NBER WORKING PAPER SERIES

Optimizing ESG rating assessment in Energy Company via Supervised Fine-Tuned LLMs

Mengyao Zhang[1], Aref Mahdavi Ardekani[1], Mark Cummins[2], and Michael Dowling[1]
[1]Dublin City University
[2]University of Strathclyde

Working Paper
September 2025

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2025

© 2025 National Bureau of Economic Research, Inc.

Abstract

This study explores supervised fine-tuning of large language models to enhance Environmental, Social and Governance (ESG) evaluation for listed energy companies. ESG ratings are increasingly central to sustainable investing, yet existing systems often lack transparency, consistency and timeliness. To address these issues, we build a dataset that links sustainability reports to two major rating sets, Refinitiv and Bloomberg, which are used as supervision targets. Five representative models are fine-tuned, namely ChatGPT-4o-mini, LLaMA, Qwen and two DeepSeek distilled variants. Performance is benchmark dependent: ChatGPT-4o-mini aligns most closely with Bloomberg ratings, while Qwen performs best on Refinitiv, with lower mean squared error and mean deviation and higher accuracy and correlation than pre-trained baselines. The fine-tuned models also produce concise, text-grounded rationales that improve interpretability. Together, the results indicate that benchmark-specific fine-tuning can deliver transparent, timely and cost-efficient ESG assessments and reduce reliance on proprietary feeds, supporting investors, regulators and issuers in informed decision-making.

Keywords: ESG, energy, large language models, fine-tuning.

Optimizing ESG rating assessment in Energy Company via Supervised Fine-Tuned LLMs
Mengyao Zhang

1 Introduction

Global challenges such as climate change, resource scarcity, and social inequality have pushed sustainable investing to the forefront of the financial industry. Environmental, Social, and Governance (ESG) is an evaluation system that takes into account environmental and social welfare while considering profitability (Dechow, 2023) and has been widely used by market participants in investment analysis and portfolio selection (Christensen et al., 2022). One sector where ESG considerations are especially salient is the energy industry, as it lies at the heart of climate solutions. In particular, renewable energy has emerged as a focal point, it is viewed as an effective solution to combat climate change and alleviate the energy crisis, and its enormous growth potential has started to attract significant investor interest in the ESG era (Mohandas, 2024). In this context, our study investigates how artificial intelligence models can be fine-tuned to improve the transparency, timeliness and interpretability of ESG evaluation in energy companies.

With the rise of sustainable finance, ESG ratings have become a primary tool for quantitatively assessing a company's commitment to sustainability. It saves substantial effort in conducting in-depth analyses of specific data when assembling a sustainable investment portfolio (Billio et al., 2021).These ratings condense complex ESG performance into a single score, offering investors a convenient benchmark. In practice, ESG scores are now widely used in portfolio selection and risk management, a recent survey found that 65% of investors consult ESG ratings at least once a week. The appeal of these ratings is their simplicity: they provide an official channel to incorporate

sustainability factors without requiring investors to perform extensive analysis themselves (Liu et al., 2025). According to Ni (2023), ESG investing can be understood as the incorporation of non-financial ESG information into the investment decision-making process by financial institutions. This approach not only encourages behavioral changes among investors and corporations but also mitigates adverse societal externalities while bolstering the sustainability of long-term investments. ESG ratings play a crucial role in guiding capital toward more sustainable companies. Better understanding and improving these ratings is not just an academic exercise, but a pressing issue for financial markets and society at large.

However, the current ESG rating ecosystem suffers from several inherent problems that undermine its usefulness and credibility. First, most ESG rating methodologies are proprietary and opaque. Data providers like Refinitiv, Bloomberg, MSCI, and S&P Global each use their own criteria and weighting schemes, making their scores something of a "black box" to outsiders (Del Vitto, Marazzina and Stocco, 2023). This lack of transparency makes it difficult for investors to trust or interpret the drivers behind a particular score. Second, there is often significant disagreement among rating agencies on what score a given company deserves. Different providers can arrive at markedly different ESG assessments for the same firm. Such inconsistencies are not merely theoretical, they can have real financial consequences. Recent research in the Chinese market found that when ESG ratings diverge substantially, idiosyncratic stock volatility increases, presumably due to investor confusion and noise trading(Liu et al., 2025). In other words, the lack of a unified framework for ESG scoring not only weakens confidence in the ratings' reliability, but may also introduce instability in asset prices.

A third limitation of traditional ESG ratings is their lack of timeliness and scope. ESG data is typically updated on an annual cycle aligned with corporate reporting, meaning scores may remain static for long periods. This makes it hard for investors to use ratings to respond to emerging news, controversies, or significant events in real time. Moreover, the conventional process relies heavily on structured disclosures and quantitative metrics, struggling to incorporate rich unstructured information such as long-form sustainability reports or news articles. Important signals embedded in textual data, like a new climate commitment in a CEO's letter or a controversy reported in the media may be overlooked or only reflected with a delay. These gaps constrain the interpretability and responsiveness of ESG evaluations, highlighting the need for more dynamic and transparent approaches. While ESG ratings are vital for channeling investment toward sustainability, their current formulation is ripe for improvement to greater consistency, clarity, and real-time adaptability would markedly enhance their value.

Advances in artificial intelligence, particularly large language models (LLMs), open up a promising avenue to address these challenges. In the past few years, LLMs like OpenAI's GPT series and Meta's LLaMA have demonstrated exceptional capability in understanding and generating human-like text. These models built on the Transformer architecture, can digest vast amounts of unstructured text from news articles to corporate reports, and produce coherent analyses or summaries. Such capabilities suggest that LLMs could directly analyze the textual data underlying ESG issues (e.g. sustainability reports, press releases, news) and infer a company's ESG performance in a more granular and up-to-date manner than traditional methods.

In theory, a language model finely tuned to the ESG domain could read through a firm's annual report, recognize relevant ESG disclosures or controversies, and output an evaluative score or narrative. This text-driven approach would leverage far more information than the limited set of indicators in standard ESG rating models, potentially yielding a more comprehensive assessment. Importantly, it could also increase transparency, an LLM could be prompted to explain why it assigned a certain score by pointing to specific evidence in the text, thereby providing a form of interpretable "audit trail" for its ratings.

However, applying general-purpose LLMs directly to specialized financial tasks comes with its own hurdles. Large language models are prone to hallucination. they might produce confident-sounding statements that are not grounded in facts, especially in domains where the training data was sparse or where precision is critical (Mündler et al., 2023). Finance and ESG content contain technical terminology and nuanced context that a generic model might not fully grasp, leading to irrelevant or incorrect outputs. Prior research has shown that even advanced models like GPT-3.5/4 can generate incorrect or nonsensical answers to domain-specific questions if not properly adapted. Moreover, as Radford et al. (2019) observed, a shortage of high-quality financial data in the training mix can hurt the accuracy and reliability of model-generated content. Therefore, to harness LLMs for ESG evaluation, it is crucial to specialize or fine-tune these models on relevant data, and to guide their reasoning process so as to minimize errors. The goal is to combine the raw power of LLMs (their ability to read and synthesize vast text) with domain-specific knowledge and constraints that curb their tendency to go off-track (Ouyang et al., 2022).

In this study, we propose an AI-driven framework that fine-tunes LLMs to perform ESG scoring for energy companies, addressing the deficiencies of traditional ESG ratings. We construct a large training corpus that integrates standard ESG datasets with real-time textual information.

In particular, we leverage two well-established ESG rating benchmarks, Refinitiv ESG scores and Bloomberg's ESG Disclosure scores as supervised targets for our models. These provide official ground truth signals on a company's sustainability performance. By combining the structured ESG labels with rich textual descriptions of companies' activities, we create a dataset suitable for teaching a language model how to read ESG-related report and predict an ESG rating from it.

