Tilburg University

BENCHMARKING ESG TEXT CLASSIFICATION IN LOW RESOURCE SETTINGS

A COMPARATIVE STUDY OF QLORA FINE-TUNING, FEW-SHOT LEARNING, AND ZERO-SHOT ENSEMBLE

POROSHAT GHASHGHAEI

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2131091

COMMITTEE

dr. Samaneh Khoshrou
MSc. Matthijs de Koning

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

May 19th, 2025

WORD COUNT

8746

ACKNOWLEDGMENTS

I would like to thank my supervisor dr. Samaneh Khoshrou for her guidance and valuable feedback throughout this research. I also thank my second reader MSc. Matthijs de Koning for the time and effort they dedicated to reviewing my work. Special thanks to Hamid Madani, former Data Science student, for his insightful input, which helped shape my research. I'm also deeply grateful to my friends and family for their constant support and encouragement.

BENCHMARKING ESG TEXT CLASSIFICATION IN LOW RESOURCE SETTINGS

A COMPARATIVE STUDY OF QLORA FINE-TUNING, FEW-SHOT LEARNING, AND ZERO-SHOT ENSEMBLE

POROSHAT GHASHGHAEI

Abstract

Automated classification of Environmental, Social, and Governance (ESG) disclosures is critical for promoting corporate transparency and combating greenwashing. However, developing reliable ESG classifiers is difficult due to the scarcity of annotated data and the high resource demands of fine-tuning large language models (LLMs). While ESG-specific models like ESG-BERT, ESG-RoBERTa, and ESG-LLaMA have shown promise in supervised settings, their performance under realistic low-resource conditions remains underexplored.

This research benchmarks three scalable strategies for ESG sentence classification with minimal labeled data: (1) QLoRA parameter-efficient fine-tuning of ESG-LLaMA, (2) a novel zero-shot ensemble of FinBERT-ESG, ESG-RoBERTa, and DeBERTa-v3-MNLI, and (3) few-shot prompting with DeepSeek-LLM-7B-Chat. Additionally, we evaluate ESG-LLaMA in a pure zero-shot setting to assess its domain-specific generalization capabilities. All models are evaluated on the same held-out test set of SASB-aligned ESG sentences using balanced accuracy and macro F1-score.

QLoRA fine-tuning achieved the strongest results (F1 = 0.885; accuracy = 0.887), particularly improving classification of Governance texts. The zero-shot ensemble (F1 = 0.698) and few-shot DeepSeek (F1 = 0.690) offered competitive alternatives without training. These findings show that effective ESG classification is achievable even with limited resources.

By unifying and comparing these approaches, this research contributes a comprehensive evaluation of low-resource ESG classification strategies. The findings support more scalable and accessible ESG analysis pipelines, especially in settings where labeled data and compute resources are limited—thereby advancing the societal goal of sustainable, transparent, and accountable finance.

1 SOURCE/CODE/ETHICS/TECHNOLOGY STATEMENT EXAMPLE

The "SASB-aligned ESG sentence" was obtained from Kaggle (Junprung, 2023) . Since the dataset is publicly available on Kaggle, the owner implicitly provides consent for its use for non-commercial research, as per Kaggle's terms of service. All figures and plots were created by the author. Some code components were adopted from publicly available examples, including documentation and tutorials from Hugging Face and Optuna. All experiments were conducted using GPU4EDU, the university's shared computing cluster providing academic access to NVIDIA GPUs. A full list of libraries and their versions is included in Appendix C on page 42. The codes for the whole research can be found here. Code generation and debugging support were provided by ChatGPT. Grammarly and ChatGPT were used to support spelling, grammar, and overall language clarity. This thesis was written in LaTeX using Overleaf and the code executed in Visual Studio Code.

2 INTRODUCTION

The term ESG (Environmental , Social, and Governance) was popularized by a 2004 United Nations report (United Nations Global Compact, 2004), emphasizing the importance of including sustainability considerations into financial decision-making. ESG metrics help assess a company's impact on society and the environment, and are increasingly critical for responsible investing (Tan et al., 2025). For example, ESG analysis can reveal whether a business prioritizes social welfare alongside profitability, and whether it acknowledges its environmental and social responsibilities in public disclosures.

Insights derived from ESG data are especially important for responsible investing, where social and environmental concerns guide financial decisions (Alliance, 2018). However, ESG-related information in corporate disclosures often lacks structure and consistency, making it difficult for investors and regulators to assess sustainability performance reliably. Reviewing such data manually takes considerable time and may introduce personal bias, which may increase the likelihood of misinterpretation or greenwashing (Mehra et al., 2022). Greenwashing occurs when a company performs poorly in environmental practices but still promotes a misleading image of sustainability through positive communication.(de Freitas Netto et al., 2020).

