Using Pre-trained Language Model for Accurate ESG Prediction

Lei Xia[1,2*], Mingming Yang[2], Qi Liu[1†],
[1]The University of Hong Kong, [2]Tencent AI Lab,
brianleixia@connect.hku.hk, shanemmyang@tencent.com, liuqi@cs.hku.hk

Abstract

Environmental, Social, and Governance (ESG) has been crucial in investment decision-making in recent years, with an increase in ESG-centric research emerging. Concurrently, Natural Language Processing (NLP) has emerged in analyzing ESG-related texts. However, there is a lack of models and datasets specifically tailored for ESG categorization. This study presents a novel approach leveraging Pretrained Language Models (PLMs) and Large Language Models (LLMs) to tackle ESG text classification tasks. We introduce a pipeline for creating specialized datasets for ESG analysis by using keyword search and LLMs APIs to label data. Through continued pre-training PLMs such as BERT, DistilRoBERTa, and RoBERTa on our datasets, our approach significantly surpasses traditional baseline performances. We also introduce ESGLlama and FinLlama, domain-specific models derived from Llama2, with FinLlama demonstrating exceptional efficacy in financial benchmarks and ESG text comprehensions [1]. Final evaluations reveal that our models achieve significant advancements in ESG classification, outperforming established baselines.

1 Introduction

Environmental, Social, and Governance (ESG) considerations represent the cornerstone of contemporary sustainable or responsible investment strategies. Over the past decade, ESG has become the preeminent framework for socially responsible investments and decision-making within the financial sector. However, a significant challenge remains relying on voluminous annual sustainability reports for informed decision-making. The comprehensive nature of these reports necessitates substantial effort for thorough analysis, highlighting the critical demand for automated solutions. In this context, Natural Language Processing (NLP) emerges as an indispensable tool, enabling navigating through extensive sustainability narratives and extracting pivotal ESG insights precisely.

Recent advancements in NLP have streamlined the identification and interpretation of ESG information, enabling more nuanced analysis. This research background sets the stage for exploring the integration of NLP in enhancing the efficiency and depth of ESG analysis. Additionally, existing research has applied pre-trained language models (PLMs) in ESG-related NLP tasks such as climate change-related text detection and controversy detection (Nugent et al., 2020; Huang et al., 2023; Schimanski et al., 2023; Webersinke et al., 2021). However, a significant gap exists in the processing and collecting textual ESG data. This results in a scarcity of publicly accessible, high-quality ESG textual datasets, especially for established text categorization tasks within the ESG domain. Recent developments in large language models (LLMs) are more powerful than small PLMs and have demonstrated their potential in performing various NLP tasks like language understanding and generation. But, no such research focuses on using LLMs to solve ESG-related tasks.

In this study, we tackle the significant gap in the availability of ESG-related datasets and apply PLMs and LLMs to challenging ESG classification tasks. We also use keyword search and LLMs APIs to annotate datasets for both 4-class and 9-class ESG classification. Further, we enriched our dataset collection with conversational history data, which proved crucial for Supervised Fine-Tuning (SFT) processes. SFT, a pivotal concept in our approach, refers to refining pre-trained models by training them on labeled datasets specific to the target task, thus enabling the models to learn task-specific patterns and adapt to the problem domain. This comprehensive fine-tuning involved

[*]Work was done when Lei Xia was interning at Tencent AI Lab.

[†]Qi Liu is the corresponding author.

[1]Our code and data can be found at https://github.com/brianleixia/LLM4ESGPrediction

Proceedings of the Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen), Jeju, South Korea, August 3, 2024

both PLMs and LLMs, significantly boosting their performance in ESG-related tasks. Moreover, we developed two fine-tuned LLMs, ESGLlama and FinLlama, based on the Llama2, which demonstrated substantial improvements over baseline models. FinLlama also excelled in financial benchmarks. In summary, our key contributions are the following:

- We propose a pipeline by utilizing keyword search and LLMs APIs to annotate data and construct three datasets for ESG analysis: pre-training corpus, classification dataset, and ESG SFT dataset.
- We introduce three domain-specific PLMs: ESG-BERT, ESG-DistilRoBERTa, and ESG-RoBERTa. These models notably surpass their base models and baseline.
- We conduct two fine-tuned Llama2 models: ESGLlama and FinLlama. FinLlama exhibits remarkable improvements in financial benchmarks.
- We compare PLMs and LLMs across various experimental settings, comprehensively analyzing their performance.

2 Datasets Construction Pipeline

In response to the notable scarcity of datasets tailored for ESG domain analysis, we propose a pipeline, as illustrated in Figure 1, which encompasses data preprocessing, labeling procedures, and model training to enhance ESG data analysis capabilities systematically. Initially, data is sourced from various open sources and cleansed according to predefined rules. During the preprocessing phase, data is preliminarily categorized using keyword searches. Subsequent labeling employs APIs from LLMs to ensure high classification accuracy. Human evaluations are conducted to validate the labeled data, which then facilitates the construction of specialized datasets for further model pre-training and fine-tuning.

Specifically, we have constructed three types of datasets to enhance the accuracy of ESG prediction tasks: (1) Pre-training Dataset. This expansive corpus of ESG-related texts is designed to bolster the initial training of domain-specific models, thereby improving their ability to interpret ESG contexts accurately. (2) Classification Datasets. These datasets are segmented into four-class and nine-class categories for ESG texts, playing a pivotal role in the fine-tuning process to enhance model precision in ESG categorization. (3) SFT Dataset. Tailored for the Supervised Fine-Tuning (SFT) of Large Language Models (LLMs), this dataset incorporates conversational data generated by LLMs during the labeling procedure to boost the models' proficiency in ESG classification tasks.

2.1 Data Collection and Processing

For data collection, we searched and collected datasets mainly from two resources: huggingface [2] and kaggle [3]. Refer to more details of our collected data in Appendix B. After data collection, we extract textual content pertinent to ESG analysis. We standardized the datasets to a **sentence-level** format in the initial data processing phase, facilitating uniform analysis across diverse data sources. Following the standardization, a data-cleaning procedure was implemented. This involved the removal of URLs and special characters from the text, ensuring that the datasets were devoid of extraneous information that could potentially skew the analysis. The processed data amounted to approximately 18 million sentences.

2.2 Data Labeling

Keyword Search The keyword search initiates data identification across ESG subdomains, segregating text relevant to Env, Soc, Gov and Non-ESG content. This meticulous process enabled us to partition the corpus into distinct segments, each corresponding to a specific aspect of ESG. While this method predominantly isolated relevant ESG-related text, it is essential to acknowledge that it might not entirely preclude the presence of Non-ESG data within these preliminary datasets. We argue that Non-ESG data within the pre-training phase could inadvertently enhance the model's robustness by exposing it to a broader spectrum of textual content. Details of keywords are in Appendix C.

After filtering the texts by keyword searching, we got the preliminary results in Table 8. To validate the effectiveness of our classification approach, these visualizations effectively confirm the appropriateness of the categorized data, with predominant terms such as "GHG emission" and "climate change" in the Environmental domain, "human

[2] https://huggingface.co/
[3] https://www.kaggle.com/

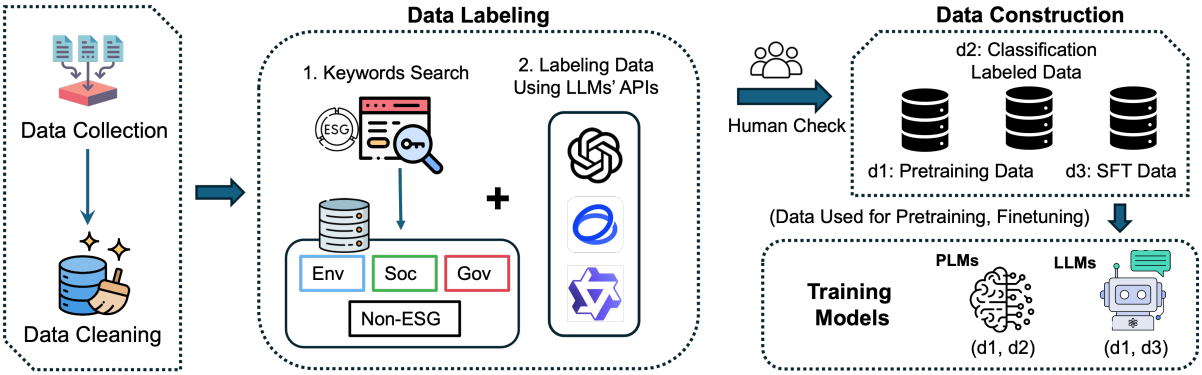


Figure 1: The work pipeline encompasses data collection, preprocessing, and labeling, followed by model training. Data is initially collected from open sources and cleansed. Using keyword searches and enhancing label accuracy through LLM's APIs, with further validation by human evaluation. The resultant dataset is used for pre-training and fine-tuning classification tasks.