We fine-tune five representative LLMs, ChatGPT-4o-mini, Meta's LLaMA, Alibaba's Qwen, and two smaller distilled versions of LLaMA and Qwen developed by DeepSeek. During fine-tuning, the models learn to map textual cues to ESG score outputs, essentially internalizing the rating logic of the Refinitiv and Bloomberg benchmarks. To enhance the models' real-world utility, we incorporate Chain-of-Thought (CoT) reasoning techniques in our framework. CoT prompting encourages the model to explain its reasoning step by step, which not only aids interpretability but also tends to reduce logical errors in the final output. Through this multi-faceted approach, our aim is to develop a system that can extract high-quality ESG insights from text and generate scores that are accurate, transparent, and timely.

We rigorously evaluate the fine-tuned models to see how well they can replicate and interpret official ESG ratings. The test bed for our experiments is a set of publicly listed energy companies. After training, we task each model with reading a company's latest ESG report and outputting ESG scores, which we compare against the official ratings.

Our results are encouraging. Performance is benchmark conditioned. On the Bloomberg dataset, the fine-tuned ChatGPT-4o-mini model achieves the best alignment, with high correlation and minimal average error. On the Refinitiv dataset, the fine-tuned Qwen model attains the top performance on the same metrics. This suggests that the leading model on each benchmark has effectively learned the underlying ESG evaluation logic from the training data and can apply it with high accuracy to new company reports. In addition, the ChatGPT-based model demonstrates a strong ability to explain its scoring decisions, often highlighting the same key factors that human analysts consider, for example a major renewable energy project launch or a governance scandal, when justifying its score. The LLaMA model shows moderate success, with outputs that trend in the right direction and improve over baseline, but with higher error rates and occasional anomalies. The smaller distilled versions struggle the most, likely due to limited capacity. We observe some hallucination in the weaker models, such as citing an ESG policy not mentioned in the text, which underscores the importance of strong base models and techniques and CoT to maintain factual accuracy.

Overall, these findings illustrate the potential of fine-tuned LLMs to revolutionize ESG evaluation. To our knowledge, this is the first study to successfully align a general large language model with established ESG rating criteria and demonstrate its applicability in a specific domain. By overcoming the opacity of traditional ESG scores, our approach offers a path toward more transparent and interpretable ESG assessments. Stakeholders could not only receive an updated ESG score in near-real-time as new information emerges, but also a summary of why that score changed, drawn directly from the latest disclosures and news. This real-time, explainable insight is particularly valuable in an era of heightened scrutiny on greenwashing, companies making misleading claims about sustainability.

An AI-driven system that continuously reads and evaluates what companies are actually doing about ESG can serve as an early-warning mechanism for inconsistencies between words and actions. Furthermore, our framework is inherently scalable, once trained, the LLM can be deployed across thousands of companies, potentially filling coverage gaps where human analysts or rating agencies have limited resources. In short, if this approach works as envisioned, it could greatly enhance investors' ability to make informed, responsible decisions and help allocate capital in alignment with global sustainability goals. It marries the breadth and depth of qualitative ESG analysis with the consistency of quantitative scoring, thereby bridging a critical gap in sustainable finance research. We believe this contribution not only advances the academic literature at the intersection of NLP and finance, but also has practical implications for how ESG information is consumed and acted upon.

The structure of this paper is as follows: Section 2 reviews the related literature; Section 3 details the fine-tuning experimental setup underpinning our methodology; in Section 4, the experimental results are discussed, and a critical analysis is offered; finally, in Section 5, the study's conclusions are presented, and future work directions are discussed.

2 Related Work

ESG indicators have attracted sustained academic and practitioner interest, with many studies linking non-financial performance to financial outcomes and investment decisions (Margot et al., 2021; Christensen et al., 2022). Ratings issued by Refinitiv, Bloomberg, MSCI and S&P Global are widely used because they compress complex information into accessible scores for portfolio construction and risk management (Billio et al., 2021). Yet two structural issues limit their usefulness. First, most providers rely on proprietary criteria and weightings, which makes the rating process opaque to users and hard to audit (Del Vitto, Marazzina and Stocco, 2023). Second, agencies often disagree on the score a firm deserves, due to differences in scope, measurement and aggregation (Berg et al., 2022a; Berg et al., 2022b; Billio et al., 2021). This divergence is not only a conceptual problem; evidence shows that greater rating disagreement is associated with higher idiosyncratic return volatility, consistent with noise trading induced by mixed signals (Liu et al., 2025). These frictions motivate methods that retain the convenience of a single score while making the mapping from evidence to rating transparent and auditable.

Early work modeled ESG or related financial outcomes with linear models, tree ensembles and neural networks, frequently achieving solid predictive accuracy on structured inputs (Sharma, Gupta and Gupta, 2022). Other studies examined market reactions to rating changes and the informational content of vendor scores (Pelizzon et al., 2021; Berg et al., 2022a). However, such approaches often under-utilise long-form textual disclosures and usually provide limited insight into why a specific score is assigned. Some recent papers begin to close this gap by extracting structured ESG signals from reports with information-extraction and LLM pipelines, improving timeliness and coverage while documenting what the model found (Bronzini et al., 2024). There is also growing evidence that fine-tuning domain models on ESG-specific corpora can raise classification and detection accuracy relative to generic baselines (Xia, Yang and Liu, 2024), and that targeted datasets for ESG activities materially improve open models after fine-tuning (Birti, Osborne and Maurino, 2025). These findings motivate language-native systems that read reports directly and return scores with reasons.

Domain pre-trained language models such as FinBERT and FLANG improved performance on finance sentiment and classification tasks by adapting to financial language (Araci, 2019; Yang et al., 2020; Shah, Shah and Ramachandran, 2022). BloombergGPT further showed the benefits of scaling domain pre-training but remains closed, which limits reproducibility and cost control (Wu et al., 2023). Evidence is mixed on whether finance-specific BERT variants consistently outperform strong general baselines once vocabulary and data are controlled (Arslan, Dengel and Ahmed, 2021). General LLMs such as ChatGPT and LLaMA exhibit powerful text understanding but can hallucinate and mis-handle specialised concepts without domain supervision (Bubeck et al., 2023; Jeong, 2023; Jeong, 2024). These observations point to a pragmatic route: supervised fine-tuning of capable base LLMs on ESG texts, optionally with reasoning guidance, to improve factual alignment and interpretability while avoiding the barriers of fully closed domain models.

Instruction-following and supervised fine-tuning enable models to follow task schemas and produce well-formed outputs; human-feedback pipelines and instruction datasets have proven effective for

controlling format and content in downstream tasks (Ouyang et al., 2022; Touvron et al., 2023; Taori et al., 2023; Chiang et al., 2023). For ESG evaluation, this suggests a design in which an LLM learns benchmark-specific rating behaviour and emits both a total score and pillar-level justifications from textual evidence. Recent work has begun to replicate vendor scoring logic with machine learning to illuminate key drivers and the balance between aspirational disclosures and realised performance (Ferro, Marazzina and Stocco, 2025), reinforcing the value of transparent, evidence-linked outputs. In a multi-benchmark context, aligning to Refinitiv and Bloomberg separately is methodologically important because their scales and criteria differ. Normalising evaluation thresholds to each scale facilitates fair comparison and helps explain why some benchmarks appear more learnable than others in practice. Retrieval or constrained prompting can further anchor outputs in cited text to reduce factual drift and enhance auditability (Bronzini et al., 2024).

Smaller or distilled models offer efficiency benefits, yet their capacity limits can reduce accuracy and explanation stability on long, evidence-heavy tasks. Studies fine-tuning compact models for ESG text detection show encouraging gains but still report gaps to larger systems on complex reasoning (Xia, Yang and Liu, 2024; Birti, Osborne and Maurino, 2025). For applications where users require tight error control and consistent rationales, mid-to-large LLMs adapted via supervised fine-tuning and reasoning guidance tend to be more reliable, whereas smaller variants can complement as cost-effective components when bandwidth and latency dominate.