The growing importance of ESG factors for promoting transparency and sustainable finance has led to a rising demand for automated ESG text classification through Machine Learning and Natural Language Processing (NLP) methods.(Sokolov, Mostovoy, et al., 2021). However, the development of reliable models is challenged by the scarcity of labeled ESG data, which limits the effectiveness of traditional supervised learning approaches.

2.1 *Scientific Relevance*

While recent studies have developed fine-tuned ESG models like ESG-BERT, ESG-RoBERTa (Xia et al., 2024) and category-specific classifiers such as Env-RoBERTa (Schimanski et al., 2024), most research evaluates these models in supervised settings against individual baselines. In contrast, this research introduces a novel approach: a zero-shot ensemble that combines multiple domain-specific and general-purpose models, namely FinBERT-ESG (Huang et al., 2022), ESG-RoBERTa (Xia et al., 2024), and DeBERTa-MNLI (He et al., 2023) for ESG text classification. So far, there has been no comprehensive evaluation comparing this type of ensemble with a domain-specific large language model like ESG-LLaMA (Xia et al., 2024) within the ESG classification domain.

This research also investigates the effectiveness of parameter-efficient fine-tuning using QLoRA (Dettmers et al., 2023), a method that significantly reduces memory usage while achieving performance comparable to full fine-tuning. Although such techniques have shown promise in domains like clinical NLP (Gema et al., 2024), they remain underexplored in the ESG context, especially when labeled data is limited.

Additionally, in low-resource settings, few-shot learning offers a practical alternative by using minimal labeled examples (Brown et al., 2020). This research tests an instruction-tuned general-purpose model, DeepSeek-LLM-7B-Chat (DeepSeek-AI, 2024), on ESG classification tasks without any prior ESG-specific fine-tuning. Its performance is compared to both the QLoRA fine-tuned ESG-LLaMA and the zero-shot ensemble, filling a clear gap in the literature.

This research fills a gap by evaluating the comparative performance of zero-shot ensemble, few-shot, and QLoRA fine-tuned models on ESG text classification; an approach not yet systematically benchmarked in the literature.

2.2 *Social Relevance*

ESG classification carries significant societal impact, extending beyond finance to shape corporate decision-making, regulatory compliance, sustainability projects, and actions addressing global issues like climate change and social inequality. Automated ESG classification can enhance corporate accountability by identifying greenwashing, aid regulatory oversight, guide ethical investment choices, and boost transparency across supply chains (Ahmad et al., 2023). To ensure reliable ESG classification despite limited labeled data, it is crucial to identify NLP models that are both effective and resource-efficient. This research therefore compares zero-shot ensemble, few-shot learning, and QLoRA fine-tuning to help stakeholders select scalable and robust solutions for automated ESG classification.

2.3 *Research Goals*

To address the research gap in ESG text classification, this research aims to answer the following main and supporting questions:

> *To what extent does fine-tuning ESG-LLaMA with QLoRA improve its performance on ESG text classification tasks compared to (a) its zero-shot performance and (b) a zero-shot ensemble of FinBERT-ESG, ESG-RoBERta, and DeBERTa-MNLI?*

A supporting research question is formulated as:

> *How effective is few-shot prompting with an instruction-tuned LLM such as DeepSeek for ESG text classification, compared to (a) zero-shot performance and (b) QLoRA-fine-tuned ESG-LLaMA?*

2.4 *Main Findings*

Among the evaluated approaches for ESG text classification, parameter-efficient fine-tuning demonstrated the highest overall effectiveness with approximately 0.89 for both macro F1-score and balanced accuracy. It outperformed all other models and, importantly, overcame previous challenges in classifying Governance sentences. Zero-shot techniques also outperformed the majority-class baseline (macro F1: 0.250; balanced accuracy: 0.333). The best performance in this category came from the ensemble of domain-specific models; FinBERT-ESG, ESG-BERT, and DeBERTa-v3-MNLI, which achieved a macro F1 of 0.698 and a balanced accuracy of 0.735, with strong results particularly in the Environment and Governance categories. Few-shot classification using DeepSeek produced moderate outcomes (macro F1: 0.690; balanced accuracy: 0.687), though it struggled to clearly distinguish between Social and Governance categories. In summary, fine-tuning provides the most stable and accurate results under data-constrained conditions, while ensemble zero-shot models deliver strong, ready-to-use performance. Few-shot prompting, though effective, showed more variation depending on the class.