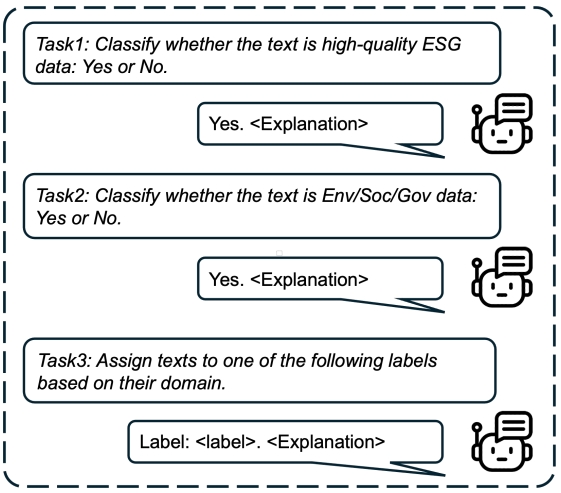


Figure 2: Representation of task decomposition and the task descriptions alongside responses from LLM

rights" and "customer" in Social, and "director" and "financial statement" in Governance, reflecting the accurate representation of domain-specific high-frequency words. Our next objective was to refine the accuracy of our labeled data further. To achieve this, we planned to leverage LLMs for an additional layer of filtering and validation. Details of visualizations are in Appendix D.

Labeling Data Using LLMs Before labeling the data, we recognized a complexity gradient in categorization tasks, where tasks with fewer categories are inherently simpler than those with more. Studies such as Bang et al. (2023) (Bang et al., 2023) suggest that LLMs may underperform in specific, challenging downstream tasks, including multi-class classification tasks. To address this, we devised a structured approach to simplify the ESG classification challenge, as depicted in Figure 2. In this stage, the overall task is divided into three simpler tasks, where Task1 and Task2 comprise the four-class task (Env, Soc, Gov, Non-ESG), and an additional Task3 is required to construct the nine-class task. Specifically, the nine-class classification involves three environmental categories (Climate Change, Natural Capital, Pollution and Waste), three social categories (Human Capital, Product Liability, Community Relations), two governance categories (Corporate Governance, Business Ethics and Values), and one Non-ESG category. The final three categories of the nine-class task are unified into a single ternary (3-class) task, applying the same categorization principles as the four-class task but with an added layer of specificity. Significantly, this ternary categorization is based on data already classified under the four-class schema, further refining the process.

For each sub-task, we employed APIs from three different LLMs: Qwen (*qwen-max*), GLM (*glm-4*), and GPT-3.5 (*gpt-3.5-turbo-instruct*). This multi-model strategy was underpinned by several rationales: Firstly, LLMs are prone to 'hallucination', often generating less reliable outputs due to their randomness. Utilizing multiple models helps mitigate significant data bias and enhances the diversity of the labeled data. Secondly, the decision to leverage several LLMs' APIs was economically driven, aiming to reduce costs associated with extensive data filtering and labeling tasks. Lastly, employing multiple models concurrently significantly enhances the efficiency of the data labeling process. Details regarding the prompt design and an example of LLM response are in Appendix F.

2.3 Data Construction and Analysis

Pre-training Dataset In constructing the pre-training dataset, we initially aggregated datasets categorized as Env, Soc, and Gov based on key-

word searches. Recognizing the challenges associated with processing excessively long texts, we implemented a filtration step to exclude these from the dataset. Then, we executed a 90-10 split to segregate the data into training and evaluation subsets. The evaluation set is crucial in monitoring the training loss and establishing an early stop during the pre-training phase.

Classification Dataset The development of the labeled classification dataset involved multiple meticulous steps. Initially, we processed the outputs from the llms used for each classification task and subjected these to a rigorous human review to verify the LLM-generated classifications. This review process was crucial as it helped refine the data for the four-class and nine-class categorizations, specifically excluding Non-ESG data due to its inherent complexities and the limitations of LLM outputs, which may not always guarantee the absolute accuracy of the responses. Consequently, the Non-ESG dataset was compiled in a two-fold approach: approximately 8,500 samples were selected from the LLM responses, and an additional 5,500 samples were isolated following a keyword search, cumulatively amounting to around 14,000 Non-ESG samples. A notable issue identified was the class imbalance within the nine-class dataset. To rectify this, we implemented a normalization strategy by capping the maximum number of instances per class at 3,000, leading to a more balanced distribution. Furthermore, we applied stratified sampling for both datasets to ensure equitable class representation. Details of dataset distribution are in Appendix E.

Supervised Fine-Tuning Dataset SFT is a critical refinement process in NLP, enhancing a large language model's adaptability to specific tasks. This alignment improves the model's precision and adaptability for specific tasks. In line with best practices like those demonstrated by the Alpaca model (Taori et al., 2023), its instruction dataset has three fields: instruction, input, and output. We constructed a SFT Dataset similarly for ESG classification tasks with the following instructions:

1. Identification of ESG-related text: "If the following text is ESG related data."

2. Four-Class classification: "Classify the following text into one of the four ESG categories: {*categories*}."

3. Nine-category Class: "Classify the following text into one of the nine ESG categories: {*categories*}."

The dataset preparation involved reformatting existing four-class and nine-class datasets to align with these instructions, generating 95,412 data points. We also employed stratified sampling to select about 28,000 data points, ensuring diverse and balanced coverage across the instructions.

3 Methodology

3.1 Pre-trained Based Method

Baseline Our baseline employs FinBERT (Huang et al., 2023), a model adapted from BERT for the financial sector. FinBERT has been extended to address ESG-related classifications.

Datasets The dataset used for pre-training, detailed in Section 2.3, comprises 5,257,347 training sentences and 584,150 validation sentences, obtained via keyword search. While keyword searches are prone to including non-ESG phrases, resulting in false positives, this is beneficial for pre-training. It allows the model to learn the broader context of sustainability topics by exposing it to relevant and irrelevant samples.

Training Models As detailed in Section 2.3, we utilized this dataset to pre-train models including BERT (Devlin et al., 2019), DistilRoBERTa (Sanh et al., 2019), and RoBERTa (Liu et al., 2019). Instead of starting from scratch, we engaged in Continual Pre-Training (CPT), a strategy that allows a model to assimilate new data while preserving previously acquired knowledge. This approach is advantageous for adapting models to evolving data streams or new, unseen data. By continuing to pre-train on an established model's checkpoint, we infused domain-specific ESG knowledge into the models. Consequently, we selected the model with the smallest validation loss as our final pre-training models: `ESG-BERT`, `ESG-DistilRoBERTa`, and `ESG-RoBERTa`. Details regarding pre-training process are in Appendix G.

3.2 LLM Based Method

Baseline Llama2 (Touvron et al., 2023), an open-source large language model. We choose Llama2 (`Llama2-7b-chat-hf`) as a baseline for the ESG classification task.

Datasets Our LLM-based methods utilize two main types of datasets: the pre-training corpus and Supervised Fine-Tuning (SFT) datasets. The pre-training corpus has been substantially expanded to include not only the ESG-related texts discussed in Section 2.3 but also a significant volume of financial texts, primarily sourced from financial reports, totaling 5,282,943 sentences. For SFT, we employed two distinct datasets. The first SFT dataset, as introduced in Section 2.3, consists of conversational data generated during the labeling of ESG data using LLMs. The second SFT dataset is more extensive, integrating the conversational data and additional financial instruction tuning data as outlined in FinGPT (Wang et al., 2023) and the ESG_Chat dataset [4]. The ESG_Chat dataset comprises dialogues between humans and LLMs, focusing on strategies to enhance ESG scores. Then, we adopted a targeted sampling strategy, producing a refined subset of 86,425 sentences.

Fine-tuning Models To enhance the LLM's understanding of ESG-related themes, we enriched the model with ESG-related knowledge, resulting in the creation of two specialized models: ***ESGLlama*** and ***FinLlama***. ESGLlama underwent fine-tuning through Supervised Fine-Tuning (SFT) using conversational data tailored for ESG classification tasks, notably improving its accuracy within ESG contexts (as discussed in datasets, the first SFT dataset). Meanwhile, FinLlama was developed to tackle a broader spectrum of financial tasks, integrating extensive financial texts and targeted instruction-tuning data, ranging from sentiment analysis to financial Question Answering (QA). For fine-tuning FinLlama, we employed a two-stage training approach. Initially, the Llama2 model underwent Continual Pre-Training (CPT) using a combined corpus of ESG-centric texts and additional financial documents, including financial news and annual reports. Subsequently, in the second stage, we conducted supervised fine-tuning on the model pre-trained in the initial phase using the second SFT dataset (as discussed in datasets).

4 Experiments

4.1 Test on Public Dataset

To evaluate the generalizability of our trained models for ESG-related tasks, we conducted tests using publicly available datasets: environmental_2k [5], social_2k[6] and governance_2k[7] which are derived from annual reports spanning 2017-2021. Each dataset is expertly annotated for binary classification, where '0' indicates "No" and '1' denotes "Yes" outcomes. We fine-tuned our models ESG-BERT, ESG-RoBERTa, and ESG-DistilRoBERTa on these datasets with a partitioning scheme of 64% training, 16% validation, and 20% testing.

