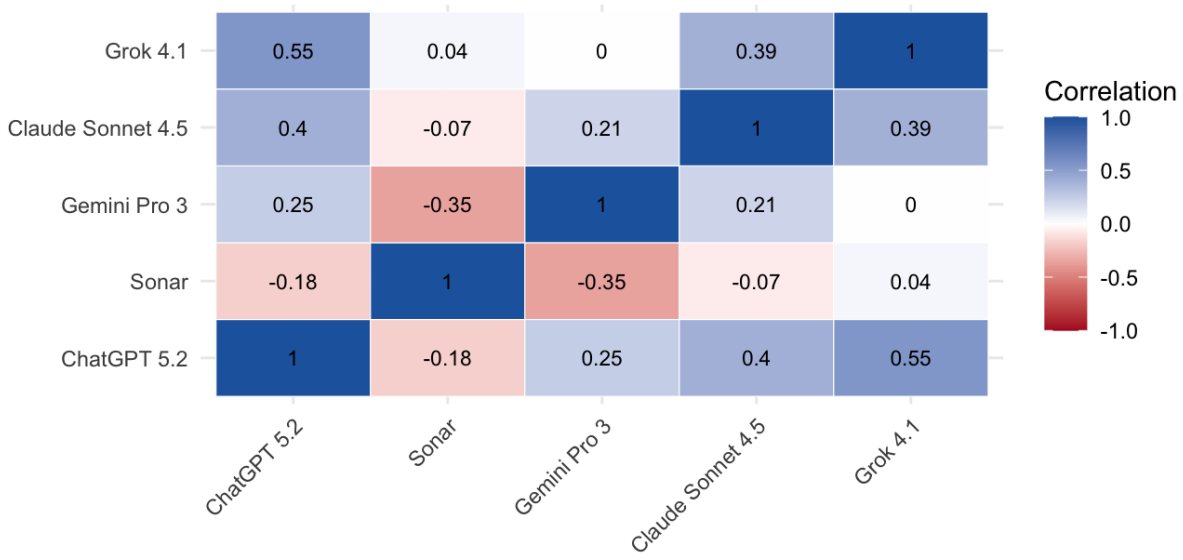
Appendix A. 16 - Correlation Matrices: United Kingdom – All Providers (STOXX600Return) (author's own construction)

APPENDIX B – Correlation Matrices Across Providers

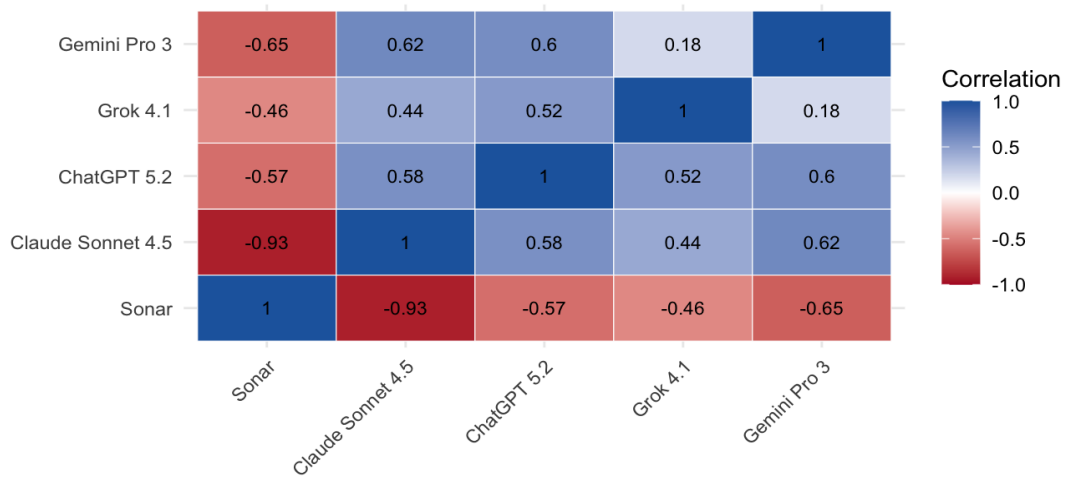
This appendix presents the complete correlation matrices between the Synthetic ESG *AggregateScore* generated by the different providers for the selected countries case study.