3 RELATED WORK

Recent progress in Machine Learning and Natural Language Processing has led to the creation of transformer models that perform effectively in ESG text classification. A major challenge in this area is the need for large labeled datasets. Fine-tuning large language models (LLMs) typically demands high computational resources and large amounts of labeled data.

However, in many real-world applications, ESG datasets tend to be small and imbalanced, making full fine-tuning impractical (J. Lee & Kim, 2023; Prottasha et al., 2024). To address these limitations, various strategies have emerged, including domain-specific transformer models, efficient fine-tuning methods, few-shot prompting, and ensemble learning.

3.1 *Domain-Specific Transformer Models and LLMs for ESG*

Pretraining language models on domain-specific data has emerged as an effective approach in ESG classification. These methods enable the models to capture specialized terminology and context. For instance, FinBERT built on the BERT architecture and trained on financial text has demonstrated enhanced performance in ESG-related sentiment and topic classification (Araci, 2019), especially in scenarios with limited labeled data. ESG-BERT, trained on ESG-specific corpora, has similarly outperformed standard BERT when classifying financial disclosures (Mehra et al., 2022).

Xia et al. (2024) introduced multiple ESG-tuned models, including ESG-BERT, ESG-RoBERTa, ESG-DistillRoBERTa, and ESG-LLaMA. Among them, ESG-RoBERTa achieved the strongest overall performance, with an F1-score of 0.9086 and accuracy of 0.9102, outperforming other models such as FinBERT (F1 = 0.7165, accuracy = 0.7222) and ESG-BERT (F1 = 0.9071, accuracy = 0.9083). Moreover, ESG-LLaMA, a fine-tuned version of LLaMA2-7b-chat (Touvron et al., 2023) trained on ESG texts, demonstrated improved zero-shot precision compared to the original LLaMA2 across multiple prompting strategies in both four-class and nine-class ESG classification tasks. While exact F1-scores or accuracy were not reported, the visual comparison highlights ESG-LLaMA's advantage in understanding ESG-specific language.

Other studies have developed category-specific classifiers that target individual ESG labels. For example, (Schimanski et al., 2024) proposed separate models for environmental (e.g., air quality), social (e.g., data security), and governance (e.g., business ethics) texts. These models enabled multi-label ESG classification, which is useful when documents span several sustainability categories.

3.2 *Parameter-Efficient Fine-Tuning (PEFT) Methods*

Although domain-specific LLMs are effective, they remain resource-intensive to fine-tune fully, especially on small ESG datasets (Birti et al., 2025). Parameter-efficient fine-tuning (PEFT) methods offer a solution by updating only a small portion of the model's parameters. One of the most widely used PEFT methods is Low-Rank Adaptation (LoRA)(Hu et al., 2021), which reduces training complexity by inserting lightweight adapter layers and updating only a small fraction of the model's parameters. For instance, Birti et al. (2025) fine-tuned LLMs on just 0.08–0.12% of their parameters and achieved strong performance on ESG classification tasks with datasets containing as few as 1,300 samples.

Building on LoRA, QLoRA was proposed by Dettmers et al. (2023) to further improve efficiency. QLoRA builds on LoRA by quantising weights to 4-bit precision and placing adapters across all transformer layers, significantly lowering memory use without major performance loss. This makes QLoRA highly suitable for low-resource ESG tasks, where fine-tuning on full-precision models is not practical.

3.3 *Few-Shot Prompting with Instruction-tuned LLMs*

Few-shot prompting offers a lightweight alternative when labeled data is scarce, using a handful of examples to guide the model. Brown et al. (2020) introduced this approach in the original GPT-3 paper, and it's use has since been extended to newer models like GPT-4 (Almatrafi & Johri, 2025) and Claude 3.5 Sonnet (Arshad et al., 2025).

For instance, Tian and Chen (2024) used GPT-4 to classify ESG news articles using carefully crafted prompts that included label definitions and a few representative examples. Although results were strong, performance varied across languages and label types, and fine-tuned models still often outperformed prompt-only approaches.

Another study used Claude 2 with semantic similarity-based example selection, using dense embeddings. They represented both inputs and examples using dense vector embeddings to compare semantic similarities (e.g., using MiniLM). This approach improved F1 scores and reduced majority-class bias, outperforming other prompt-only baselines like Dolly v2.