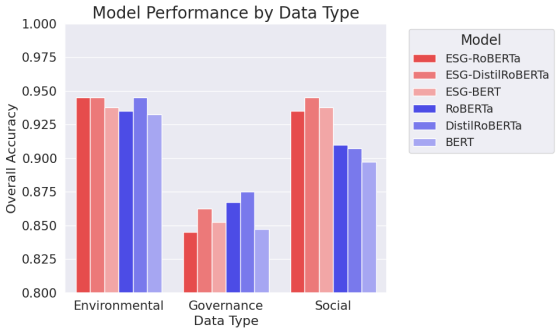


Figure 3: Overall perf. of models on public datasets

To evaluate the effectiveness of our models, the results are shown in Table 1. As we can see, the pre-trained models demonstrate superior performance compared to the baselines across the publicly accessible dataset. Notably, all pre-trained models consistently outperform their corresponding baseline models within the Social domain shown in Figure 3. Among them, ESG-DistilRoBERTa stands out with the highest precision (0.9415), recall (0.9449), and F1 score (0.9431), indicating robust performance. In the Environmental domain, ESG-RoBERTa shows remarkable precision (0.9436) and an equivalent F1 score, underscoring its effectiveness.

However, the Governance domain exhibits a contrasting scenario, with mixed results despite pre-training enhancements. The baseline DistilRoBERTa model outperforms the pre-trained versions in this domain, achieving the highest metrics with a precision of 0.8404, recall of 0.8444, and F1 score of 0.8424. This discrepancy suggests that while pre-training generally enhances model capabilities, its impact is less pronounced in the Governance domain. The observed variance may stem from misalignments between the pre-training content and the specifics of the publicly available

[4] https://huggingface.co/datasets/zadhart/ESG_Chat

[5] https://huggingface.co/datasets/ESGBERT/environmental_2k

[6] https://huggingface.co/datasets/ESGBERT/social_2k

[7] https://huggingface.co/datasets/ESGBERT/governance_2k

Model	Env			Soc			Gov		
	P	R	F1	P	R	F1	P	R	F1
BERT	0.9207	**0.9285**	0.9244	0.8960	0.8899	0.8927	0.8048	**0.8168**	0.8104
ESG-BERT	**0.9300**	0.9284	**0.9292**	**0.9354**	**0.9345**	**0.935**	**0.8141**	0.8085	**0.8112**
DistilRoBERTa	0.9340	**0.9436**	**0.9385**	0.9035	0.9044	0.9039	**0.8404**	**0.8444**	**0.8424**
ESG-DisrilRoBERTa	**0.9364**	0.9397	0.9380	**0.9415**	**0.9449**	**0.9431**	0.8252	0.8271	0.8261
RoBERTa	0.9279	0.9246	0.9262	0.9041	0.9135	0.9076	**0.8292**	**0.8421**	**0.8352**
ESG-RoBERTa	**0.9340**	**0.9436**	**0.9385**	**0.9311**	**0.9345**	**0.9327**	0.8048	0.7976	0.8011

Table 1: Performance metrics across environmental, social, and governance domains on public datasets. **Bold** shows the best results among baseline and corresponding pre-trained model, and underlined indicates the best results in each column.

governance data, suggesting a need to refine the fine-tuning parameters better to tailor the models to this domain's nuances.

4.2 Test on Classification Datasets

Evaluate PLMs We fine-tuned our pre-trained models on ESG classification tasks (four-class and nine-class) using our constructed classification data. The training parameters were standardized at a batch size of 32 across 50 epochs while learning rates were adjusted based on model and task specifics. For the four-class classification, the learning rates were set at 3e-6 for the BERT model and 1.25e-6 for both DistilRoBERTa and RoBERTa. For the nine-class task, BERT was fine-tuned at 3e-6, DistilRoBERTa at 1.75e-6, and RoBERTa at 1.15e-6. These rates were meticulously selected to optimize each model's performance on its respective task. An *early stopping* mechanism was implemented during fine-tuning to curb overfitting and enhance computational efficiency. The models chosen for further utilization demonstrated the best performance on the validation set across the 50 epochs, specifically those achieving the lowest validation loss.

Table 2: Four-Class Evaluation Results of PLMs

Model	P	R	F1	Acc
FinBERT	0.7357	0.7150	0.7165	0.7222
BERT	0.8668	0.8658	0.8641	0.8667
dtRoBERTa	0.8672	0.8687	0.8662	0.8684
RoBERTa	0.8610	0.8596	0.8582	0.8602
ESG-BERT	0.9074	0.9077	0.9071	0.9083
ESG-dtRoBERTa	0.9027	0.9040	0.9014	0.9034
ESG-RoBERTa	**0.9086**	**0.9100**	**0.9086**	**0.9102**

To assess the effectiveness of our pretrained models, we conducted tests on two sets: a four-class and a nine-class classification task, with results detailed in Table 2 and Table 3, respectively. The

Table 3: Nine-Class Evaluation Results of PLMs

Model	P	R	F1	Acc
FinBERT	0.7160	0.7154	0.7081	0.7273
BERT	0.8393	0.8357	0.8361	0.8419
dtRoBERTa	0.8240	0.8153	0.8179	0.8239
RoBERTa	0.8187	0.8196	0.8174	0.8275
ESG-BERT	0.8606	**0.8637**	**0.8617**	**0.8693**
ESG-dtRoBERTa	0.8575	0.8552	0.8556	0.8616
ESG-RoBERTa	**0.8611**	0.8591	0.8592	0.8662

evaluations included baseline models, our specifically pre-trained models, and their base models. For the four-class task, ESG-RoBERTa excelled, achieving the highest metrics with a precision of 0.9086, a recall of 0.9100, an F1 score of 0.9086, and an accuracy of 0.9102, significantly surpassing the baseline finbert-esg model, which only reached an accuracy of 0.7222. This demonstrates a clear superiority over the baseline, with even the base models outperforming finbert-esg when fine-tuned. In the nine-class task, ESG-BERT led with the highest recall of 0.8637 and an F1 score of 0.8617, while ESG-RoBERTa achieved the top accuracy of 0.8662. These results highlight the advantages of our ESG-specific pretraining and fine-tuning strategy, markedly improving upon the performance of the baseline finbert-esg-9-categories model.

Evaluate LLMs We will evaluate the performance of the baseline and our fine-tuned models across six different *experimental settings*: Zero-Shot, One-Shot, In-Context Learning (ICL), Zero-Shot with Chain of Thought (CoT) (Kojima et al., 2023), One-Shot with CoT, and ICL with CoT. The dataset used for SFT in ESG text classification was constructed from ESG SFT data, as detailed in Section 2.3. More details about the ESG classification SFT dataset can be found in Appendix I. To process the results from our models, particularly the baseline, we utilized a *regular expression* matching technique to extract predicted labels from model

outputs. Details regarding classification prompts design are in Appendix J.

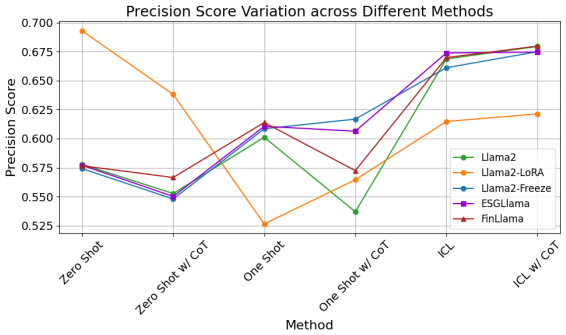


Figure 4: Four-Class Precision of LLMs

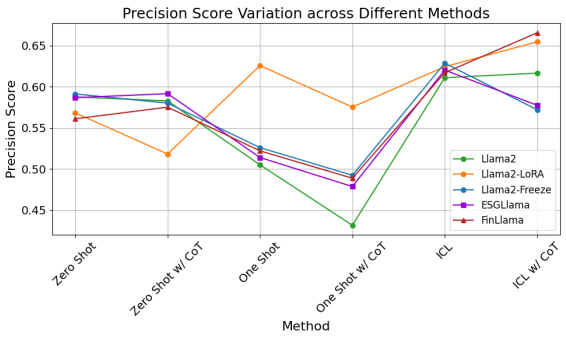


Figure 5: Nour-Class Precisions of LLMs

For four-class classification, Figure 4 shows that both models consistently outperform the baseline across most experimental settings. Notably, even the baseline model improves significantly when subjected to SFT with our ESG classification dataset. Interestingly, the Freeze fine-tuning method generally surpasses the LoRA approach, except in zero-shot scenarios where LoRA excels, possibly indicating its tendency to overfit slightly. The integration of CoT prompts typically reduces performance in zero-shot and one-shot settings, except for ICL tasks. This reduction may stem from CoT's incompatibility with classification tasks, which require straightforward decision-making rather than stepwise logic processing. However, incorporating demonstration examples in ICL tasks enhances the model's grasp of classification logic, significantly improving outcomes in ICL-CoT settings by providing richer context and sample diversity. Furthermore, FinLlama achieves superior precision over ESGLlama with the addition of CoT.