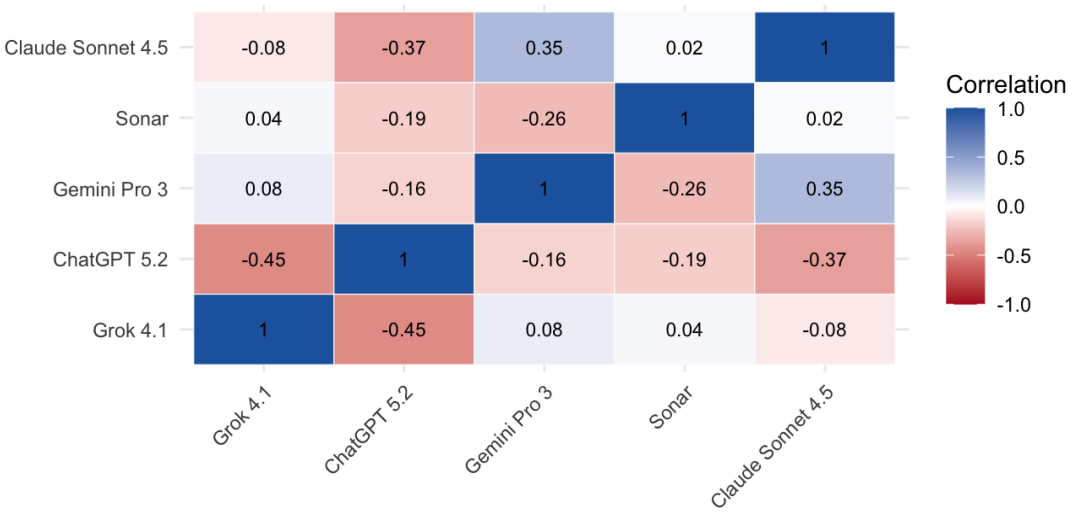
Appendix B. 1 - Correlation of ESG Aggregate Scores Across Providers - France (author's own construction)



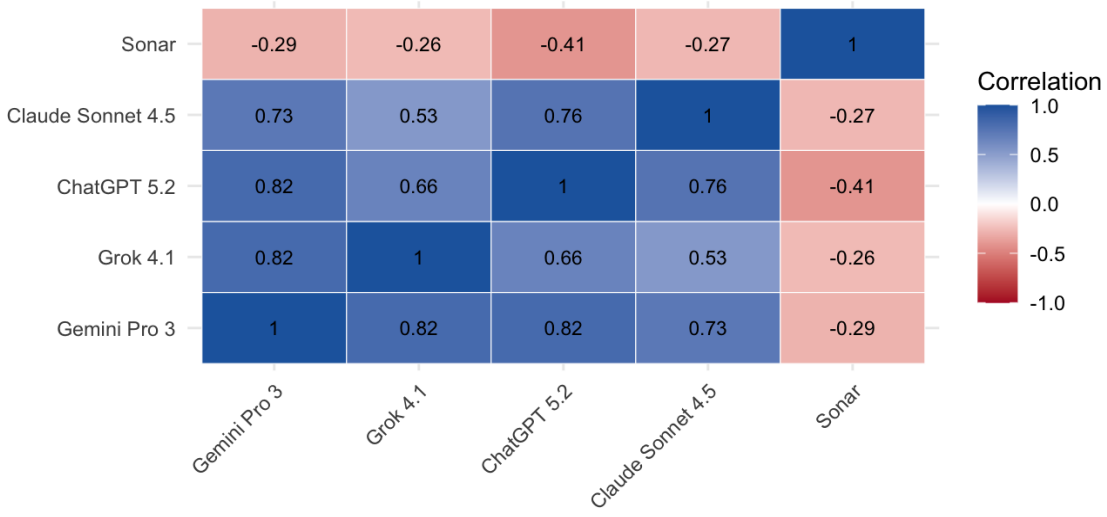
Appendix B. 2 - Correlation of ESG Aggregate Scores Across Providers - Germany (author's own construction)