3.4 *Ensemble Methods and Zero-Shot Classification*

Zero-shot learning (ZSL) allows models to generalize to unseen classes without needing labeled examples for those specific categories (Palaniappan M et al., 2023). In the context of ESG classification, this is particularly useful because ESG-related categories evolve over time, and obtaining labeled training data is expensive and often inconsistent across sources (Sokolov, Caverly, et al., 2021). By using pretrained knowledge, ZSL methods can perform classification tasks without the need for task-specific supervision while still achieving a reasonable degree of accuracy.

Previous research has shown that ensemble learning can improve model robustness by combining the outputs of multiple predictive models (Rane et al., 2024). Furthermore, the paper demonstrated that ensembles enhance performance across NLP tasks such as sentiment analysis and fraud detection. However, Rane et al. (2024) focused on supervised ensemble methods and did not explore the use of zero-shot models or domain-specific transformers.

In this research, a zero-shot ensemble strategy is adopted, combining three transformer models: FinBERT-ESG (Huang et al., 2022), ESG-RoBERTa (Xia et al., 2024), and DeBERTa-v3-MNLI (He et al., 2023). FinBERT and ESG-RoBERTa have been pretrained or fine-tuned on ESG-related corpora, while DeBERTa-v3-MNLI is a general-purpose model trained on natural language inference tasks. In this research, these models are treated as zero-shot models because they have not been further trained or adapted to the specific ESG classification task or dataset at any stage of the process. Their predictions are generated directly based on their pretrained knowledge.

In the ESG domain specifically, (H. Lee et al., 2024) demonstrated that ensemble methods combining fine-tuned transformers such as BERT, RoBERTa, and ALBERT improved ESG classification performance. In their study, each model was fine-tuned on ESG data before ensembling. When evaluated at a batch size of 20, the best-performing ensemble (BERT + RoBERTa + ALBERT) achieved an F1-score of 0.79 (Table 3), outperforming all individual models.

3.5 *Gap Identification and Research Contribution*

While several studies have explored ESG-specific transformer models, zero-shot prompting, and LoRA-based fine-tuning, most of them evaluate these techniques in isolation. There is currently a lack of comprehensive research comparing these strategies under realistic data scarcity conditions using small ESG datasets. Additionally, although ensemble models have been shown to improve performance in ESG classification tasks, their use in zero-shot settings remains underexplored; particularly with the inclusion of domain-adapted models such as FinBERT-ESG and ESG-RoBERTa.

Another underexplored area is the application of parameter-efficient fine-tuning techniques like QLoRA to ESG-specific large language models. While LoRA has shown promising results in general NLP and specialized domains like healthcare, the use of quantised LoRA in ESG classification, particularly on domain-adapted LLMs, like ESG-LLaMA has not yet been widely studied.

To address these gaps, this research evaluates three strategies for ESG classification under realistic data scarcity:

1. A zero-shot ensemble of FinBERT-ESG, ESG-RoBERTa, and DeBERTa-v3-MNLI, compared to ESG-LLaMA's performance in its zero-shot state (before fine-tuning).
2. QLoRA fine-tuning of ESG-LLaMA, assessing whether this parameter-efficient method can effectively improve the model's performance when trained on a small dataset.
3. Few-shot prompting with DeepSeek-LLM-7B-Chat, a general-purpose instruction-tuned model, tested without domain-specific adaptation.

4 METHOD

This section describes the experimental pipeline used to evaluate three approaches for ESG text classification: zero-shot inference, parameter-efficient fine-tuning, and few-shot learning. All methods are tested on the same held-out test set to ensure a fair comparison and assess how well they generalize. A visual overview of the full workflow is shown in Figure 1, and each step is explained in detail below.

- Dataset and Preprocessing: Includes data structure, cleaning, label mapping, and data augmentation.
- Zero-Shot: Ensemble Models: combines FinBERT, ESG-RoBERTa, and DeBERTa-v3-MNLI using soft voting.
- Model Inference and Prediction Extraction: Standardizes model outputs across different architectures.
- Zero-Shot: ESG-LLaMA: uses prompt-based classification.
- ESG-LLaMA QLoRA Fine-Tuning: Parameter-efficient fine-tuning using quantization and LoRA.
- Few-Shot Prompting with DeepSeek: Uses a general-purpose instruction-tuned model with in-context examples.

- Evaluation Strategy: Describes the metrics used, baseline comparisons, and qualitative error analysis.