In the nine-class classification, the increase in category complexity and diversity presents more significant challenges, as indicated by lower overall performance metrics than in the four-class scenario. Performance visualization in Figure 5 shows that both ESGLlama and FinLlama substantially outperform the baseline across most configurations, affirming the enhanced capability of our fine-tuned models in handling ESG-related texts. FinLlama excels in ICL, mainly when provided with ample examples, showcasing its deep understanding of the financial domain. Conversely, the performance notably drops in one-shot learning scenarios, where providing a single instance per class introduces significant bias and variability, impairing the model's accuracy. However, increasing the number of examples markedly improves performance, underscoring the benefits of more extensive training datasets. The comparison between LoRA and Freeze methods reveals that LoRA outperforms Freeze in one-shot settings, suggesting that LoRA's parameter adjustments are better suited for absorbing limited class-specific information efficiently. Additional analyses are in Appendix K.

4.3 Test on Financial Benchmark

To assess the FinLlama model's performance in financial NLP tasks, we evaluate it on FinGPT benchmark (Wang et al., 2023). Our evaluation concentrated on two critical tasks: financial text sentiment analysis and headline classification, utilizing the fingpt-headline dataset [8]. Results, presented in Table 4, clearly show that FinLlama significantly outperforms the baseline Llama2 model across these tasks. This superior performance across financial sentiment analysis and headline classification tasks validates the effectiveness of our constructed pre-training and Supervised Fine-Tuning (SFT) datasets.

Table 4: Perf. of models on Financial Benchmarks

	Llama2		FinLlama	
Dataset	**Acc**	**F1**	**Acc**	**F1**
FPB	0.4703	0.4140	**0.7855**	**0.7838**
FiQA	**0.7964**	0.7744	0.7782	**0.8096**
TFNS	0.3811	0.3037	**0.8405**	**0.8408**
NWGI	0.5656	0.4833	**0.6501**	**0.6445**
Headline	0.4314	0.6182	**0.8783**	**0.6975**

5 Results Analysis

Performance of Pre-trained Models Our analysis highlighted that classification task complexity increases with the number of categories. This was evident from the lower convergence rates in

[8] https://huggingface.co/datasets/FinGPT/fingpt-headline

the nine-class task compared to the four-class task. ESG-RoBERTa excelled in the four-class task due to its larger parameter set, which enhances its text understanding capabilities. In contrast, ESG-BERT performed better in the nine-class task, suggesting that its pretraining objectives and architecture might offer superior generalization across more diverse categories. Performance evaluations on a publicly available dataset confirmed the effectiveness of our pre-trained models, as shown in Table 1. Particularly in the Social domain, reflecting the quality of our pre-trained models. The extensive testing on a public dataset validated our pretraining dataset's quality and demonstrated our models' improved comprehension of ESG-related content, enhancing classification accuracy.

Performance of Large Models Both ESGLlama and FinLlama consistently outperform the baseline across most testing scenarios, with notable improvements in the baseline model following SFT with our ESG classification dataset. This enhancement highlights the dataset's quality and the effectiveness of SFT. A distinct observation is Freeze is generally better than LoRA because the trend of line changes in its results is consistent with those of other experimental setups, and more examples can improve its results. The integration of CoT typically reduces performance in zero-shot and one-shot scenarios. Still, it improves outcomes in ICL tasks due to additional context and examples provided. Transitioning to a nine-class framework increases task complexity, generally lowering performance metrics. In ICL tasks, FinLlama shows superior proficiency, particularly when additional samples are included, reflecting its adeptness at navigating complex classification landscapes. Conversely, performance drops in one-shot scenarios, underscoring the challenges of minimal data learning. Moreover, testing FinLlama on financial benchmark further validates its superiority in financial NLP, highlighting FinLlama's effectiveness in financial discourse analysis.

Case Study In this analysis, we compare the performance of ESG-RoBERTa and FinLlama in deciphering complex ESG-related texts. An example from Table 5 shows ESG-RoBERTa misclassifying a case as "Business Ethics and Values," likely due to its reliance on keywords such as "reputation" and "financial costs" associated with ethical issues. Conversely, FinLlama employs ICL with CoT reasoning to accurately categorize the exact text as "Product Liability." It provides a detailed rationale, dissecting the text to link it to product liability issues and showcasing its ability to perform stepwise, human-like reasoning for accurate and interpretable results. FinLlama's response highlights its adeptness at integrating extensive contextual information and performing better than PLMs in this scenario. This case study underscores fine-tuned LLMs' advanced comprehension and reasoning capabilities in detailed, context-rich environments. Despite some limitations, it demonstrates considerable scope to further enhance LLMs for more effective ESG text categorization.

Text: *A warranty or product liability claim whether or not meritorious could harm our reputation result in substantial financial costs or divert managements attention which could have an adverse effect on our business operating results and financial condition.* **True Label:** *Product Liability*
ESG-RoBERTa **Answer:** *Business Ethics and Values* ✗
FinLlama (ICL-CoT): **Response:** Label: Product Liability\n\nExplanation: The text mentions the potential for a warranty or product liability claim to harm the company's reputation, result in financial costs, and divert management's attention, which could have an adverse effect on the company's business operations and financial condition. This falls under the category of Product Liability, as it relates to the potential legal liability of a company for any defects or hazards in its products. **Answer:** *Product Liability* ✓

Table 5: An example of results of PLM and LLM.

6 Conclusion and Future Work

We proposed a pipeline to address the lack of ESG-related datasets, utilizing keyword searches and LLM APIs to annotate and construct three types of data for ESG text classification tasks. This approach has significantly enhanced the performance of pre-trained models on ESG classification tasks. We also introduced domain-specific LLMs, ESGLlama and FinLlama, which were fine-tuned on our datasets, marking a major advancement in applying LLMs to ESG-related challenges. Notably, FinLlama has surpassed existing financial benchmarks. Comparative analysis reveals that while PLMs generally perform better, LLMs offer greater interpretability and adeptly handle complex contexts by integrating contextual information. Moving forward, we will further evaluate our developed datasets, and leverage the superior classification accuracy of PLMs to enhance and refine LLMs' performance in ESG analysis.

Limitations

The limitations of the current work include: (1) The present model is not equipped to handle long text data (e.g., document-level data) as our data are normalized to the sentence level. Future work will focus on model training and inference with long text data. (2) The current dataset presents a cross-domain issue, where a text may pertain to both environmental and governance categories. In the future, we will refine our dataset to enhance its classification clarity and granularity, ensuring texts are either distinctly classified into specific categories or appropriately labeled as belonging to multiple categories.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings. *arXiv preprint*. ArXiv:1904.03323 [cs].

Alex Andonian, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Shivanshu Purohit, Tri Songz, Wang Phil, and Samuel Weinbach. 2021. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch.

Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint*. ArXiv:1908.10063 [cs].

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv preprint*. ArXiv:1903.10676 [cs].

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. ArXiv:1810.04805 [cs].

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv preprint*. ArXiv:2004.10964 [cs].

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large Language Models in Finance: A Survey. *arXiv preprint*. ArXiv:2311.10723 [cs, q-fin].

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. ArXiv:1907.11692 [cs].

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. volume 5, pages 4513–4519. ISSN: 1045-0823.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. 2023. BBT-Fin: Comprehensive Construction of Chinese Financial Domain Pre-trained Language Model, Corpus and Benchmark. *arXiv preprint*. ArXiv:2302.09432 [cs].

Lu Lu, Jinghang Gu, and Chu-Ren Huang. 2022. Inclusion in csr reports: The lens from a data-driven machine learning model. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 46–51.

Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. 2020. Analyzing sustainability reports using natural language processing. *arXiv preprint arXiv:2011.08073*.

Tim Nugent, Nicole Stelea, and Jochen L. Leidner. 2020. Detecting ESG topics using domain-specific language models and data augmentation approaches. *arXiv preprint*. ArXiv:2010.08319 [cs].

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Purver, Matej Martinc, Riste Ichev, Igor Lončarski, Katarina Sitar Šuštar, Aljoša Valentinčič, and Senja Pollak. 2022. Tracking changes in esg representation: Initial investigations in uk annual reports. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 9–14.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Zacharias Sautner, Laurence Van Lent, Grigory Vilkov, and Ruishen Zhang. 2023. Firm-level climate change exposure. *The Journal of Finance*, 78(3):1449–1498.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Bridging the Gap in ESG Measurement: Using NLP to Quantify Environmental, Social, and Governance Communication.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2022. A dataset for detecting real-world environmental claims. *Center for Law & Economics Working Paper Series*, 2022(07).

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Francesco S. Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2021. ClimaText: A Dataset for Climate Change Topic Detection. *arXiv preprint*. ArXiv:2012.00483 [cs].

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. FinGPT: Instruction Tuning Benchmark for Open-Source Large Language Models in Financial Datasets. *arXiv preprint*. ArXiv:2310.04793 [cs, q-fin].