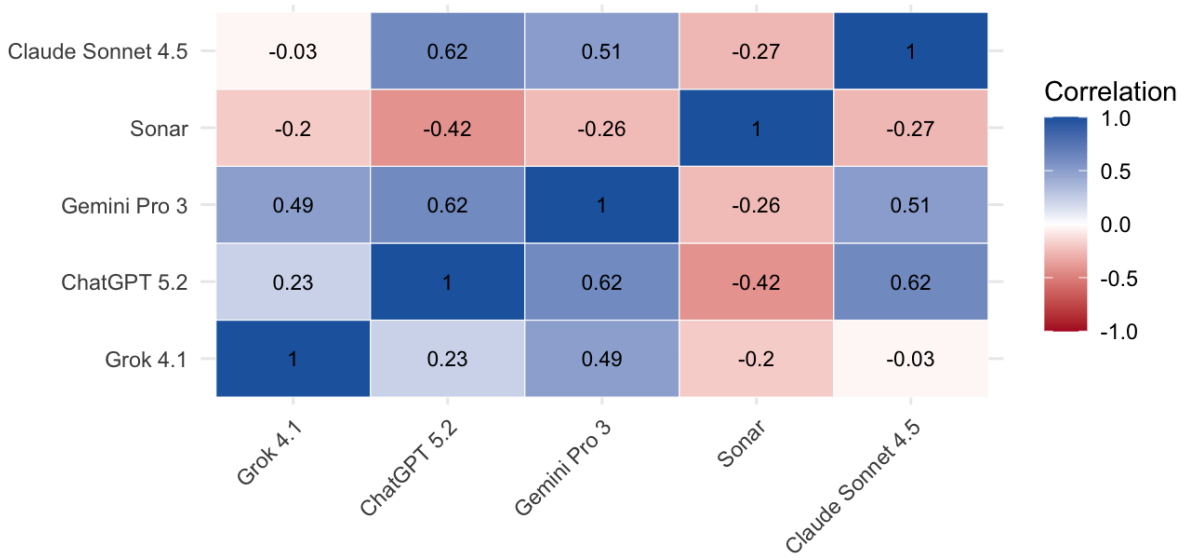
Appendix B. 3 - Correlation of ESG Aggregate Scores Across Providers – Italy (author's own construction)



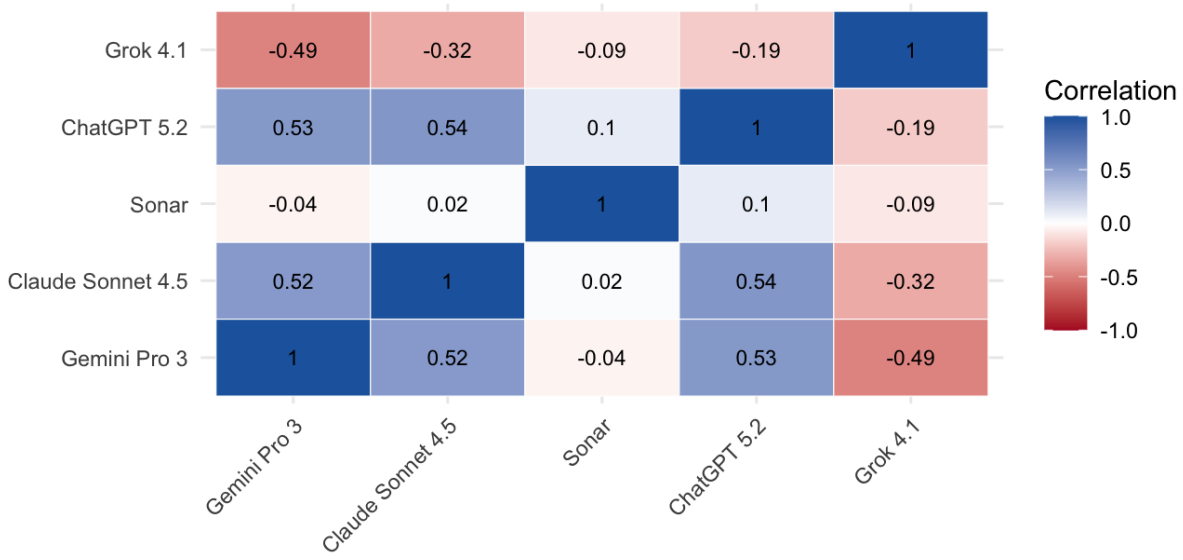
Appendix B. 4 - Correlation of ESG Aggregate Scores Across Providers – Norway (author's own construction)



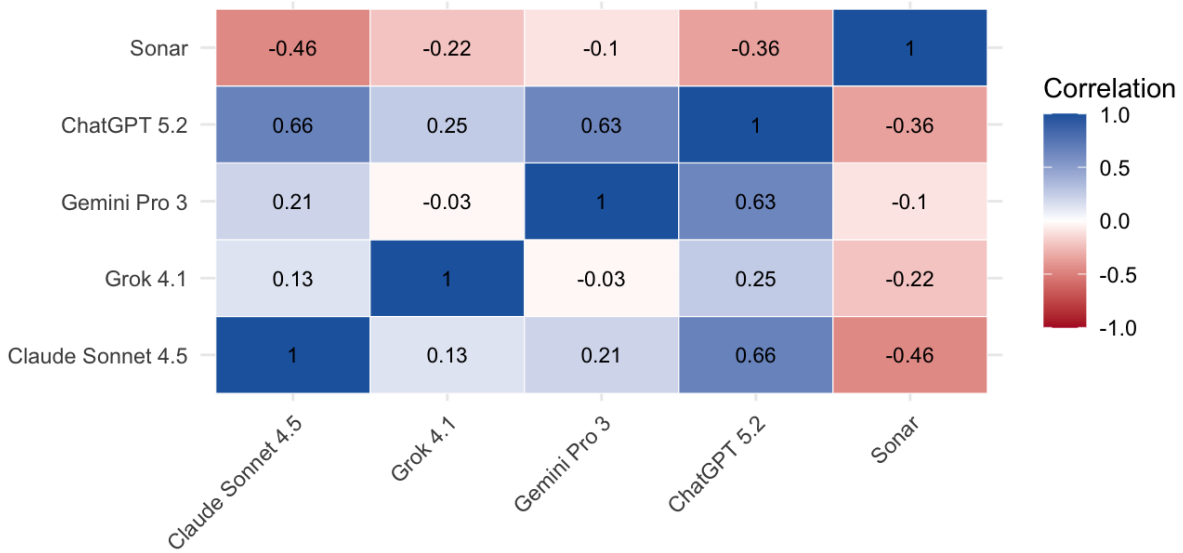
Appendix B. 5 - Correlation of ESG Aggregate Scores Across Providers – Portugal (author's own construction)