The literature shows that ESG ratings are widely used yet suffer from opacity and cross-vendor disagreement, with measurable market consequences (Billio et al., 2021; Berg et al., 2022a; Liu et al., 2025). Traditional ML approaches predict ESG or financial outcomes but under-exploit text and often lack interpretable, evidence-linked justifications. Finance-domain PLMs improve on targeted tasks yet face reproducibility and scope constraints, while general LLMs require adaptation to avoid hallucination and domain errors. Recent advances in LLM-based information extraction and ESG-focused fine-tuning demonstrate that text-grounded, explainable ESG analytics are feasible (Bronzini et al., 2024; Xia, Yang and Liu, 2024; Birti, Osborne and Maurino, 2025). What remains scarce is a dual-benchmark, language-native framework that learns the distinct logics of Refinitiv and Bloomberg from public disclosures, outputs structured scores with concise rationales, and evaluates alignment under scale-appropriate metrics. Our study addresses this gap by fine-tuning strong base models on ESG texts from energy firms, training separately against Refinitiv and Bloomberg targets, and analysing accuracy, deviation and correlation alongside explanation quality.

3 Methodology

3.1 Data Collection and Processing

3.1.1 Dataset Construction

The sample selected for this study comprises publicly listed energy companies from Asia, Europe, North America, and Australia. Following a process of screening and comparison, a total of 332 companies were identified. These companies possess ESG scores from both Refinitiv and Bloomberg,

two recognised rating agencies. To enhance the accuracy of fine-tuned models, reduce hallucination, and address computational constraints, we collected the ESG reports and financial reports of these companies for the year 2024. Based on this information, and their respective 2024 ESG scores from Refinitiv and Bloomberg, two datasets were constructed, named Ref_company and Blg_company, respectively. These datasets were used independently for fine-tuning in order to evaluate model performance when subjected to differing ESG rating methodologies.

We generated a corpus-level word cloud from all ESG reports in our dataset after lowercasing, lemmatizing and removing stopwords and boilerplate. The visualization highlights the most frequent disclosure terms and themes, with prominent tokens such as management, company, energy, employees, emissions, risk, safety, water, board, operations, supply, sustainability and financial. Taken together, these keywords summarise the dominant topics across E, S and G pillars and provide a quick validity check on coverage; we also use them to refine our KPI lexicon and prompt templates for downstream fine-tuning.

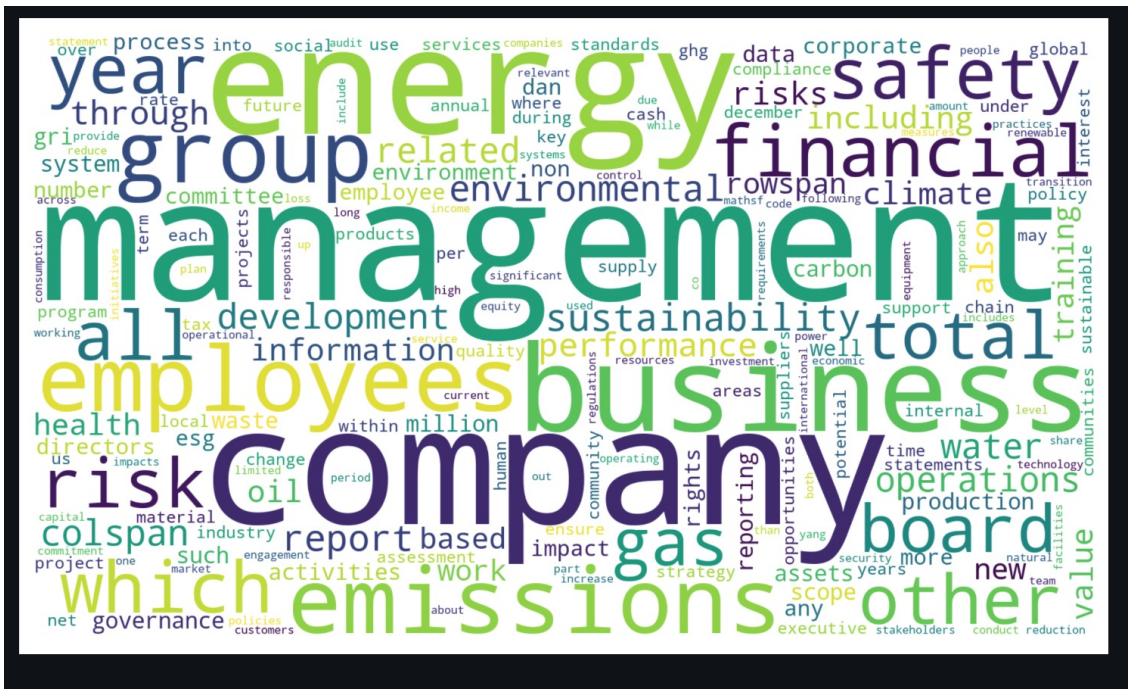


Figure 1: Word cloud of ESG reports highlighting frequent disclosure terms and themes.

Among the main ESG rating agencies currently providing official ESG scores are Refinitiv, Bloomberg, MSCI, and Sustainalytics. These agencies apply different evaluation methodologies, including distinct computational logics, weighting schemes, and data sources. This study focuses on ESG scores from Refinitiv and Bloomberg. Refinitiv was acquired by the London Stock Exchange Group (LSEG) in 2021; however, as the database still retains the name Refinitiv, this name is used throughout the study (London Stock Exchange Group, 2021).

ESG scores from Refinitiv are designed to transparently and objectively measure a company's relative ESG performance, commitment and effectiveness across 10 main themes (emissions, environmental

product innovation, human rights, shareholders, etc.) based on publicly-reported data. As one of the earliest providers to systematically compile and evaluate environmental and social data, Refinitiv offers coverage of more than 12,000 publicly listed and privately held companies, representing over 80% of global market capitalization (Refinitiv, 2023)
To generate its ESG ratings, Bloomberg uses real-time financial data, helping the firm to create dynamic and market-responsive insights that traditional survey-based providers struggle to replicate. The company's strength lies in its ability to combine alternative data sources, including satellite imagery and social media sentiment analysis, with conventional ESG metrics to provide unique insights unavailable elsewhere.

3.1.2 Data Preprocessing

ESG and financial reports typically contain large volumes of heterogeneous information, including structured tables, boilerplate text, and unstructured narrative content.
As a preliminary step, we implemented a multi-stage preprocessing pipeline to prepare the documents for model consumption. Initially, all reports were converted to the markdown text and non-informative content such as disclaimers and repetitive formatting artifacts was removed. Second, the documents were segmented by semantic topic using rule-based heuristics (e.g., section headers such as "Environmental Performance"), which enabled focused processing of ESG-relevant segments. Third, key performance indicators were extracted and embedded as special tokens to facilitate model recognition of numerical information. In addition, we employed Retrieval-Augmented Generation (RAG) to associate each prompt with the most contextually relevant text segment, thereby improving the quality of supervision. This pipeline yields consistent, information-dense input representations that improve fine-tuning efficiency and alignment with rating benchmarks. Compared to prior studies that rely on raw document ingestion or shallow bag-of-words representations, our preprocessing approach ensures higher semantic density, better model conditioning, and greater alignment with ESG rating logic.