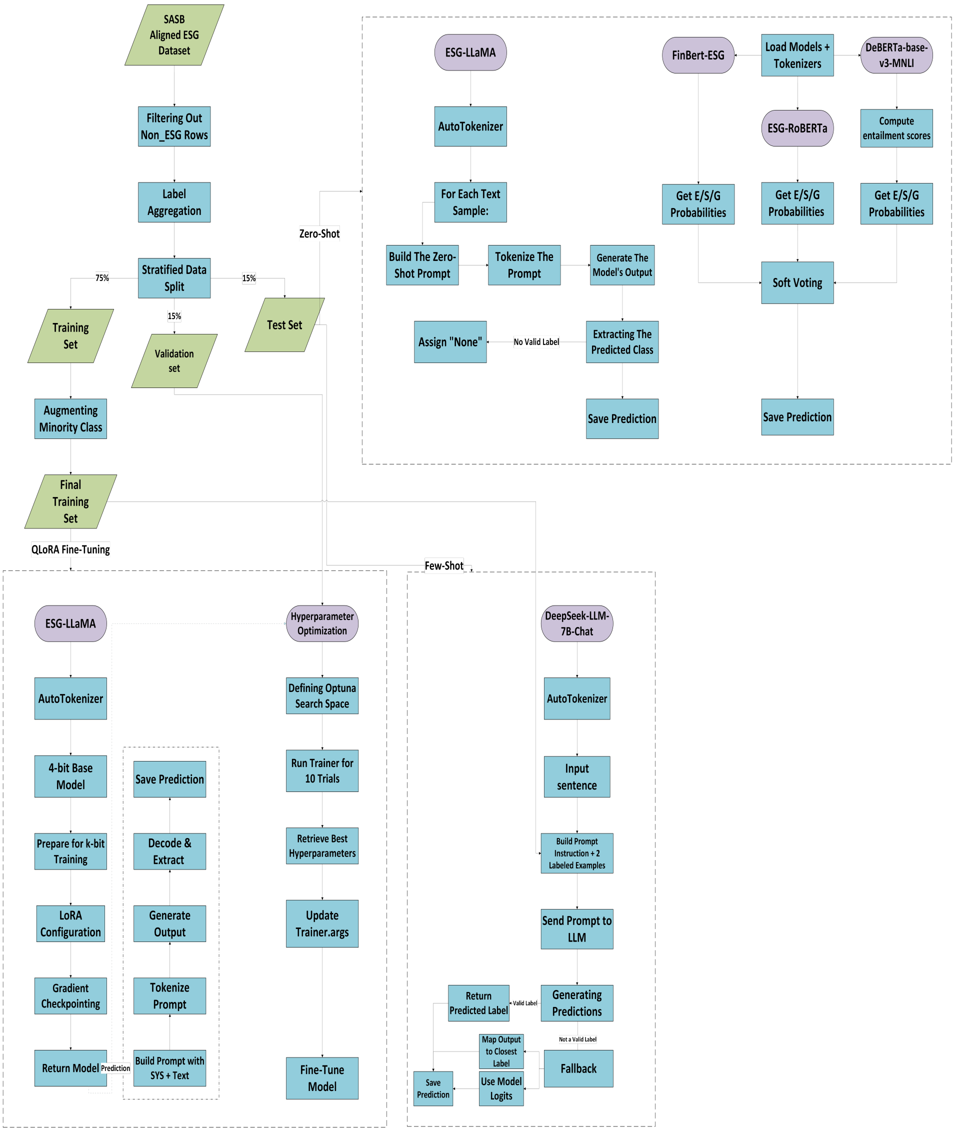

Figure 1: Research Methodological Overview

4.1 *Dataset and Preprocessing*

The SASB-aligned ESG dataset includes 6,460 labeled sentences across six original categories: Social Capital (882), Leadership & Governance (721), Environment (567), Human Capital (466), Business Model & Innovation (278), and Non-ESG (3,546).

To better align with standard industry classifications, Business Model & Innovation was combined with Leadership & Governance (Goswami et al., 2023), and Human Capital with Social (Moreira et al., 2025) due to their overlapping themes in corporate decision-making. This grouping reflects standard practices in sustainability reporting and improves comparability with established frameworks (Sustainability Accounting Standards Board (SASB), 2017).

This label consolidation is supported by ESG literature. Rating inconsistencies often stem from what Khan et al. (2016) describes as *scope divergence*; when agencies include different topics. Grouping the original five categories into three broader dimensions enhances label consistency and comparability. Overlapping reporting frameworks also challenge structured extraction, reinforcing the need for standardized ESG taxonomies (Organisation for Economic Co-operation and Development, 2025).

There were no missing values, so imputation was not needed. Non-ESG entries were removed to focus the model on three classes: Environment, Social, and Governance. This choice was based on: (1) difficulty in accurately labeling Non-ESG content (Xia et al., 2024), (2) its class imbalance (55% of data), (3) the goal of this research, which is distinguishing ESG categories,(4) reduced computational cost, and (5) ensuring consistent class definitions across evaluation settings. A limitation is that the final model cannot detect non-ESG text at inference; future work could add a first-stage ESG detector.