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. ArXiv:2303.17564 [cs, q-fin].

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance. *arXiv preprint*. ArXiv:2306.05443 [cs].

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. FinBERT: A Pretrained Language Model for Financial Communications. *arXiv preprint*. ArXiv:2006.08097 [cs].

Yuwei Yin, Yazheng Yang, Jian Yang, and Qi Liu. 2023. FinPT: Financial Risk Prediction with Profile Tuning on Pretrained Foundation Models. *arXiv preprint*. ArXiv:2308.00065 [cs, q-fin].

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2023. XuanYuan 2.0: A Large Chinese Financial Chat Model with Hundreds of Billions Parameters. *arXiv preprint*. ArXiv:2305.12002 [cs].

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Related Work

ESG Related NLP The exploration of textual data in ESG reports has seen a marked increase in interest, covering various research topics. Recent studies have expanded beyond traditional analyses by adopting machine learning models to address societal issues such as stereotypes and inclusivity (Lu et al., 2022). Furthermore, diachronic distributional techniques have been utilized to trace the evolution of ESG terminology, revealing shifts in discourse (Purver et al., 2022). Traditional research often employs keyword-based analysis methods (Sautner et al., 2023), which lack contextual sensitivity (Varini et al., 2021). Recent shifts toward context-aware machine learning models have improved performance in diverse tasks such as climate content classification (Webersinke et al., 2021), topic detection (Varini et al., 2021), Q&A systems (Luccioni et al., 2020), and claim detection and verification (Stammbach et al., 2022). Deploying fine-tuned BERT models, especially those trained on extensive business and financial news corpora like the Reuters News Archive, has effectively identified ESG controversies (Nugent et al., 2020).

Pre-trained Language Models The advent of robust Pre-trained Language Models (PLMs) such as BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), RoBERTa (Liu et al., 2019) has significantly boosted NLP task performance across diverse domains. While domain-specific pre-training further augments their performance in specialized fields (Gururangan et al., 2020), with dedicated models like BioBERT (Lee et al., 2020) for biomedicine, ClinicalBERT (Alsentzer et al., 2019) for clinical care, and SciBERT (Beltagy et al., 2019) for scientific texts demonstrating targeted advancements. Additionally, ClimateBERT (Bingler et al., 2022) specifically addresses climate risk assessment. The landscape of Large Language Models (LLMs) encompasses models like T5 (Raffel et al., 2020), and the OpenAI GPT series, beginning with GPT-3 (Brown et al., 2020), renowned for setting benchmarks in generative tasks. Other notable GPT-style models include PaLM (Chowdhery et al., 2023), and GPT-NeoX (Andonian et al., 2021), alongside GLM (Du et al., 2021). Despite many LLMs being proprietary, open-source models like OPT (Zhang et al., 2022) and LLaMA (Touvron et al., 2023) foster extensive research and practical applications. Despite these advances, the application of PLMs in the nuanced ESG domain remains nascent. Our

work seeks to bridge this gap, leveraging PLMs to enhance ESG analysis and categorization.

Financial Language Models The application of language models in finance is rapidly expanding, as these models are increasingly used for specialized functions such as risk assessment and information extraction (Li et al., 2023). For instance, BloombergGPT (Wu et al., 2023) was initially trained with a mix of general and finance-specific datasets using BLOOM176B, while Xuan Yuan 2.0 (Zhang et al., 2023) and Fin-T5 (Lu et al., 2023) focus on the Chinese financial market, leveraging specialized pre-training. Fine-tuning for financial models predominantly targets sentiment analysis, news categorization, question-answering, summarization, and entity recognition. Noteworthy adaptations include FinBERT (Liu et al., 2020; Yang et al., 2020; Araci, 2019; Huang et al., 2023). Emerging models like PIXIU (Xie et al., 2023), and FinGPT (Yin et al., 2023) exemplify the advanced application of LLaMA architectures tailored for financial tasks. Unlike previous work, we fine-tuned LLMs to address ESG classification in finance and conduct a comprehensive analysis.

B Details of Collected Data

Below are the descriptions of the datasets we collected:

- *ESG-Prospectus-Clarity-Category*[9]: This dataset comprising 1,155 entries categorized into four ESG language classes: Specific, Ambiguous, Generic, and Risk. These entries were systematically extracted from the "Principal Investment Strategy" sections of sustainable (ESG) fund prospectuses through a specialized data extraction pipeline.

- *Esg-sentiment*[10]: Featuring text across nine emotion classes within the ESG spectrum (<*Environmental, Social, Governance*> * <*Negative, Neutral, Positive*>), each emotion assigns binary labels (0/1).

- *ESGBERT base-data*[11]: This dataset extracted 13,846,000 sentences from annual reports (13,079,890 sentences), responsibility reports (695,631 sentences), sustainable reports (259,163 sentences) and articles (143,289 sentences).

- *Environmental_claims*[12]: This dataset focuses on the binary classification of environmental claims made by publicly listed companies, containing 2,647 entries. It is designed to detect real-world environmental assertions.

- *DAX ESG Media Dataset* [13]: Comprising approximately 11k recent English language ESG documents (text is document level) related to German DAX companies, this dataset includes both company issued reports and third party data, alongside an auxiliary file detailing the Sustainable Development Goals (SDGs).

- *CLIMATE-FEVER* [14]: This dataset consists of 1,535 real-world climate change claims. Each claim is supported by five Wikipedia-sourced evidence sentences annotated to either support, refute, resulting in a total of 7,675 claim-evidence pairs.

Our data extraction involved the retrieval of the `'text'` field across datasets, except the *DAX ESG Media Dataset*, from which the `'content'` field was extracted, and the *CLIMATE-FEVER*, where both the `'claim'` and the `'evidence'` fields within the `'evidences'` array were extracted. The summary of datasets is shown in Table 6.

C ESG Keywords

All keywords we used shown in Table 7 refer to (Schimanski et al., 2023).

D Word Clouds of keyword search

After filtering the texts by keywords searching. The texts are categorized into Environmental (Env), Social (Soc), Governance (Gov), and Non-ESG groups. The word clouds generated from these texts shown in Figure 6 offer a visual representation of the predominant themes within each category. In the Environmental domain, the word cloud prominently features terms such as "GHG emission" and "climate change," highlighting the focus on environmental impact. Socially oriented texts

[9] https://huggingface.co/Abhijeet3922

[10] https://huggingface.co/datasets/TrajanovRisto/esg-sentiment

[11] https://huggingface.co/datasets/ESGBERT/base_data

[12] https://huggingface.co/datasets/climatebert/environmental_claims

[13] https://www.kaggle.com/datasets/equintel/dax-esg-media-dataset

[14] https://www.sustainablefinance.uzh.ch/en/research/climate-fever.html

Table 6: Summary of Collected ESG-Related Datasets

Dataset Name	Content Format	Size
ESG-Prospectus-Clarity-Category	<Text, Label>	2310 rows (546 kB)
Esg-sentiment	<Text, Environmental Negative,...,Social Positive>	679 rows (80.1 kB)
ESGBERT base-data	<Text>	13,846,000 rows (2.33 GB)
Environmental_claims	<text, label>	2647 rows (272 kB)
DAX ESG Media	<company, content, datatype, data, domain, esg_topics, internal, symbol, title>	11455 rows (130.11 MB)
CLIMATE-FEVER	<claim_id, claim, claim_label, evidences>	1,535 rows (3 MB)

Table 7: ESG Keywords Across Domains

Domain	Keywords
Environmental	adaptation, agricultural, air quality, biodiversity, biomass, climate, CO2, conservation, consumption, diversity, ecosystem, emissions, energy, environmental, flood, forest, fossil fuel, GHG, global warming, green, greenhouse, land use, methane, mitigation, nature, ozone, pollution, renewable, soil, solar, sustainability, water, recycling, clean energy, natural
Social	age, culture, race, accessibility, accident, accountability, awareness, charity, community, consumer protection, cyber security, data privacy, discrimination, diversity, education, employee benefit, empowerment, equality, ethics, fairness, gender, health, inclusion, mental well-being, parity, privacy, quality of life, religion, safety, social impact, volunteerism, welfare, wellbeing, workforce
Governance	audit, authority, bribery, compliance, corporate governance, corruption, crisis management, due diligence, ethics, framework, integrity, legal, lobby, oversight, policy, regulation, reporting, risk management, stakeholder engagement, transparency, whistleblower, board diversity, executive pay, shareholder rights, sustainable governance, corporate transparency, anti-corruption, business ethics

are characterized by frequent mentions of "human rights," "product," and "customer," reflecting the emphasis on societal concerns and stakeholder welfare. In the Governance category, words like "director," "financial statement," "management," and "shareholder" dominate, aligning with expectations for governance-related discourse. These visual insights from the word clouds roughly correspond with our anticipated high-frequency words for each ESG classification, underscoring the effectiveness of our keyword-based filtering approach. we got the preliminary results shown in Table 8.