3.1.3 Model Selection and Benchmark Design

In this study, in order to capture a representative cross-section of current LLM capabilities, we selected five models varying in architecture, parameter scale, and release modality: GPT-4o-mini, LLaMA 3.1-8B, Qwen 2.5-7B, and two distilled variants developed by DeepSeek.
These models were selected to reflect a spectrum of capabilities in instruction following, domain generalisation, and computational efficiency. While ChatGPT represents a state of the art proprietary model, LLaMA and Qwen are leading open-source alternatives with growing adoption in academic and industrial research. The inclusion of distilled versions allows us to examine performance trade-offs associated with model compression.
Besides, we adopt Refinitiv and Bloomberg ESG ratings as supervision targets to ground model outputs in institutional-grade scoring frameworks. These benchmarks differ significantly in scoring logic, dimensional emphasis, and data transparency, providing a valuable testbed for robustness analysis. Refinitiv is known for broader company coverage and greater reliance on disclosed

Table 1: Model Characteristics Comparison

GPT-4o-mini (18/07/2024)	GPT-4 is developed by OpenAI. It boasts 1.75 trillion parameters, making it one of the most powerful language models in the world. Based on the Generative Pre-trained Transformer (GPT) architecture, it is particularly well-suited for specialized textual interactions and tasks. GPT-4 can be applied in domains such as NLP, machine translation, and code generation. It has demonstrated the ability to produce text at near-human levels, translate languages accurately, and generate various types of creative content.
LLaMA-3.1-8B (12/01/2025)	LLaMA is developed by Meta, with parameter sizes ranging from 7 to 65 billion. Although it is smaller than GPT-4, it remains highly capable. A key feature of LLaMA is its open-source nature, allowing anyone to use it for research or commercial purposes. This accessibility makes it particularly appealing to researchers and developers who wish to build their own language models.
Qwen-2.5-7B (26/09/2024)	Qwen is developed by Alibaba. Its flagship version, Qwen-14B, comprises approximately 14 billion parameters. Qwen excels at Chinese semantic understanding, multi-turn dialogue, and knowledge reasoning, and it also demonstrates strong capabilities in cross-domain information integration and creative text generation. These advantages stem from its efficient architecture and open interfaces, which further enhance its adaptability and performance.
Deepseek R1-distill-Qwen-7B (24/02/2025) and Deepseek R1-distill-LLaMA-8B (24/02/2025)	These two models are distilled models based on the Qwen and LLaMA architectures, respectively, developed using Deepseek's proprietary large-scale corpus. Through knowledge distillation, both models achieve greater efficiency while retaining strong performance in language understanding and reasoning. Their training on diverse, high-quality data sources enhances their versatility across a wide range of NLP tasks.

KPIs, while Bloomberg incorporates more qualitative judgment and stricter rating scales. This dual-benchmark setup enables us to evaluate whether fine-tuned models internalize rating-specific logic or exhibit generalized ESG reasoning capability.

By fine-tuning each model separately on Refinitiv and Bloomberg ESG ratings, we create two distinct supervision regimes that reflect fundamentally different scoring logics. This setup allows us to examine not only whether models can reproduce benchmark-specific criteria, but also how their outputs vary in response to divergent institutional standards, thereby revealing the extent to which fine-tuned LLMs internalize or adapt to ESG evaluation frameworks.

As one of the earliest providers to systematically compile and evaluate environmental and social data, Refinitiv offers coverage of more than 12,000 publicly listed and privately held companies, representing over 80% of global market capitalization (Refinitiv, 2023). In parallel, the study incorporates Bloomberg's ESG rating data, allowing for comparative evaluation of fine-tuned models across datasets derived from rating providers that differ in both coverage and methodological frameworks.

3.2 Fine-tuning

The development of LLMs represents a significant breakthrough in generative Artificial Intelligence (AI). Models based on the Transformer architecture, such as BERT and ChatGPT, have acquired robust representational capabilities through pre-training on large-scale text datasets, rendering them suitable for various NLP tasks.

Recently, LLMs have demonstrated remarkable proficiency in natural language understanding (NLU) and executing a range of tasks via natural language instructions without additional training data (Brown et al., 2020). In these foundational models, LLMs are employed as Pre-trained Language Models (PLMs) using self-supervised or semi-supervised learning on extensive general knowledge corpora. PLMs are defined as models trained on large-scale datasets while remaining domain-agnostic regarding their subsequent application. In practical applications, PLMs must have their final output layers adapted to the tasks of specific specialised domains, a process commonly referred to as fine-tuning (Raschka, 2023). Fine-tuning aims to adjust the existing knowledge of PLMs to improve their performance on specific tasks or within specific domains. For specialised tasks, fine-tuned models almost always outperform general-purpose pre-trained models. While PLMs are capable of generating coherent text, they may struggle to follow precise instructions and often lack the ability to perform complex, multi-step reasoning. Fine-tuning allows the model to focus on patterns and information relevant to a particular context.

3.2.1 Fine-tuning Setup

Given that the quality of the pre-training data significantly influences the performance of LLMs, high-quality data is deemed crucial for LLMs, even more so than for smaller language models. The capacity of these models largely depends on the pre-training corpus and the methodologies employed in its preprocessing (Jeong, 2023).

Supervised Fine-tuning (SFT) is the most common and the most useful form of fine-tuning. This approach enables the model to learn how to map specific inputs to expected structured outputs based on provided examples. The "supervision" comes from associating each input with its correct labeled output, effectively continuing the process of supervised learning. This method defines clear expectations for model outputs and is particularly effective for adapting models to specialized tasks. This techniques are expected to further enhance this learning process and improve model performance (Dilmegani, 2023).

In this study, we apply supervised fine-tuning to align each model's output with ESG rating logic by using structured input, the output pairs that reflect benchmark scores and rationale. Each training

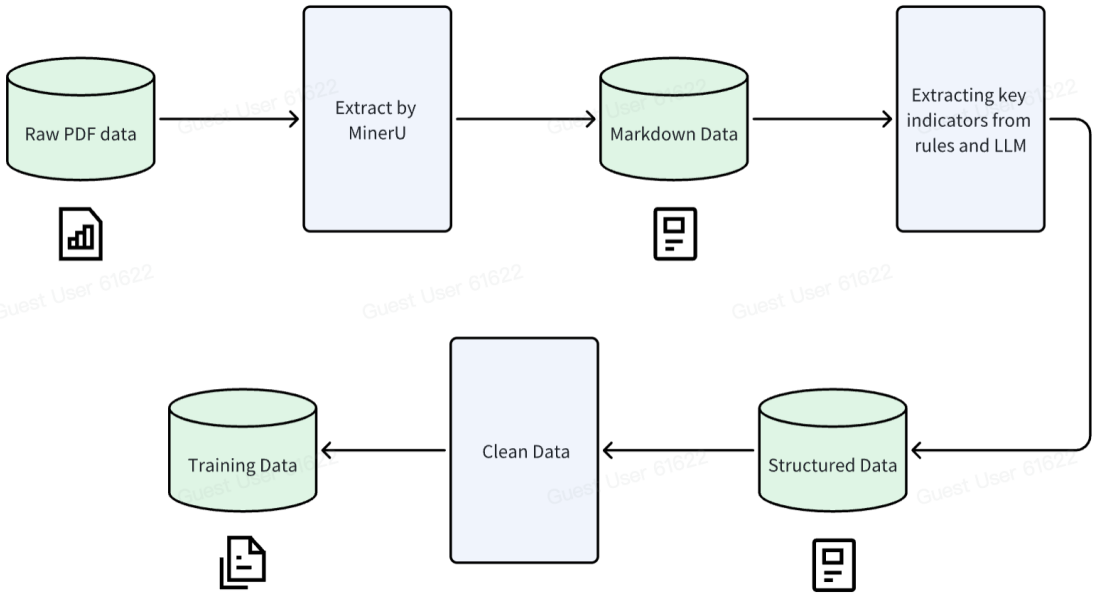


Figure 2: Data preprocessing pipeline.

sample consists of a concatenated ESG and financial report as the input and a structured output comprising a total ESG score, sub-scores for the environmental, social, and governance dimensions, and a textual explanation supporting the scores. The output format is explicitly designed to reflect the logic of institutional ESG ratings, with fixed slots for each score and constraints on explanation length and structure. This facilitates both numerical alignment and interpretability.

3.2.2 Prompt Design

In our implementation, we apply supervised fine-tuning to five LLMs using ESG reports as inputs and benchmark scores as structured targets, with the objective of teaching the models to replicate institutional ESG rating behavior. Each training instance comprises a concatenated document that includes both the ESG and financial reports of a company, serving as the input, and a structured output consisting of a total ESG score, sub-scores for the environmental, social, and governance dimensions, and a concise explanation supporting the assigned values. The output format is carefully designed to reflect the logic and dimensional separation emphasized in benchmark methodologies, ensuring that the models learn not only to predict numerical scores but also to justify them in a manner consistent with Refinitiv and Bloomberg rating principles.