Appendix B. 6 - Correlation of ESG Aggregate Scores Across Providers – Spain (author's own construction)



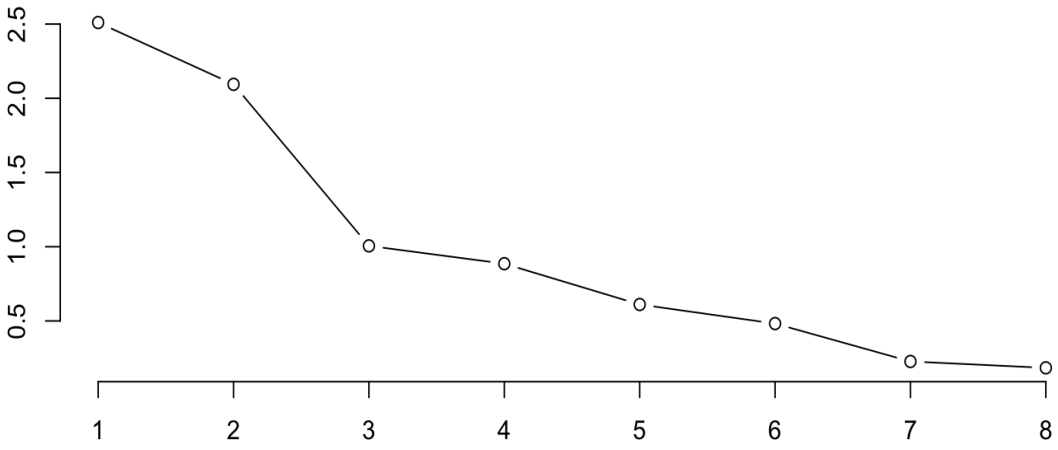
Appendix B. 7 - Correlation of ESG Aggregate Scores Across Providers - United Kingdom (author's own construction)



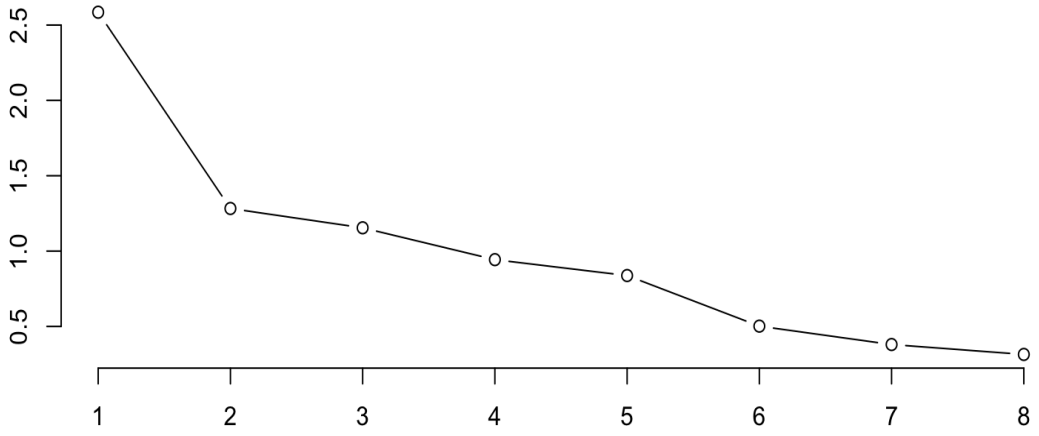
Appendix B. 8 - Correlation of ESG Aggregate Scores Across Providers - Poland (author's own construction)