After filtering, a stratified 70/15/15 split was applied following the approach by Xia et al. (2024). The resulting training set contained 943 Social, 699 Governance, and only 397 Environment samples. To address this imbalance, the Environment class was augmented using paraphrasing with the T5 model (Raffel et al., 2020), generating one paraphrase per sentence. T5 is known for producing high-quality text-to-text outputs and was chosen based on prior research showing its effectiveness in improving data diversity and model performance (Bird et al., 2020). Augmentation was applied only to the training set to avoid inflating test results.

Figure 2 demonstrates the initial dataset and class distribution before label aggregation, cleaning, and augmentation.

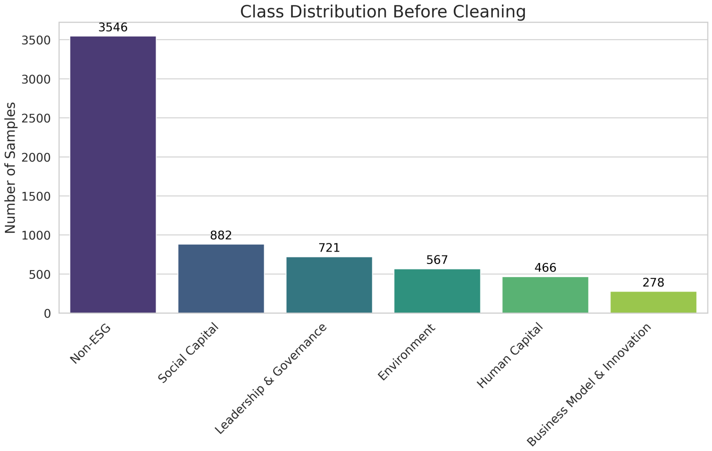

Figure 2: Class Distribution Before Cleaning

Figure 3 shows the class distribution in the training set after cleaning and augmenting the minority class (Environment) for handling class imbalance.

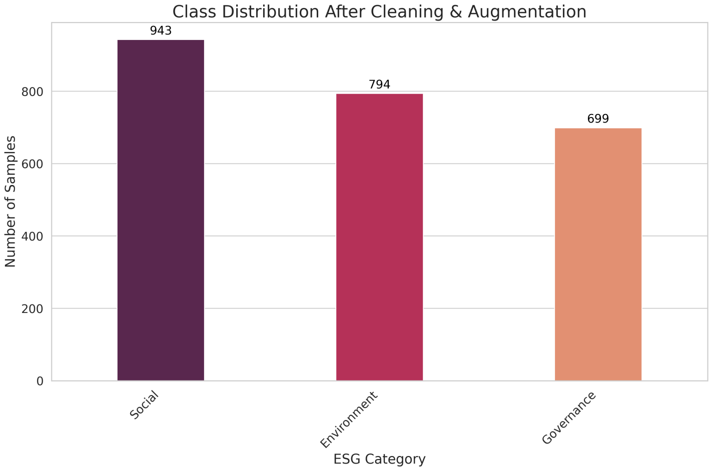

Figure 3: Class Distribution After Cleaning and Augmentation

4.2 *Zero-Shot: Ensemble Models*

This component builds on the approach of H. Lee et al. (2024), who explored ESG classification using an ensemble of fine-tuned BERT-based models. While their study focused on supervised learning, this research adapts their ensemble strategy to a strict zero-shot setting, meaning no task-specific training is applied. The selected models, FinBERT-ESG, ESG-

RoBERTa, and DeBERTa-v3-MNLI, were chosen based on their accessibility, domain relevance, and suitability for zero-shot classification.

- FinBERT-ESG: An ESG-specific model based on FinBERT, fine-tuned on 2,000 manually labeled sentences from corporate ESG and annual reports (Huang et al., 2022). It is available on Hugging Face and can be found here.

- ESG-RoBERTa: A domain-specific variant of RoBERTa, trained on over 18 million ESG-related sentences using LLM-assisted keyword filtering.g (Xia et al., 2024). The model can be found here.

- DeBERTa-v3-base-MNLI: A general-purpose model developed by Microsoft and fine-tuned on natural language inference (NLI) tasks.(He et al., 2023). Though not ESG-specific, it performs well in zero-shot setups when labels are framed as entailment prompts. (Yin et al., 2019). The model is available here.