This approach encourages the models not only to approximate institutional ratings numerically but also to internalize the reasoning patterns embedded in those ratings. Supervision signals are derived directly from Refinitiv and Bloomberg 2024 ESG ratings. To ensure consistency and encourage reasoning-based alignment, the output format follows a templated structure inspired by institutional rating criteria. Supervision signals are constructed independently for two dataset, one based on Refinitiv ratings and the other on Bloomberg, such that each model is fine-tuned separately under

each benchmark regime. These benchmarks differ significantly in rating methodology, indicator weighting, and textual emphasis, allowing the fine-tuned models to internalize divergent ESG evaluation frameworks.To facilitate this, we designed a structured prompt template that guides the model in producing ESG ratings alongside textual explanations. Each prompt follows an instruction-based format that explicitly requests both quantitative assessments and qualitative analysis, simulating the decision-making logic of institutional raters.
The prompt begins by assigning the model the role of an ESG evaluation expert and then provides the full disclosure text (concatenated ESG and financial reports) as the evaluation subject. The model is instructed to return a standardized response containing a total ESG score, three dimension-specific sub-scores, and an analysis that identifies both positive and negative aspects of the input document. This format supports alignment with benchmark logic while also encouraging interpretable output.

```
You are an expert in ESG Score, I will give you a text and the score for
that text, you will give me an analysis of the text that matches the score.
NOTE:
1. Your response must have only two parts.
2. The first part is the score, you will start with: "I will rate this
text as a ESG score of: {total_score}/10, environmental score of: {e_score}/10,
social score of: {s_score}/10, and governance score of: {g_score}/10."
3. In the second part, you will analyze the text's positive and negative
aspects, starting with: "The positive aspects of this text are: xxx" and
"The negative aspects of this text are: xxx".
4. Your analysis will be no longer than 500 words.
TEXT:
{text}
```

3.2.3 Training Procedure

The fine-tuning process was conducted separately for each benchmark dataset (Refinitiv and Bloomberg), ensuring that the model learned scoring behavior aligned with only one institutional rating logic at a time. Fine-tuning was performed on five models: GPT-4o-mini, LLaMA 3.1-8B, Qwen 2.5-7B, and two distilled variants: DeepSeek-R1-distill-LLaMA-8B and DeepSeek-R1-distill-Qwen-7B.

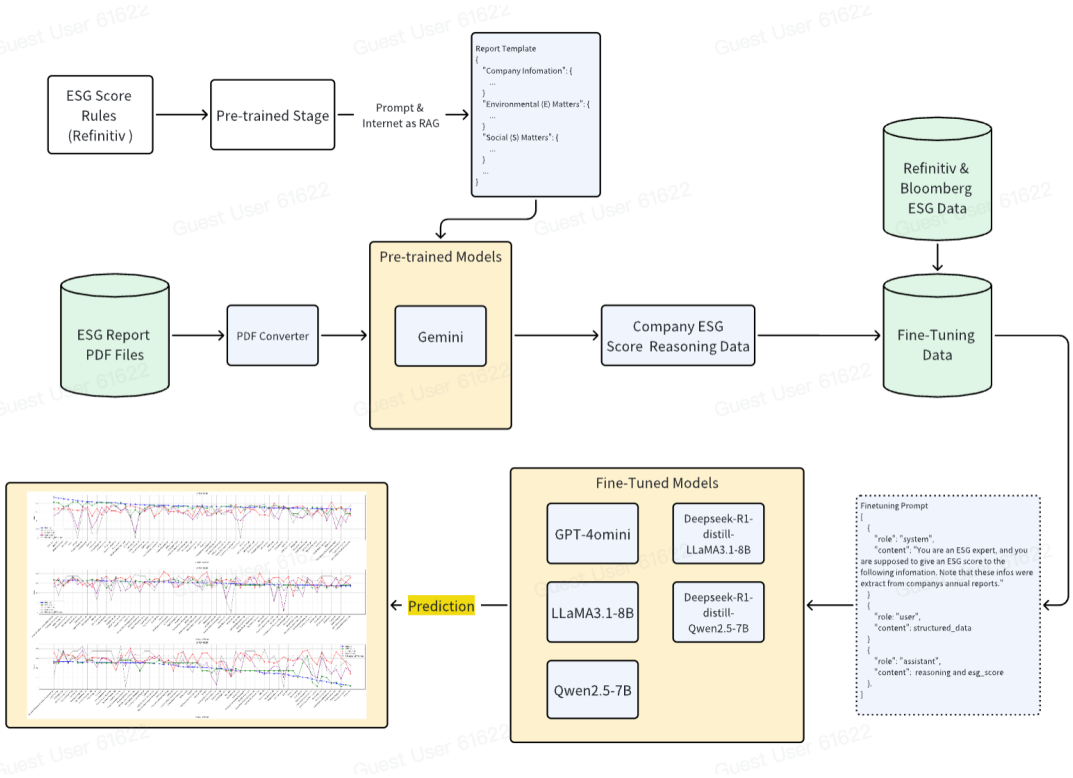


Figure 3: Overall fine-tuning workflow.

For GPT-4o-mini, fine-tuning was conducted using OpenAI's API-based supervised instruction interface, with each document truncated or chunked to a maximum length of 10,000 tokens. The training corpus consisted of 609,748 tokens derived from company-level Bloomberg ESG reports, paired with ESG scores and structured textual explanations. The explanations were generated using Gemini 2.5, based on extracted key passages and heuristic rules. The model was fine-tuned over 3 epochs with a batch size of 1, using cross-entropy loss as the optimization objective.

For the open-source models, LLaMA, Qwen, and their distilled variants, fine-tuning was implemented using the LoRA (Low-Rank Adaptation) technique via the LLaMA-Factory framework. This parameter-efficient approach introduces trainable low-rank matrices into the transformer architecture, significantly reducing computational cost while preserving base model capabilities. The models were trained with a learning rate of 1e-4, batch size of 8, over 5 epochs, and employed cosine annealing learning rate scheduling to stabilize training and prevent overfitting. Fine-tuning was performed using NVIDIA A100 GPUs, with each training job taking approximately 3-4 hours depending on the model size and corpus complexity.

Throughout the fine-tuning process, models were evaluated on held-out validation sets from each dataset using benchmark-aligned metrics such as mean squared error (MSE), accuracy, and explanation coherence. Early stopping was not applied due to the small size of the training corpus and the need for full convergence within limited epochs.

In addition to this, for enhance the interpretability and logical traceability of the ESG scoring process, Chain-of-Thought (CoT) prompting was selectively applied to two distilled models: DeepSeek-R1-

distill-Qwen and DeepSeek-R1-distill-LLaMA. This method is designed to emulate the analytical reasoning procedures employed by human evaluators in the ESG domain, thereby ensuring clearer alignment with the logic underlying institutional benchmark ratings. The CoT prompting approach decomposes the task into an initial reasoning phase, referred to as "thinking", and a subsequent output phase that includes score assignment and justification. By separating reasoning from output generation, this method compels the model to base its evaluations on verifiable information disclosed in ESG and financial reports. In doing so, it explores whether reasoning-enhanced prompting can mitigate the inherent representational limitations of distilled models and reduce the likelihood of hallucinated or spurious reasoning.

4 Results

This essay evaluates five fine-tuned large language models using two ESG rating benchmarks, Bloomberg and Refinitiv. The models assessed include GPT-4o-mini, LLaMA 3.1, Qwen 2.5, and two DeepSeek distilled variants based on the LLaMA and Qwen architectures.