Table 8: Summary of Processed Data

Domain	Num. of Sent.	Avg. Num. of Words		
		Q1	Mean	Q3
Env	2,143,453	19	30.43	36
Soc	2,796,077	20	31.46	37
Gov	1,851,303	20	31.75	38
Non-ESG	11,392,832	-	-	-
Total	18,183,665	-	-	-

E Data Distribution

Four-class and nine-class categorization criteria defined by Huang et al. (Huang et al., 2023)

(a) Environmental (b) Social (c) Governance

Figure 6: ESG Domain Word Clouds After Keywords Search

Pre-training Dataset. We performed a 90-10 train-eval split to create the training and evaluation datasets, as shown in Table 9.

Table 9: Pre-training Dataset Statistics

Dataset	Num. of Sentences
Train	5,257,347
Valid	584,150
Total	5,841,497

For the four-class dataset. We used a 70:15:15 splitting ratio to construct the train-dev-test sets. The training set consisted of 37,155 instances, with 10,144 'Soc', 9,799 'Non-ESG', 9,192 'Env', and 8,020 'Gov'. The validation and test set each contained 7,962 instances, with 2,174 'Soc', 2,100 'Non-ESG', 1,969 'Env', and 1,719 'Gov' for validation, and 2,174 'Soc', 2,100 'Non-ESG', 1,970 'Env', and 1,718 'Gov' for testing. Results are shown in Figure 7.

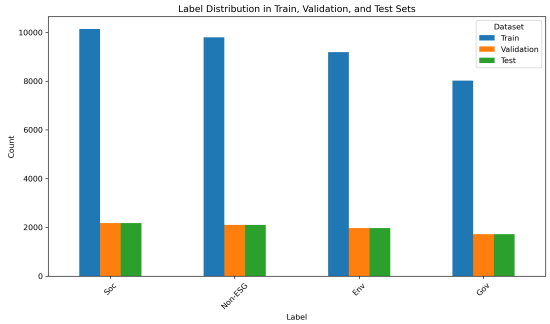


Figure 7: Four-class Label Distribution in Train, Val, Test Sets

For the nine-class dataset. We applied an 81:09:10 splitting ratio. The training set had 17,419 instances, with each label ('Human Capital', 'Product Liability', 'Pollution and Waste', 'Business Ethics and Values', 'Corporate Governance', 'Community Relations', 'Non-ESG', 'Climate Change', 'Natural Capital'). The validation set contained 2,151 instances. Similarly, the test set had 1,936 instances. These datasets were constructed using stratified sampling to ensure a balanced representation of each class in the train-dev-test splits. Lastly, we fine-tuned our pre-trained models on these two datasets to adapt them for the four-class and nine-class ESG text classification tasks. Results are shown in Figure 8.

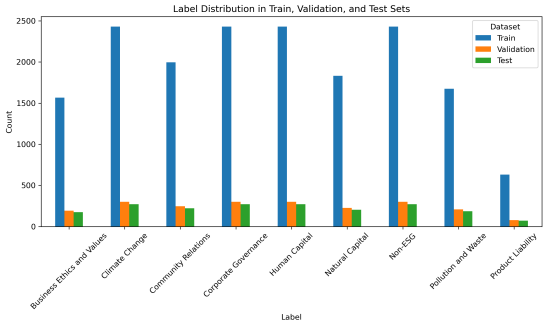


Figure 8: Nine-class Label Distribution in Train, Val, Test Sets

F LLM Labeling Prompts Design

We primarily utilize a combination of **few-shot learning** and Chain of Thought (**CoT**) in prompts design. Few-shot learning enables the model to learn from a limited quantity of text to align the acquired knowledge with our specific purpose. CoT (Wei et al., 2022) is a reasoning strategy that involves breaking down a problem into sub-problems and connecting them in a specific logical order based on a chain structure. The purpose of using a few shots is to familiarize the model with the ESG classification strategy using a small sample. Using

CoT is intended to enhance the model's reasoning process.

For task 1: *Classify whether the text is high-quality ESG data: Yes or No.* The {Criteria} will be replaced by certain criteria, which are generated by GPT-4, and {Data} will be replaced by certain text we want to be classified.

```
System Prompt: "You are a helpful
assistant in data managing, and good at
using high-quality data criteria"
User Prompt: "To identify high-quality
ESG data, we should consider the
following criteria: {Criteria}
The following sentence is the data
needed to define: {Data}
Answer 'Yes' or 'No' first,
then give an explanation.
Let's think step by step."
```

For task 2: *Classify whether the text is Env/Soc/-Gov data: Yes or No.*

```
System Prompt: "You are an expert in ESG
data classification, especially {Type}
ESG data classification."
User Prompt: "To identify {Type} ESG data,
we should consider the
following criteria: {Criteria}
Answer 'Yes' or 'No' first,
then give an explanation.
Demonstrations:
{Demonstrations}
Text: {Data}
Answer: Let's think step by step."
```

For task 3: *Assign texts to one of the following labels based on their domain.* The {Classes} are corresponding to {Type}.

```
System Prompt: "You are an expert in
classifying {Type} ESG data into
finer categories such as {Classes}."
User Prompt: "For {Type} ESG data
classification, consider: {Criteria}
First assign a class label based on
these categories, or 'other' if
uncertain. Then give an explanation.
Demonstrations:
{Demonstrations}
Text: {Data}
Answer: Let's think step by step."
```

F.1 Annotation Examples

To illustrate the LLM's responses to each task, we use text *"The possible threat of natural disasters affecting the ability to trade."* as an example, the responses are shown in Table 10.

G CPT PLMs Details

The continue pre-training parameters for our models were set as follows: a batch size of 16 for both training and evaluation, a total of 25 epochs, a seed value of 42 for reproducibility, and a maximum sequence length of 512. The models were trained with a gradient accumulation step of 2, weight decay of 0.01, and 16-bit precision (fp16) to optimize memory usage and accelerate the training process. The model checkpoints were saved and evaluated at the end of each epoch, as specified by the evaluation_strategy and save_strategy parameters set to 'epoch'. We conducted experiments on BERT, DistilRoBERTa, and RoBERTa at three different stages: epoch 5, epoch 15, and epoch 25. The results indicated that the models performed best at epoch 25. We utilized `torch.distributed.launch` for single-machine multi-GPU training. The pretraining of BERT, DistilRoBERTa, and RoBERTa took approximately 27 hours, 23.5 hours, and 33 hours, respectively. The training loss, validation loss, and validation accuracy during the training process are visualized in Figure 9. The smooth decrease in the loss curve and the gentle rise in the accuracy curve suggest that the models were converging and the training process was proceeding as expected. Consequently, we selected the model with the smallest

Table 10: Example of Data Through Three Tasks

Item	Content
Text	The possible threat of natural disasters affecting the ability to trade.
Task 1 Response	Yes. 1. Relevance: The sentence is directly related to environmental issues, specifically the potential impact of natural disasters on trade. 2. Accuracy: It is difficult to determine the accuracy of this statement without additional context or information about the specific location and type of natural disaster being referred to. However, in general, natural disasters can disrupt trade and supply chains, so the statement is plausible. 3. Source Credibility: Without knowing the source of this statement, it is difficult to assess its credibility. 4. Specificity: The statement provides a general idea of the potential threat posed by natural disasters to trade, but does not provide detailed insights or examples. 5. Objectivity: The statement appears to be objective and presents a balanced view of the potential impact of natural disasters on trade.
Task 2 Response	Answer: Yes, the possible threat of natural disasters affecting the ability to trade is an environmental factor. Natural disasters can have a significant impact on the environment, including infrastructure, supply chains, and the ability to conduct business operations.
Task 3 Response	Label: Climate Change. Explanation: The text mentions the possible threat of natural disasters affecting the ability to trade. This indicates a concern related to climate change, as natural disasters such as hurricanes, floods, and wildfires are often associated with climate-related events. Climate change can disrupt trade by damaging infrastructure, disrupting supply chains, and impacting transportation systems. Therefore, the text falls under the category of Climate Change.

validation loss at epoch 25 as our final pretraining models: `ESG-BERT`, `ESG-DistilRoBERTa`, and `ESG-RoBERTa`.

H FinLlama Training Details

Datasets. This fine-tuning was conducted on a specialized instruction-tuning dataset on financial domain delineated in FinGPT (Wang et al., 2023). Furthermore, we enhanced the dataset by incorporating the ESG_Chat dataset, which consists of dialogues between humans and Large Language Models (LLMs) focusing on methodologies to improve ESG scores. These conversations are structured to provide step-by-step guidance, with the LLM responses specifically tailored to offer structured, actionable advice. The characteristics of these datasets are detailed in Table 11.