APPENDIX C – PCA Results

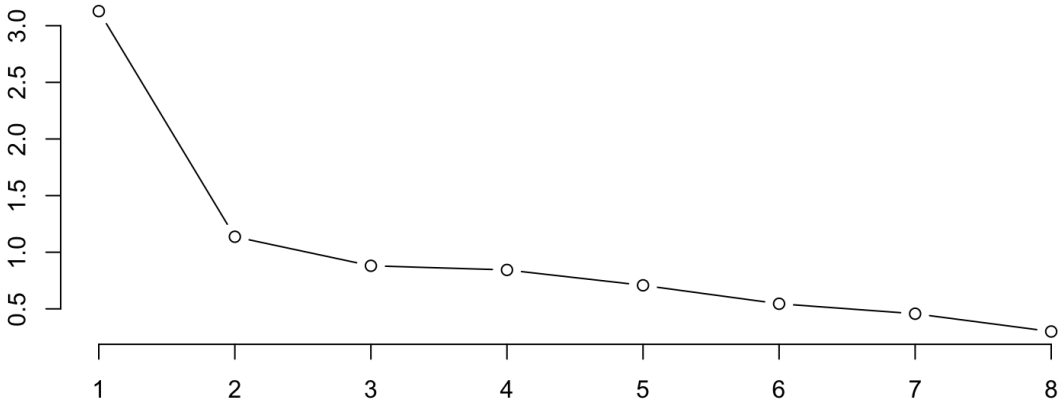
For each provider, the full scree plots of the principal components are reported in this section, illustrating how the proportion of variance declines across successive components and visually confirming the contrast between models with a dominant first principal component (PC1) and those with a more gradual decay.



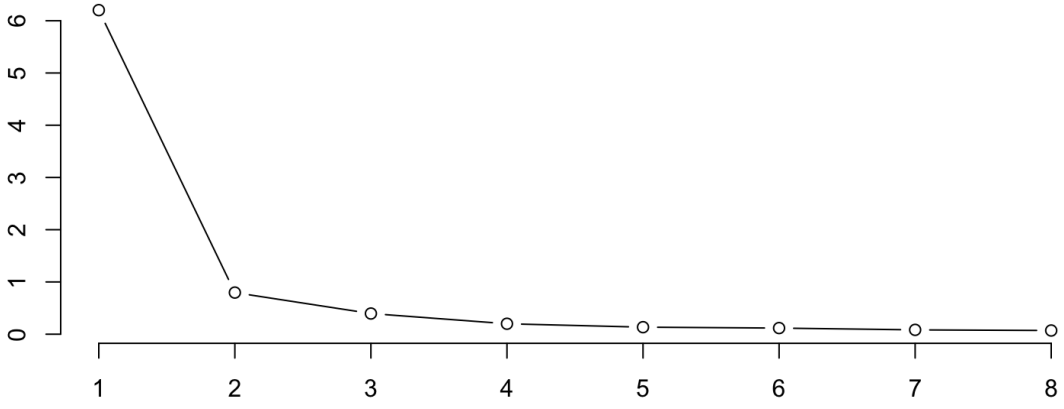
Appendix C. 1 - Scree Plot Claude Sonnet 4.5 – PCA Aggregate ESG (author's own construction)



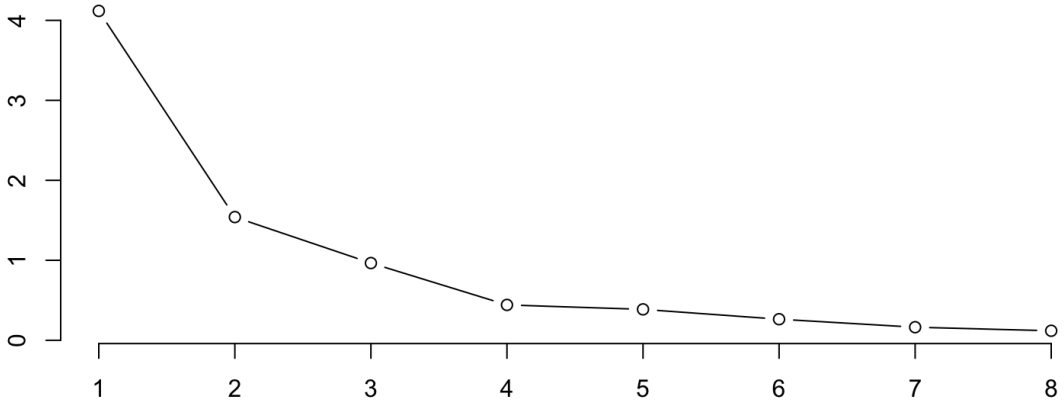
Appendix C. 2 - Scree Plot Gemini 3 Pro - PCA Aggregate ESG (author's own construction)