4.1 Benchmark Consistency Evaluation

We used 80 percent of the company list dataset for training and randomly selected the remaining 20 percent as a validation set to assess the scoring performance of each model.
The validation set results provide a comprehensive comparison of the generalization performance of five fine-tuned large language models across two major ESG benchmarks: Refinitiv and Bloomberg. While all models were evaluated using consistent metrics (correlation coefficient, mean deviation, accuracy, and MSE), their relative performance differed substantially across benchmarks, highlighting the interaction between model architecture, training method, and benchmark characteristics.
Overall, models demonstrated stronger predictive alignment and numerical stability on the Bloomberg validation set than on Refinitiv. Notably, Qwen 2.5-7B and GPT4o mini achieved the most consistent results across both datasets.

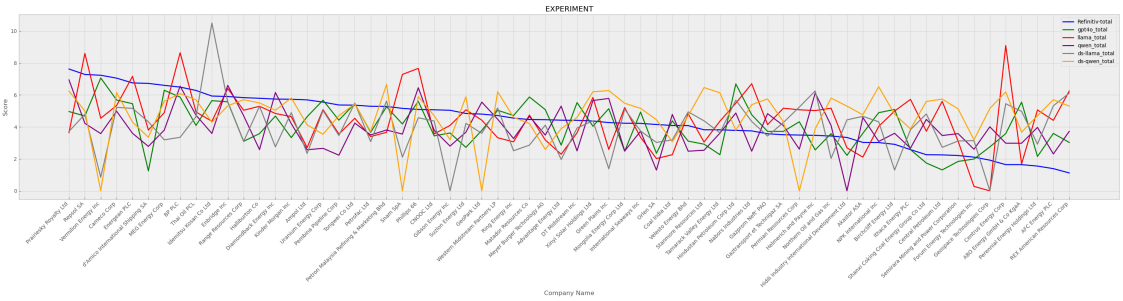


Figure 4: Test Set Benchmark Comparison Line Chart (Bloomberg)

Table 2: Performance Metrics Comparison Across Different Models and Datasets

	ESG Total Score					E Score					S Score					G Score				
	GPT4o	LLaMA	QWen	DS-QWen	DS-LLaMA	GPT4o	LLaMA	QWen	DS-QWen	DS-LLaMA	GPT4o	LLaMA	QWen	DS-QWen	DS-LLaMA	GPT4o	LLaMA	QWen	DS-QWen	DS-LLaMA
Bloomberg Dataset																				
CR ↑	0.26	0.15	0.18	0.01	0.03	0.11	0.21	0.02	0.02	0.04	0.20	0.26	0.05	-0.04	0.05	0.04	0.01	0.02	0.04	0.04
MD ↓	1.40	1.61	1.46	1.79	2.44	1.93	2.16	1.83	2.36	2.48	2.06	2.50	2.50	2.14	2.50	1.61	1.87	1.73	2.03	1.78
ACC ↑	0.17	0.38	0.40	0.28	0.34	0.31	0.28	0.34	0.25	0.30	0.31	0.30	0.22	0.30	0.30	0.41	0.41	0.37	0.39	0.35
MSE ↓	3.19	4.17	3.26	3.82	5.91	5.82	7.39	5.21	8.30	24.1	6.51	6.78	6.00	9.20	7.16	4.39	5.78	4.81	14.76	5.24
Refinitiv Dataset																				
CR ↑	0.06	0.06	0.22	0.01	0.15	0.12	0.02	0.09	0.19	0.55	0.25	0.22	0.20	0.27	0.29	0.22	0.22	0.25	0.23	0.26
MD ↓	22.14	20.23	16.05	22.87	51.81	26.46	25.79	20.30	25.83	26.02	25.20	23.37	19.37	25.69	24.91	25.24	27.22	22.08	27.15	26.04
ACC ↑	0.29	0.33	0.40	0.28	0.32	0.25	0.30	0.31	0.26	0.24	0.26	0.27	0.29	0.24	0.27	0.25	0.25	0.23	0.26	0.26
MSE ↓	753.04	657.54	414.33	855.83	214396.18	1023.88	649.17	1064.58	1028.54	970.15	856.21	561.81	1048.28	979.12	977.49	1209.68	719.46	1179.02	1052.73	1052.73

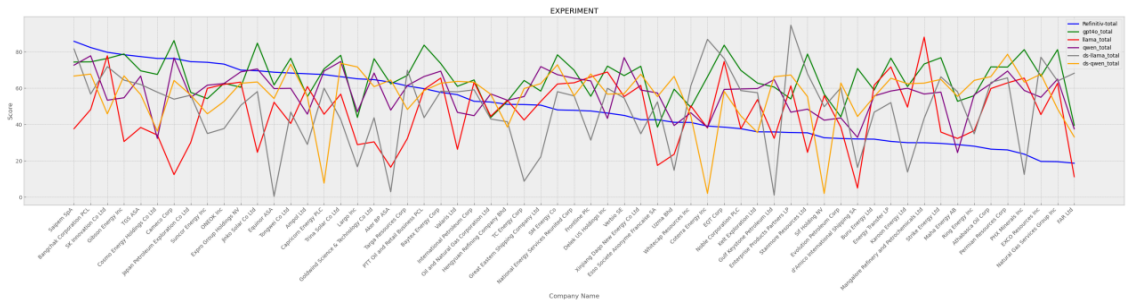


Figure 5: Test Set Benchmark Comparison Line Chart (Refinitiv)

Performance is measured using four standard metrics: mean squared error (MSE), accuracy within a ±0.5 range (ACC), mean deviation (MD), and Pearson correlation coefficient (CR).

Given the discrepancy in rating scales between Refinitiv (ranging from 0 to 100) and Bloomberg (ranging from 0 to 10), we adopt scale-adjusted evaluation thresholds to ensure ACC evaluation consistency across benchmarks. For Refinitiv-based models, a prediction is considered accurate if it falls within ±10 points of the ground truth score. For Bloomberg-based models, the corresponding threshold is set at ±1.0 point. These thresholds represent approximately 10 percent of the respective rating scales, thereby enabling a fair comparison of model performance across distinct institutional standards.

On the Refinitiv benchmark (0-100 scale), Qwen achieved the strongest overall performance among all models. It reported the lowest total MSE (414.33), the highest correlation coefficient (CR = 0.22), and the lowest mean deviation (MD = 16.05), alongside the highest proportion of predictions within a ±10-point tolerance band (ACC = 0.40). These indicators suggest Qwen effectively internalized Refinitiv's scoring framework. GPT4o mini followed with an accuracy of 0.29 and moderate error metrics (MSE = 753.04, MD = 22.14), although its correlation remained low (CR = 0.06), indicating its predictions were close in magnitude but misaligned in relative ranking. LLaMA 3.1-8B showed comparable trends (ACC = 0.33, MSE = 657.54), but exhibited higher deviations in subdimensions, particularly in governance, where CR dropped to 0.02. The distilled variants (DS-Qwen, DS-LLaMA) performed significantly worse across all metrics. For instance, DS-LLaMA yielded an MSE of 214396.18 and negative correlation (-0.01), suggesting limited capacity to capture Refinitiv's heterogeneous scoring logic.

In contrast, on the Bloomberg benchmark (0-10 scale), all models demonstrated improved performance, reflecting the relative clarity and structure of Bloomberg's rating methodology. GPT4o-mini ranked first, achieving a total MSE of 3.26, CR of 0.18, MD of 1.46, and accuracy within ±1.0 of 0.40. GPT4o-mini followed closely, with an MSE of 3.19, accuracy of 0.47 (highest among all models), MD = 1.40, and CR = 0.26, indicating a better balance between magnitude alignment and relative ranking. LLaMA's results were again moderate (ACC = 0.38), with a consistent but lower CR (0.15) and elevated MSE (4.17). Distilled variants continued to lag, particularly DS-LLaMA, which exhibited high MSE (50.5), lowest CR (0.03), and the lowest accuracy (0.34).