Unlike masked language models, DeBERTa-v3-base-MNLI, rather than simply predicting masked tokens, is trained to detect word swaps and sequence alterations, improving contextual understanding (He et al., 2023). Because of its fine-tuning on the MNLI dataset, it is well-suited for zero-shot classification tasks when framed as entailment problems. In Natural Language Inference (NLI), entailment refers to a relationship where a given hypothesis logically follows from a premise. For zero-shot classification, this setup allows the model to predict whether a label description (as a hypothesis) is entailed by the input text (the premise) (Yin et al., 2019).

While DeBERTa-v3 offers general language understanding, FinBERT-ESG and ESG-RoBERTa provide ESG-specific knowledge. This mix of models is motivated by findings that suggest ensemble diversity can contribute to improved generalization performance(Dietterich, 2000).

4.2.1 *Model Inference and Prediction Extraction*

All models are evaluated directly on the test set in a strict zero-shot setting, with no additional training or fine-tuning. To ensure consistent class names across models, a label alias dictionary is applied (e.g., mapping "environmental" and "environment" to "Environment"). To ensure reproducibility, fixed seeds were set for Python, NumPy(Harris et al., 2020), and PyTorch(Paszke et al., 2019).

FinBERT and ESG-RoBERTa return logits for four categories: Environment, Social, Governance, and "None." Since Non-ESG content was removed during preprocessing, the "None" category is discarded. The remaining logits are normalized via a softmax function, and the highest-probability class is selected as the prediction.

DeBERTa-v3-MNLI operates via natural language inference (NLI), evaluating whether a sentence entails one of three hypotheses: "This sentence is about Environment," "...Social," or "...Governance." The entailment scores are treated as class probabilities, allowing comparison with the other models.

Final predictions are generated using soft voting, in which the class probability vectors from all three models are averaged. While (H. Lee et al., 2024) used F1-weighted voting based on validation performance, this research applies equal weighting to remain consistent with strict zero-shot conditions. Equal weighting or uniform voting is supported by early ensemble research (Hansen & Salamon, 1990; Zhou, 2012) and remains a common strategy in recent zero-shot classification literature (Allingham et al., 2023; Caruana et al., 2004).

4.3 *Zero-Shot Classification Using ESG-LLaMA*

This research also evaluates ESG-LLaMA in a zero-shot setting, meaning testing the model prior to fine-tuning. ESG-LLaMA is a domain-specific variant of the LLaMA2 model, fine-tuned using supervised instruction tuning (SFT) on over five million ESG-related sentences from public sources.Xia et al., 2024.

Due to its domain-specific pretraining and instruction-focused fine-tuning, ESG-LLaMA shows strong zero-shot performance in ESG text classification tasks. It demonstrates a deeper understanding of ESG-specific language compared to general-purpose LLaMA models and supports multi-class classification with prompt-based instructions. However, it is not optimized for multi-label tasks.

ESG-LLaMA was fine-tuned using a simple instruction-based format Xia et al. (2024), similar to Alpaca (Taori et al., 2023). In their original work, the authors evaluated ESG-LLaMA in various setups, including zero-shot and in-context learning. Their results showed improved performance when demonstrations were provided in the prompt. However, since this part of the research focuses on zero-shot classification, we initially adapted the exact prompt without any demonstrations. Surprisingly, the performance was below average, with a particularly low F1-score, and the *Environment* class that was rarely predicted. Details of the prompt and classification report are provided in Appendix B. 41

Based on this outcome, the prompt was revised, and the best-performing version was used in further experiments. The final format is:

```
Instruction: Classify the following
text into one of these ESG categories:
Environment, Social, or Governance.
Input: [text]
Output: [model prediction]
```

This prompt is tokenized and converted into input IDs, which are passed to the model for generation. The model generates tokens autoregressively using greedy decoding, a strategy that selects the most likely next token at each step. This ensures deterministic and concise outputs suitable for classification.

Because the model's output includes the input prompt at the beginning, token slicing was applied to remove those initial tokens and isolate only the generated part.Specifically, the first N tokens, corresponding to the prompt length, are discarded to isolate the generated response.

The model's output is processed using regular expression to extract the predicted ESG category. If one of the valid class labels ("Environment", "Social", or "Governance") is found, it is recorded as the prediction. Otherwise, the label is set to "None." Although a fallback mechanism was included to assign the label "None" when the model failed to output a valid ESG category, this condition was never triggered in practice. All predictions fell within the expected label set, so no additional filtering or corrections were required. All predictions are then saved in a CSV file for evaluation.