Appendix C. 3 - Scree Plot Sonar - PCA Aggregate ESG (author's own construction)

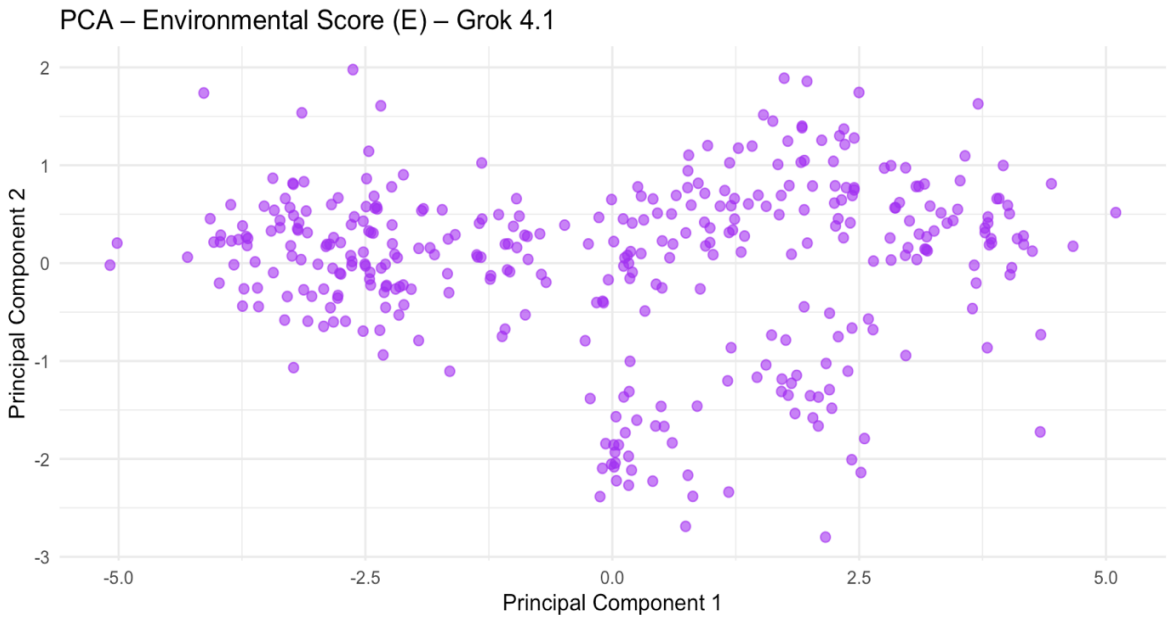


Appendix C. 4 - Scree Plot Grok 4.1 - PCA Aggregate ESG (author's own construction)

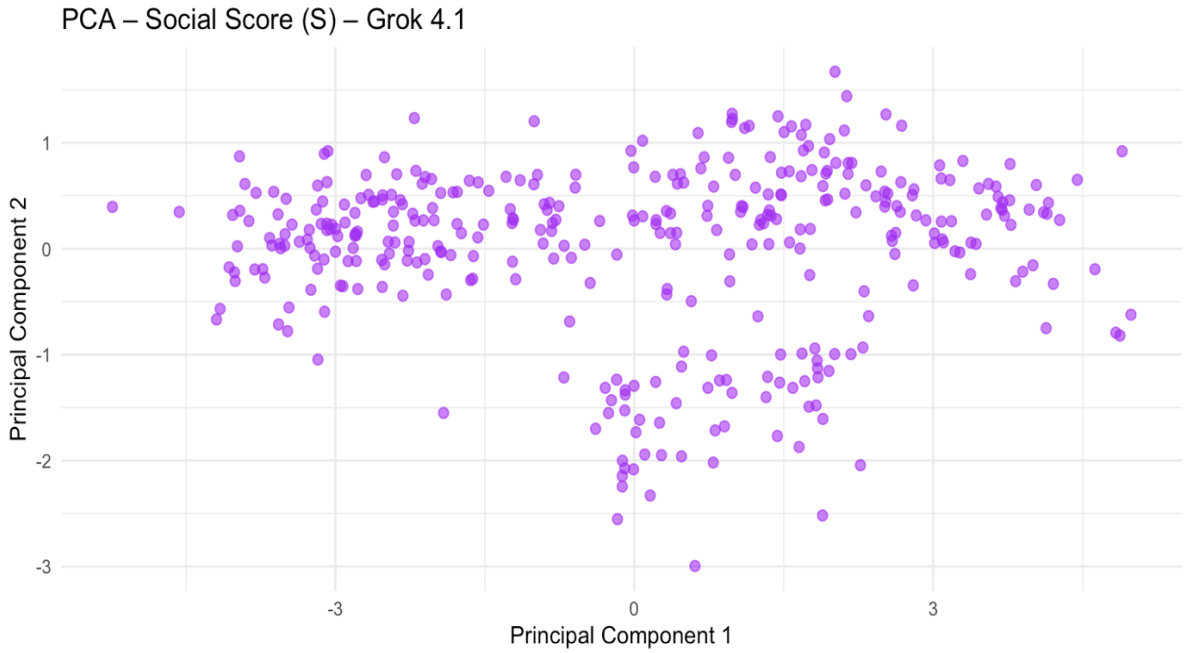


Appendix C. 5 - Scree Plot GPT-5.2 - PCA Aggregate ESG (author's own construction)