These results highlight two key findings. First, Qwen and GPT4o-mini consistently outperforms across both benchmarks, combining strong numerical fit with relatively high correlation, reflecting a

better internalization of ESG rating logic. Second, Bloomberg scores appear more learnable across all models, likely due to their structured and compact format, whereas Refinitiv's broader scope and noisier structure introduce challenges in learning both magnitude and relative ranking. Distilled models, while efficient, show significant degradation in both fit and generalization, underscoring the trade-off between parameter efficiency and reasoning capability in ESG scoring tasks.

The consistently poor performance of the distilled models across both datasets warrants further examination. Despite their computational efficiency, these models appear to lack the representational capacity necessary for accurate ESG evaluation. One possible explanation lies in the knowledge distillation process itself. While distillation aims to transfer the output behavior of a larger teacher model into a smaller student model, this transfer often prioritizes surface-level alignment rather than deep semantic reasoning. ESG scoring, by its nature, requires not only numerical estimation but also structured interpretation of qualitative information found in sustainability reports and financial disclosures. If the distillation process fails to preserve the reasoning mechanisms of the teacher model, particularly the CoT structures or internal scoring logic, then the student model may replicate outputs without fully understanding the underlying rationale. Furthermore, given the multi-dimensional and context-dependent nature of ESG criteria, smaller models may struggle to capture cross-dimensional dependencies or domain-specific nuances, especially when trained with limited supervision. These limitations collectively reduce the interpretability, robustness, and benchmark alignment of the distilled models, as reflected in their lower accuracy, higher deviation, and weaker correlation with reference ratings.

A comparison of the two datasets indicates that model performance is sensitive to the characteristics of the benchmark rating system. GPT-4o-mini performs best when trained and evaluated on the Bloomberg dataset. This is likely due to Bloomberg's relatively stable and transparent scoring logic. Qwen 2.5, on the other hand, adapts better to the Refinitiv benchmark, which presents more variation and less structural clarity. The two distilled models do not reach comparable levels of performance on either dataset. These findings suggest that both model architecture and the nature of the benchmark influence scoring reliability. They also demonstrate that fine-tuned large language models are capable of producing not only ESG ratings but also coherent reasoning explanations that reflect benchmark-specific logic.

4.2 Explanation alignment Quality Analysis

Building upon the performance differences observed in Section 4.1, we find that all five fine-tuned models exhibit stronger predictive consistency and explanation alignment on the Bloomberg dataset compared to Refinitiv. To illustrate this further, we select Permian Resources Corp, a representative U.S. energy company, as a case example. Using the Bloomberg-aligned fine-tuned models, we present a comparative analysis of the ESG scores and model-generated justifications. This example offers a clear lens through which to examine the qualitative divergence in model reasoning, despite relatively close quantitative scores. The comparative analysis presented in Table 3 highlights substantial variation in the quality, structure, and depth of ESG justifications generated by the five fine-tuned models.

Table 3: Explanation Alignment Comparison Across Models (Bloomberg Benchmark)

Model	ESG Total	E/S/G Sub Score	Summary of Explanation Content	Dominant Dimension	Explanation Quality
Bloomberg Benchmark	3.49	3.39/1.20/6.72			
GPT-4o-mini	4.32	2.26/5.52/6.13	Structured, multi-pillar reasoning with emphasis on environmental KPIs and incentive-linked governance practices.	Governance	High: dimensionally complete but stylistically generic.
Qwen 2.5-7B	2.60	2.67/2.37/3.10	Balanced across dimensions; KPI references present but narrative remains template-driven.	Environmental	Moderate–High: dimensionally balanced and benchmark-aware, but lacks nuance and interpretive depth.
LLaMA 3.1-8B	5.07	5.35/4.38/5.06	Covers E and G; explanation contains repetition and omits social sub-score.	Social	Moderate: thematically narrow, lacks structural depth.
DS-Qwen	6.74	6.76/Na/7.30	Discusses progress in all ESG dimensions; notes structured GHG data and limited stakeholder engagement. (Note: S score omitted in output)	E + G (*S missing*)	Low–Moderate: incomplete output with structural flaws.
DS-LLaMA	5.24	6.20/4.66/6.15	Generic statements with minimal justification; social and governance content underdeveloped.	Environmental	Low: low-resolution output with signs of compression artifacts.

First, explanation granularity differs significantly by model architecture. GPT-4o-mini consistently demonstrates the most coherent and human-aligned interpretive reasoning, demonstrating a coherent multi-pillar reasoning framework. Qwen 2.5-7B, while also dimensionally balanced, tends to adopt a template-like structure with limited nuance. LLaMA 3.1-8B exhibits a narrow evaluative lens, focusing disproportionately on social deficiencies, while both distilled models display significant quality degradation. Specifically, DS-Qwen omits the social score entirely in its structured output and presents a repetitive explanation, whereas DS-LLaMA suffers from generic phrasing and low-dimensional coverage. Second, the interpretability does not necessarily follow from score proximity. Models such as DS-LLaMA, although reasonably close in score, produce low-resolution and repetitive explanations, likely a byproduct of model compression and reduced representational capacity.

These findings highlight an important methodological insight: numerical convergence in ESG scoring is insufficient without corresponding alignment in interpretive logic. High-quality model outputs must reflect not only statistical learning of labeled benchmarks but also semantic internalization of institutional rating criteria. Moreover, the observed volatility in the explanations generated by distilled models suggests that compression, while computationally efficient, may compromise model robustness and undermine trustworthiness in decision-critical ESG applications such as portfolio construction, regulatory oversight, and sustainability reporting.

4.3 Ablation Test

To assess the impact of fine-tuning on ESG evaluation performance, we conducted an ablation test comparing fine-tuned models with their pre-trained counterparts. This comparison focuses on two

models that achieved the best performance within their respective benchmark datasets: GPT-4o-mini on the Bloomberg dataset and Qwen 2.5 on the Refinitiv dataset. By contrasting fine-tuned and pre-trained versions of each model, we aim to evaluate whether supervised fine-tuning meaningfully improves alignment with institutional ESG scoring frameworks.

Table 4: Comparison between Pre-trained and Fine-tuned Models on Bloomberg Dataset

Model	Dimension	MSE	ACC (±1.0)	MD	CR
GPT-4o	TOTAL	3.19	0.47	1.40	0.26
	E	5.82	0.31	1.93	0.26
	S	6.51	0.31	2.03	0.20
	G	4.39	0.41	1.61	0.04
GPT-4o_pretrain	TOTAL	17.07	0.04	3.82	0.03
	E	25.63	0.05	4.58	0.06
	S	22.08	0.07	4.24	0.04
	G	4.91	0.33	1.80	0.02

Table 5: Comparison between Pre-trained and Fine-tuned Models on Refinitiv Dataset

Model	Dimension	MSE	ACC (±10.0)	MD	CR
Qwen	TOTAL	414.33	0.40	16.05	0.22
	E	649.17	0.31	20.30	0.25
	S	561.81	0.29	19.37	0.19
	G	719.46	0.25	22.08	0.06
DS-Qwen	TOTAL	603.15	0.32	19.40	0.04
	E	861.74	0.29	23.24	0.09
	S	777.26	0.27	22.29	0.10
	G	912.41	0.25	23.91	0.02
Qwen_pretrain	TOTAL	1441.33	0.11	33.41	0.12
	E	2116.43	0.12	40.13	0.13
	S	1353.88	0.19	30.76	0.10
	G	1237.88	0.24	28.30	0.08

As shown in Tables, fine-tuning led to substantial improvements across all four evaluation metrics: mean squared error (MSE), accuracy (ACC), mean deviation (MD), and Pearson correlation coefficient (CR). On the Bloomberg dataset, the fine-tuned GPT-4o-mini model achieved an MSE of 3.19 and an ACC of 0.47, compared to 17.07 and 0.04, respectively, in the pre-trained version. The correlation coefficient improved from 0.03 to 0.26, and mean deviation decreased from 3.82 to 1.40. Similar trends were observed for Qwen 2.5 on the Refinitiv dataset. The fine-tuned model reduced MSE from 1441.33 to 414.33 and improved ACC from 0.11 to 0.40. Additionally, MD decreased from 33.41 to 16.05, and CR increased from 0.12 to 0.22.