Hyperparameters. Each stage was meticulously conducted throughout the training regimen over 3 epochs to ensure the models' robust assimilation of the task-specific nuances. A consistent set of hyperparameters characterized the training to maintain uniformity across the models. Specifically, the batch size per device was set to 4, coupled with a gradient accumulation strategy involving four steps. This setup facilitated optimal resource utilization and stable training dynamics. The learning rate scheduler employed was of the cosine type, which aided in gradual learning rate adjustments, contributing to smoother convergence. For monitoring and model checkpointing, logging intervals were established at every 10 steps, and model states were preserved at every 100 steps, ensuring detailed progress tracking and the ability to revert to the most effective model state. The learning rate was judiciously chosen as 5×10^{-5}, balancing rapid adaptation and the preservation of prelearned representations. The training progression

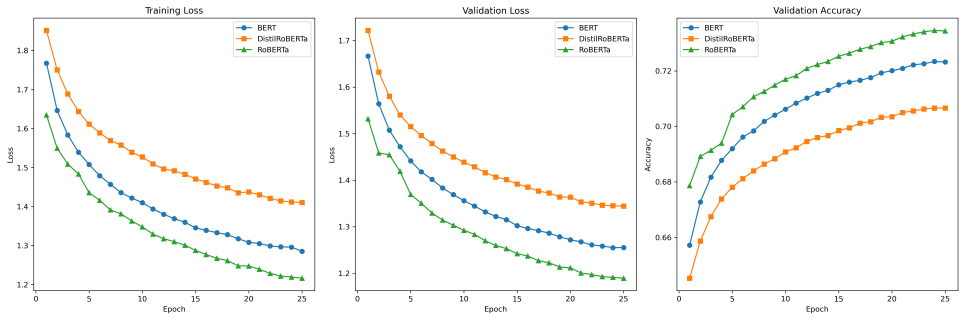


Figure 9: Continue Pre-training Log Loss and Accuracy across epochs

for both models was visually documented through loss curves, providing insightful glimpses into the models' learning trajectories. Notably, a significant loss reduction was observed after the initial epoch for both models, indicative of their swift adaptation to the training objectives. For ESGLlama, the training culminated with the loss stabilizing around 0.4, shown in Figure 10a, suggesting effective learning. Conversely, FinLlama exhibited a distinct two-phase training dynamic; the initial pretraining phase concluded with a loss of around 2.4, shown in Figure 10b, which, upon undergoing the subsequent Supervised Fine-Tuning (SFT) phase, settled at approximately 1.15 shown in Figure 10c. This delineation in training phases for FinLlama underscores the layered approach to model refinement, first broadening its financial domain comprehension, followed by targeted instruction-based fine-tuning to hone its capabilities for specific financial tasks. These models will be tested on our labeled ESG classification data. All experiments were conducted on NVIDIA V100 Tensor Core GPUs. Due to LLMs' substantial parameter size and complex structure, fine-tuning and inference can be particularly time-intensive. We employed Parameter-Efficient Fine-Tuning (PEFT) techniques such as Low-Rank Adaptation (LoRA) and freeze during SFT phases to enhance efficiency. Additionally, we utilized LLaMA-Factory (Zheng et al., 2024) framework and vLLM (Kwon et al., 2023) to accelerate pre-training SFT and inference processes.

I ESG Classification SFT Dataset

Format: [{"instruction": "...", "input": "...", "output": "..."}]
Four-class Classification: **instruction**: Classify the following text into one of the four ESG categories, choose an answer from {Env/Soc/Gov/Non-ESG}. **input**: We maintain a health and safety management system aligned to ISO legal requirements in Australia and New Zealand. **output**: Soc
Nine-class Classification: **instruction**: Classify the following text into one of the nine ESG categories, choose an answer from {Climate Change/Natural Capital/Pollution and Waste/Human Capital/Product Liability/Community Relations/-Corporate Governance/Business Ethics and Values/Non-ESG}. **input**: Grievance mechanisms forms an important part of our stakeholder engagement process, and our human rights policy states that we will provide, or cooperate in providing, appropriate remediation if we have caused or contributed to adverse human rights impacts. **output**: Human Capital

Table 12: An exam. of ESG classification SFT dataset.

The dataset we used for supervised fine-tuning is constructed from ESG SFT data in Section 2.3. The ESG classification SFT data was sampled and reconstructed from ESG SFT data by only selecting classification data and simplifying the result by retaining the text label without any additional ex-

[1] https://huggingface.co/datasets/FinGPT/fingpt-sentiment-train

[2] https://huggingface.co/datasets/FinGPT/fingpt-finred

[3] https://huggingface.co/datasets/FinGPT/fingpt-headline

[4] https://huggingface.co/datasets/FinGPT/fingpt-ner

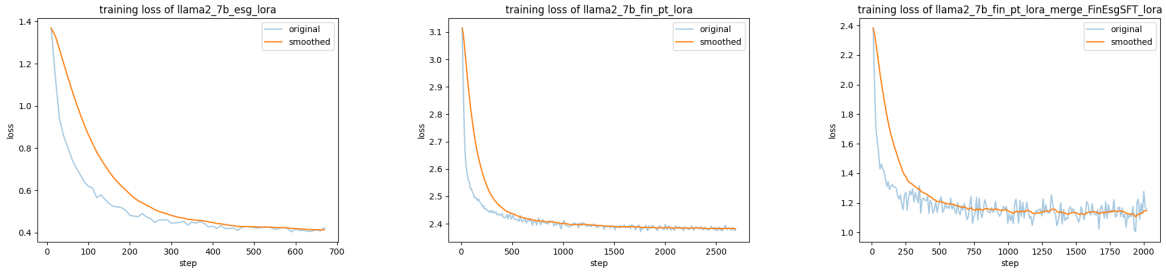
[5] https://huggingface.co/datasets/FinGPT/fingpt-fiqa_qa

[6] https://huggingface.co/datasets/FinGPT/fingpt-fineval

[7] https://huggingface.co/datasets/zadhart/ESG_Chat

Table 11: Instruction Financial Dataset Overview

Datasets	Train Rows	Test Rows	Description
fingpt-sentiment-train[1]	76.8K	N/A	Sentiment Analysis Training Instructions
fingpt-finred[2]	27.6K	5.11K	Financial Relation Extraction Instructions
fingpt-headline[3]	82.2K	20.5K	Financial Headline Analysis Instructions
fingpt-ner[4]	511	98	Financial Named-Entity Recognition Instructions
fingpt-fiqa_qa[5]	17.1K	N/A	Financial Q&A Instructions
fingpt-fineval[6]	1.06K	265	Chinese Multiple-Choice Questions Instructions
ESG_Chat[7]	914	N/A	Chat History about Improve ESG Score step-by-step



(a) ESGLlama Training Loss (b) FinLlama Training Loss S1 (c) FinLlama Training Loss S2

Figure 10: Training loss analysis during each stage of fine-tuning

planations. There are two main classification tasks contained in this dataset: four-class classification and nine-class classification. Finally, we obtained approximately 24k ESG Classification SFT Data. An example of the ESG classification SFT dataset regarding these two tasks is shown in Table 12. Using this dataset, we aim to enhance the baseline's ESG text classification capability. This is because the baseline's pre-training data may include financial text data partially related to ESG. We intend to modify the baseline for this task and evaluate its performance during the SFT training phase.

J Classification Prompts

```
System Prompt: "You are an expert in
classifying ESG data. You will start
your response with 'Label:'."
User Prompt: "Classify the following
text into one of the four ESG
categories, choose an answer from
{Categories}
Demonstrations:
{Demonstrations}
Text: {Text}
Label: So, the answer is"
```

For Four-Class classification task, we should specify the `{Categories}` by:

```
{Env/Soc/Gov/Non-ESG}
```

For Nine-Class classification task, we should

specify the {Categories} by:

```
{Climate Change/Natural Capital/
Pollution and Waste/Human Capital/
Product Liability/Community Relations/
Corporate Governance/Non-ESG
Business Ethics and Values}
```

To employ a chain-of-thought (CoT) setting, we need to slightly modify the system prompt and add let's think step by step at the end of the user prompt:

```
System Prompt: "You are an expert in
classifying ESG data. You will
response in this format:
'Label:xxx. Explanation:xxx'.
Your responses should be
precise and concise."
User Prompt: "...
Label: Let's think step by step.
So, the answer is"
```

K Additional LLM Classification Analysis

For Four-class classification. In evaluating our models, ESGLlama and FinLlama, within our experimental framework, we employed Precision, Recall, F1 Score, and Accuracy as our performance metrics. Initially, let us delve into the precision aspect, which serves to illustrate the models' exactness in classification tasks. Through the analysis of precision scores and the accompanying graphical representations shown in Figure 4, it becomes evident that both ESGLlama and FinLlama surpass the baseline model across most experimental configurations. Furthermore, even the baseline model, when subjected to Supervised Fine-Tuning (SFT) using our constructed ESG classification dataset, demonstrates enhanced performance compared to its original state. Interestingly, the Freeze fine-tuning approach generally outperforms the LoRA method, except in zero-shot settings. This observation could be attributed to the Freeze technique requiring a broader range of parameters for fine-tuning, thereby facilitating a deeper understanding of downstream tasks. In contrast, LoRA's superior performance in zero-shot scenarios might hint at a slight overfitting issue; external demonstration examples, not included in the training set, could potentially disrupt the model's inference processes. The Freeze approach, in this context, better preserves the model's generalization capabilities and intrinsic reasoning faculties.