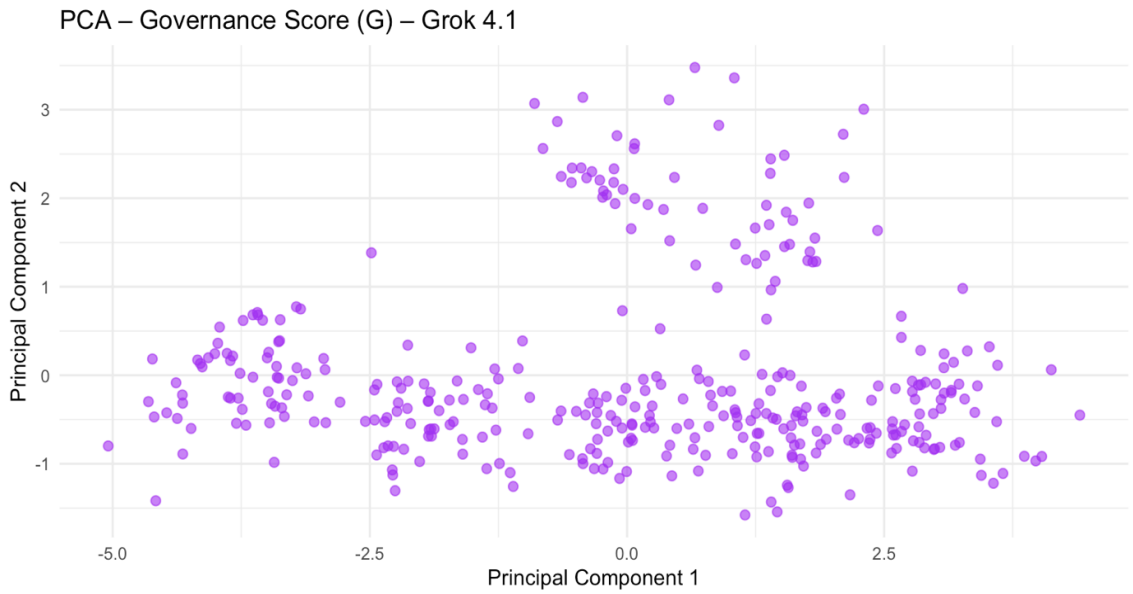
A two-dimensional PCA map of three-dimension E, S, G for Grok 4.1 is also provided showing a relatively compact and structured configuration of country points that is consistent with the strong dominance of the first component documented above.



Appendix C. 6 - PCA – Environmental Score (E) – Grok 4.1 (author's own construction)



Appendix C. 7 - PCA – Social Score (S) – Grok 4.1 (author's own construction)



Appendix C. 8 - PCA – Governance Score (G) – Grok 4.1 (author's own construction)

REFERENCES

Afolabi, H., Ram, R., & Rimmel, G. (2022). Harmonization of sustainability reporting regulation: Analysis of a contested arena. Sustainability, 14(9), 5517.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. Quarterly Journal of Economics, 84(3), 488–500.

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv.

Bahadır, O., & Akarsu, S. (2024). Does company information environment affect ESG–financial performance relationship? Evidence from European markets. Sustainability, 16(7), 2701.

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv.

Baker, M., & Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. Journal of Finance, 61(4), 1645–1680.

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. Review of Finance.

Birti, M., Maurino, A., & Osborne, F. (2025). Optimizing large language models for ESG activity detection in financial texts. In Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF '25). ACM. https://doi.org/10.1145/3768292.3770371

Boffo, R., & Patalano, R. (2020). ESG investing: Practices, progress and challenges. OECD Publishing.

Brown, T. B., et al. (2020). Language models are few-shot learners. arXiv.

CFA Institute. (2015). Environmental, social, and governance issues in investing: A guide for investment professionals.

CFA Institute. (2020). Sustainability in investment management at a pivotal juncture.

CFA Institute. (2023). Guidance for integrating ESG information.

Derrick, K. (2024). ESG sentiment analysis: Comparing human and language model performance including GPT. arXiv.