These results confirm that fine-tuning enhances both the numerical accuracy and the consistency of ESG scoring, enabling models to internalize benchmark-specific logic and provide outputs that are more reliable and interpretable. The performance gap between fine-tuned and pre-trained models underscores the importance of domain-specific supervision when deploying large language models for ESG assessment.

Beyond statistical improvements, the practical implications of fine-tuning are significant. Fine-tuned models offer the potential to support more transparent and scalable ESG evaluations, reducing reliance on proprietary rating agencies. This capability can assist investors in making informed decisions, aid regulatory bodies in monitoring corporate sustainability disclosures, and enable real-time ESG analysis in dynamic market environments. Additionally, such models can facilitate early detection of ESG risks, support corporate reporting, and contribute to sustainable finance strategies by offering timely and interpretable assessments of company performance.

In summary, These three parts demonstrate that supervised fine-tuning not only improves alignment with established ESG benchmarks but also enhances the practical utility of language models in real-world ESG applications. These findings suggest that fine-tuned LLMs can serve as effective tools for investment evaluation, regulatory support, and continuous ESG monitoring, thus contributing to more informed and sustainable decision-making across the financial ecosystem.

5 Conclusion

This paper shows that supervised fine-tuned large language models can replicate and explain institutional ESG rating logic from public disclosures in the energy sector. Using two independent benchmarks, we find benchmark-conditioned results: ChatGPT-4o-mini aligns best with Bloomberg ESG Disclosure scores, while Qwen achieves the strongest performance on Refinitiv ESG scores. Fine-tuning improves error, correlation and explanation quality over pre-trained baselines and yields structured outputs that include a total score, pillar scores and a concise text-grounded rationale.

The approach matters for sustainable investing because it turns opaque scores into auditable, evidence-linked assessments. It lowers marginal cost by relying on public text and parameter-efficient adaptation, which allows coverage to extend to thinly rated or unrated issuers. Scores can be refreshed when new reports or material news appear, so users obtain more timely signals rather than waiting for annual cycles. The benefits are concrete for each stakeholder. Investors and asset managers can use near real-time rescoring for screening, event monitoring and stewardship with rationales that cite the sentences used. Risk and compliance teams gain traceable justifications for alerts and controls. Issuers can run low-cost self-assessments against each benchmark, identify disclosure gaps and simulate which actions would move the score. Regulators can triage potential greenwashing by checking whether claims are supported in the primary document.

Limitations remain. The models emulate vendor scores rather than arbitrate between methodologies. Evidence comes from one sector and one reporting year, and part of the training rationales followed a template. Future work will broaden sectors and years, test multilingual settings, add human-in-the-loop review and conduct robustness and fairness audits under distribution shift.

Overall, the framework delivers ESG analytics that are accurate, interpretable and timely, reduces dependence on proprietary feeds and expands coverage at low marginal cost. By making the mapping from evidence to scores explicit under each benchmark, it provides market participants and supervisors with clearer signals to allocate capital and oversee sustainability performance, reduce noise created by rating disagreement and allocate capital more effectively toward sustainable outcomes.

Acknowledgements

We thank our supervisors for their guidance and detailed comments, and colleagues at the Irish Academy of Finance (IAF) for valuable seminar feedback. This research was supported by the Sustainable Energy Authority of Ireland (SEAI).

References

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

[2] Arslan, Y., Ahmed, S., Dengel, A. and Ahmed, S. (2021) 'A comparison of pre-trained language models for multi-class text classification in the financial domain', in *Companion Proceedings of the Web Conference 2021*. ACM, pp. 260–268. https://doi.org/10.1145/3442442.3451375

[3] Berg, F., Kölbel, J.F. and Rigobon, R. (2022) 'Aggregate confusion: The divergence of ESG ratings', *Review of Finance*, 26(6), pp. 1315–1344.

[4] Billio, M., Costola, M., Hristova, I. and Pelizzon, L. (2021) 'Inside the ESG ratings: (Dis)agreement and performance', *Corporate Social Responsibility and Environmental Management*, 28(5), pp. 1426–1445.

[5] Bronzini, M., Nicolini, C., Lepri, B., Passerini, A. and Staiano, J. (2024) 'Glitter or gold? Deriving structured insights from sustainability reports via large language models', *EPJ Data Science*, 13, Article 41. https://doi.org/10.1140/epjds/s13688-024-00481-2

[6] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T. and Zhang, Y. (2023) 'Sparks of artificial general intelligence: Early experiments with GPT-4', *arXiv preprint* arXiv:2303.12712.

[7] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, M., Zhu, H., Li, M., Skiena, S. and Wang, L. (2023) 'Vicuna: An open-source chatbot impressing GPT-4', *arXiv preprint* arXiv:2304.01224.

[8] Clark, K., Luong, M.-T., Le, Q.V. and Manning, C.D. (2020) 'ELECTRA: Pre-training text encoders as discriminators rather than generators', in *Proceedings of ICLR 2020*. OpenReview.

[9] Del Vitto, A., Marazzina, D. and Stocco, D. (2023) 'ESG ratings explainability through machine learning techniques', *Annals of Operations Research* (Early view). https://doi.org/10.1007/s10479-023-05514-z

[10] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of NAACL-HLT 2019*, Minneapolis, MN, pp. 4171–4186.

[11] Ferro, A., Marazzina, D. and Stocco, D. (2025) 'Uncovering ESG ratings: The (im)balance of aspirational and performance features', *Corporate Social Responsibility and Environmental Management*, 32(—), pp. — (Early view).

[12] Liu, X., Yang, Q., Wei, K. and Dai, P.-F. (2024) 'ESG rating disagreement and idiosyncratic return volatility: Evidence from China', *Research in International Business and Finance*, 70, 102368. https://doi.org/10.1016/j.ribaf.2024.102368

[13] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J. and Ziegler, D.M. (2022) 'Training language models to follow instructions with human feedback', *Advances in Neural Information Processing Systems*, 35, pp. 27730–27744.

[14] Refinitiv (2023) 'Environmental, Social and Governance Scores: Methodology'. London Stock Exchange Group. https://www.refinitiv.com/content/dam/marketing/en_us/documents/methodology/refinitiv-esg-scores-methodology.pdf

[15] Shah, N., Shah, S. and Ramachandran, R. (2022) 'FLANG: Financial language understanding with pre-trained models', *arXiv preprint* arXiv:2210.13077.

[16] Sustainability (2020) 'Rate the Raters 2020: Investor survey and interview results'. https://www.sustainability.com/globalassets/sustainability.com/thinking/pdfs/sustainability-ratetheraters2020-report.pdf

[17] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P. and Hashimoto, T.B. (2023) 'Alpaca: A strong, replicable instruction-following model', Stanford Center for Research on Foundation Models, Technical Report.

[18] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E. and Lample, G. (2023) 'LLaMA: Open and efficient foundation language models', *arXiv preprint* arXiv:2302.13971.

[19] Wu, S., Irsoy, O., Lu, S., Anchuri, K., Khabsa, M., Ioannou, Y., Clement, C.B., Lahnala, A., Chang, M.-W., Li, X., Li, S., Ma, W., Muralidharan, S., Li, X., Zhang, Y., Liu, J., Nallapati, R., Arnold, A.O., Xu, J., Subramanian, S., Gao, J., Dreyer, M., Liden, L., Pereira, J., Reganti, A. and Ainsworth, S. (2023) 'BloombergGPT: A large language model for finance', *arXiv preprint* arXiv:2303.17564.

[20] Xia, L., Yang, M. and Liu, Q. (2024) 'Using pre-trained language model for accurate ESG prediction', in *Proceedings of the 8th FinNLP Workshop (Financial NLP) at IJCAI*, pp. —.

[21] Yang, X., Sun, X., Shen, H. and He, B. (2020) 'FinBERT: A pretrained language model for financial communications', in *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, Barcelona, pp. 3393–3403.