The incorporation of Chain of Thought (CoT) prompts leads to a performance decline in zero-shot and one-shot settings, except for the Iterated Chain of Learning (ICL) tasks. This decline could stem from the absence of stepwise reasoning chains in our training data, coupled with the inherent incompatibility of the CoT methodology with classification tasks—CoT primarily suits logic-based problem-solving. Nevertheless, the addition of demonstrations in ICL tasks enriches the model's learning of classification logic through increased sample exposure, culminating in the most favorable outcomes under ICL CoT configurations.

Further examination of performance metrics, as detailed in the corresponding table shown in Table 13, reveals that the LoRA method, applied directly to the baseline on our ESG classification dataset, achieves the highest precision (0.6928), recall (0.5557), F1 score (0.5488), and accuracy (0.5697) in zero-shot tasks. This outcome not only underscores the constructed dataset's validity but also establishes a benchmark for subsequent comparisons. Furthermore, the bold formatting in the table highlights the highest precision scores across six method settings for each model, underscoring the best-performing configurations. The underlined values denote the top performance metrics across all models and settings, establishing a benchmark for comparison. The star symbol (*) identifies the best baseline result for the LoRA and Freeze fine-tuning methods, serving as a reference point for assessing the fine-tuned models' enhancements. The directional arrows (↑↓) provide a visual cue for performance fluctuations in comparison to the baseline, elucidating the impact of our fine-tuning strategies on model precision. Against this backdrop, both ESGLlama and FinLlama exhibit a decline, albeit still outperforming the baseline, especially in ICL settings. Notably, FinLlama achieves superior precision over ESGLlama with the addition of CoT, underscoring the nuanced impact of our training methodologies on model performance. In summary, the table elucidates the nuanced interplay between fine-tuning methodologies, the inclusion of CoT prompts, and the iterative learning approach on model precision. The discernible improvement in precision with ESGLlama and FinLlama, particularly in ICL settings, reaffirms the efficacy of our fine-tuning strategies in embedding ESG-specific

knowledge into large language models.

For Nine-class classification, the analysis of performance metrics, particularly precision, elucidates a notable trend: as the complexity and diversity of classification categories increase, the task inherently becomes more challenging, as evidenced by the overall diminished performance compared to the four-class scenario. This trend underscores the escalated difficulty in distinguishing among a greater number of classes.

The precision score visualization (Figure 5) demonstrates that both ESGLlama and FinLlama significantly outperform the baseline model across most methodological settings. This superiority highlights our fine-tuned models' enhanced understanding and classification capability in the context of ESG-related texts. FinLlama demonstrates superior proficiency in iterative contrastive learning (ICL), particularly in scenarios with increased sample availability, indicating a profound comprehension of financial texts and their nuances. The analysis further reveals a pronounced decrement in performance for the one-shot learning setting across more granular classification tasks. Providing only one example per class introduces considerable bias and may confound the model's judgment due to the high variance associated with minimal data. Conversely, enriching the model with a broader set of examples significantly ameliorates performance, aligning with the expected benefits of expanded training data. This intricate classification landscape observes a notable divergence in the efficacy of the LoRA and Freeze fine-tuning methods. Interestingly, The LoRA approach exhibits superior performance in the one-shot setting compared to Freeze, suggesting that LoRA's parameter adaptation might be more conducive to effectively assimilating sparse class-specific information.

Delving deeper into the details presented in the accompanying Table 14, the most commendable performance is attributed to FinLlama under the ICL with Chain of Thought (CoT) augmentation, achieving a precision score of 0.6654. This result significantly surpasses the baseline precision of 0.6164 and even outstrips the baseline model fine-tuned with LoRA on the ESG classification data, which scored 0.6544. This evidence conclusively demonstrates the potent efficacy of FinLlama, particularly when augmented with CoT in complex classification scenarios.

Model	Methods	Overall			
		Precision	**Recall**	**F1 Score**	**Accuracy**
Llama2	Zero Shot	0.5778	0.5025	0.4815	0.5093
	w/ CoT	0.5527	0.4613	0.4252	0.4776
	One Shot	0.6012	0.5056	0.4706	0.5109
	w/ CoT	0.5370	0.3767	0.2680	0.3931
	ICL	**0.6687**	**0.5408**	**0.5077**	**0.5446**
	w/ CoT	0.6794	0.5193	0.4803	0.5229
LoRA	Zero Shot	**0.6928***	**0.5557***	**0.5488***	**0.5697***
	w/ CoT	0.6381	0.4973	0.5128	0.5053
	One Shot	0.5265	0.3896	0.2924	0.3976
	w/ CoT	0.5646	0.3291	0.2442	0.3360
	ICL	0.6148	0.5157	0.4821	0.5232
	w/ CoT	0.6213	0.3971	0.3247	0.4019
Freeze	Zero Shot	0.5741	0.5000	0.4787	0.5068
	w/ CoT	0.5480	0.4613	0.4276	0.4775
	One Shot	0.6085	0.5113	0.4761	0.5168
	w/ CoT	0.6168	0.3932	0.2873	0.4073
	ICL	0.6611	**0.5382**	**0.5036**	**0.5422**
	w/ CoT	**0.6749**	0.5181	0.4767	0.5216
ESGLlama	Zero Shot	0.5770	0.4997	0.4768	0.5054
	w/ CoT	0.5502	0.4594	0.4205	0.4753
	One Shot	0.6106	0.5373	0.5140	0.5389
	w/ CoT	0.6064	0.3984	0.3128	0.4147
	ICL	0.6738	**0.5508**↓	**0.5203**↓	**0.5548**↓
	w/ CoT	**0.6746**↓	0.4882	0.4323	0.4935
FinLlama	Zero Shot	0.5766	0.4961	0.4745	0.5024
	w/ CoT	0.5665	0.4669	0.4297	0.4828
	One Shot	0.6139	0.5375	0.5139	0.5394
	w/ CoT	0.5724	0.3856	0.3011	0.4017
	ICL	0.6698	**0.5497**↓	**0.5174**↓	**0.5535**↓
	w/ CoT	**0.6797**↓	0.4917	0.4365	0.4971

Table 13: Four-class evaluation results compare with baseline and our fine-tuned LLMs. **Blod** shows the best results in six method settings according to each model, and underline illustrates the best performance in each column. Star (*) is the best baseline result for two fine-tuning methods (LoRA and Freeze). Arrow (↑↓) signifies performance compared with Star (*).

Model	Methods	Overall			
		Precision	**Recall**	**F1 Score**	**Accuracy**
Llama2	Zero Shot	0.5875	0.4404	0.4454	0.4886
	w/ CoT	0.5826	0.4106	0.4171	0.4654
	One Shot	0.5049	0.4322	0.3877	0.4737
	w/ CoT	0.4314	0.3556	0.2895	0.3838
	ICL	0.6108	0.4029	0.4017	0.4411
	w/ CoT	**0.6164**	**0.4624**	**0.4932**	**0.5057**
LoRA	Zero Shot	0.5681	0.4901	0.4759	**0.5294***
	w/ CoT	0.5180	0.4112	0.3895	0.4473
	One Shot	0.6256	**0.5347***	**0.4795***	0.5186
	w/ CoT	0.5751	0.3915	0.3450	0.3972
	ICL	0.6242	0.1946	0.1340	0.2257
	w/ CoT	**0.6544***	0.1834	0.1465	0.2123
Freeze	Zero Shot	0.5911	**0.4458**	**0.4488**	**0.4974**
	w/ CoT	0.5799	0.4122	0.4161	0.4664
	One Shot	0.5258	0.4445	0.4148	0.4866
	w/ CoT	0.4922	0.4005	0.3353	0.4323
	ICL	**0.6285**	0.4189	0.4265	0.4649
	w/ CoT	0.5719	0.2432	0.2337	0.2862
ESGLlama	Zero Shot	0.5866	0.4271	**0.4340**↓	0.4778
	w/ CoT	0.5914	0.4190	0.4258	0.4726
	One Shot	0.5138	**0.4446**↓	0.4136	**0.4855**↓
	w/ CoT	0.4785	0.4031	0.3373	0.4318
	ICL	**0.6201**↓	0.4143	0.4235	0.4576
	w/ CoT	0.5773	0.2533	0.2470	0.2965
FinLlama	Zero Shot	0.5608	0.4293	**0.4301**↓	**0.4830**↓
	w/ CoT	0.5750	0.4123	0.4164	0.4664
	One Shot	0.5219	**0.4376**↓	0.4069	0.4757
	w/ CoT	0.4886	0.4062	0.3399	0.4349
	ICL	0.6168	0.4127	0.4163	0.4638
	w/ CoT	**0.6654**↑	0.2504	0.2478	0.2908

Table 14: Nine-class evaluation results compare with baseline and our fine-tuned LLMs. **Bold** shows the best results in six method settings according to each model, and underline illustrates the best performance in each column. Star (*) is the best baseline result for two fine-tuning methods (LoRA and Freeze). Arrow (↑↓) signifies performance compared with Star (*).