Drempetic, S., Klein, C., & Zwergel, B. (2019). The Influence of Firm Size on the ESG Score: Corporate Sustainability Ratings Under Review. Journal of Business Ethics, 167(2), 333–360. https://doi.org/10.1007/s10551-019-04164-1

Escobar-Saldívar, L. J., Villarreal-Samaniego, D., & Santillán-Salgado, R. J. (2025). The effects of ESG scores and ESG momentum on stock returns and volatility. Journal of Risk and Financial Management, 18(7), 367.

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

Gavrilakis, N., & Floros, C. (2023). ESG performance, herding behavior and stock market returns: Evidence from Europe. Operational Research, 23(1).

Gibson, R., Krueger, P., & Schmidt, P. S. (2021). ESG rating disagreement and stock returns. Review of Finance, 25(5), 1315–1356.

Goldstein, I., Spatt, C. S., & Ye, M. (2021). Big data in finance. Review of Financial Studies, 34(7), 3213–3225.

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. Contemporary Accounting Research, 40(2), 806–841.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of ICWSM.

Kim, J. W., & Park, C. K. (2023). ESG performance and asymmetry. Applied Economics, 55(26), 2993–3007.

Kojima, T., et al. (2022). Large language models are zero-shot reasoners. arXiv.

Kotsantonis, S., & Serafeim, G. (2019). Four things no one will tell you about ESG data. Journal of Applied Corporate Finance, 31(2), 50–58.

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv.

Lu, Y., et al. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. arXiv.

Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. arXiv.

Reid, M., et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv.

Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., & Leippold, M. (2024). Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. Finance Research Letters, 61, 104979.

STOXX Ltd. (2025). STOXX® Index Methodology Guide (Portfolio Based Indices). STOXX.

Sultana, N. (2026). Global ESG sentiment, policy commitments, and sustainability uncertainty. Sustainable Futures, 11, 101636.

Sun, Y., et al. (2024). Alternative data in finance and business: Emerging applications and theory analysis. Financial Innovation, 10(1), 127.

Vaswani, A., et al. (2017). Attention is all you need. arXiv.

Wei, J., et al. (2022a). Chain-of-thought prompting elicits reasoning in large language models. arXiv.

Wei, J., et al. (2022b). Emergent abilities of large language models. arXiv.

Yu, H., Liang, C., Liu, Z., & Wang, H. (2023). News-based ESG sentiment and stock price crash risk. International Review of Financial Analysis, 88, 102646.

Zhao, W. X., et al. (2023). A survey of large language models. arXiv.

J. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning with Applications in R and Python, 2nd ed., Springer, 2023. ISLRv2_corrected_June_2023.pdf

ONLINE SOURCES

Anthropic. (2025). Introducing Claude Sonnet 4.5. Retrieved from https://www.anthropic.com/news/claude-sonnet-4-5

EcoVadis. (2025). ESG reporting – Key frameworks, compliance and best practices. Retrieved from https://ecovadis.com

ESG360. (2024). ESG rating agency: la sfida di misurare la sostenibilità. Retrieved from https://www.esg360.it

Google Cloud. (2025). Gemini 3 Pro. Retrieved from https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-pro?hl=it

IBM. (2023). What is big data? Retrieved from https://www.ibm.com/topics/big-data

IBM. (2023). What is environmental, social, and governance (ESG)? Retrieved from https://www.ibm.com

London Stock Exchange Group (2024). LSEG ESG Scores Methodology. Retrieved from https://www.lseg.com/content/dam/data-analytics/en_us/documents/methodology/lseg-esg-scores-methodology.pdf

MindStudio. (n.d.). Sonar — AI Model. Retrieved from https://www.mindstudio.ai/models/sonar/

MSCI (2024). MSCI ESG Ratings Methodology. Retrieved from https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Ratings+Methodology.pdf

OpenAI. (2025). Introducing GPT-5.2. Retrieved from https://openai.com/index/introducing-gpt-5-2/

Principles for Responsible Investment. (n.d.). What are the Principles for Responsible Investment? Retrieved from https://www.unpri.org

Statista. (2025). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2029. Retrieved from https://www.statista.com/statistics/871513/worldwide-data-created/

Sustainalytics (2024). ESG Risk Ratings Methodology, Version 3.1. Retrieved from https://www.sustainalytics.com/docs/knowledgehublibraries/default-document-library/sustainalytics_-esg-risk-ratings_-version-3-1_-methodology-abstract_-june-2024.pdf

xAI. (2025). Grok 4.1. Retrieved from https://x.com/xai/status/1990530499752980638

STOXX Ltd. (2025). STOXX® Index Methodology Guide (Portfolio Based Indices). STOXX. https://www.stoxx.com