4.4 *ESG-LLaMA QLoRA Fine-Tuning*

This part of the research investigates how parameter-efficient fine-tuning using QLoRA affects the performance of ESG-LLaMA on ESG text classification. The same base model used for zero-shot inference, ESG-LLaMA, is further adapted using the training set for fine-tuning, while the validation set is used for hyperparameter optimization.

QLoRA enables efficient fine-tuning of large models by combining 4-bit quantization (Dettmers & Zettlemoyer, 2023) and LoRA adapters (Hu et al., 2021). The model is loaded in a special 4-bit format called NF4 (NormalFloat4), which is designed to match the typical distribution of neural network weights (i.e., similar to a normal distribution). This is achieved using Hugging Face's `BitsAndBytesConfig`, with double quantization enabled to compress both the model weights and their scaling values, further reducing memory usage.

Input texts were tokenized with `truncation=True`, which ensures that sequences exceeding the model's maximum token limit are shortened from the end, preventing input overflow and maintaining compatibility with model constraints. The pad token was set to the end-of-sequence (EOS) token to ensure compatibility with the language model. Rather than using fixed-length padding, a custom data collator applied dynamic padding to the longest sequence in each batch, optimizing memory usage while preserving consistent batch structure.

During training, the 4-bit weights are temporarily cast to BFloat16 format to preserve numerical accuracy. Instead of updating all parameters, LoRA adapters, which are small trainable modules, are inserted into specific layers of the transformer. In this setup, only the adapters are trained while the original model weights remain frozen. The adapters are added using the `peft` (Mangrulkar et al., 2022) library, and the training process uses an 8-bit optimizer from the `bitsandbytes` library (Dettmers et al., 2022), which helps avoid memory crashes during longer sequences.

Following Hu et al. (2021), we insert LoRA adapters with rank $r = 8$ and scaling factor $\alpha = 16$ into the query and value projections $(q_{\text{proj}}, v_{\text{proj}})$ of the model. The rank controls how many directions the adapter can learn. According to their ablation studies, setting the ranks between 4 and 8 already provides strong performance when only q and v are adapted.[1]

The scaling factor α just scales how big the adapter's contribution is relative to the original weight. Hu et al. (2021) argues that once the learning rate is fixed, changing α has the same effect as changing the learning rate, so they never tune it. They simply keep α equal to, or a simple multiple of, the first rank they try.

We adopted the configuration used for RoBERTa-large in the original paper, selecting the same rank and alpha values, as they were shown to be effective. [2]

The `bias="none"` setting was used to freeze the original model's bias parameters during training, updating only the LoRA adapter weights. This follows the default and memory-efficient configuration suggested by (Hu et al., 2021).

To reduce the risk of overfitting, a dropout rate of 0.1 is applied within the LoRA adapters. This introduces mild regularization by temporarily freezing a subset of neurons during training. The QLoRA methodology recommends this dropout rate for models up to 13 billion parameters, achieving strong performance without significantly reducing capacity. (Dettmers et al., 2023).

[1] See Table 6 in Hu et al. (2021).

[2] See Table 9 in Hu et al. (2021).

Additionally, gradient checkpointing is used during the fine-tuning process to save GPU memory (Chen et al., 2016). This method works by saving only part of the activations during the forward pass and recalculating the rest during the backward pass. As described by Chen et al. (2016), this reduces memory use from $\mathcal{O}(n)$ to $\mathcal{O}(\sqrt{n})$, with only a modest increase in computation.

Altogether, this fine-tuning configuration is informed by both theoretical findings and practical results. Table 1 summarizes the key components used in the QLoRA implementation.

Component	**Configuration**
Tokenizer	`use_fast=False`, `pad_token = eos_token`
Model quantization	`load_in_4bit=True`, `torch_dtype=torch.float16`
LoRA adapters	`r=8`, `lora_alpha=16`, `target_modules=["q_proj","v_proj"]`, `dropout=0.1`,bias="none"
Gradient checkpointing	`model.gradient_checkpointing_enable()`, `use_cache=False`
Hyperparameter search	Optuna TPE sampler (`seed=42`), 10 trials
Seeds & reproducibility	`random.seed(42)`, `np.random.seed(42)`, `torch.manual_seed(42)`, HF `seed=42`

Table 1: Fine-Tuning Configuration Details

4.4.1 *Justification of Adapter Selection*

During initial experiments, I attempted to fine-tune LoRA adapters of rank 8 in all four self-attention projections (*q_proj*, *k_proj*, *v_proj*, and *o_proj*). However, this setup exceeded the available memory on our shared GPU, causing the process to fail. This occurred despite efforts to reduce memory fragmentation. To complete the experiments within the available runtime and memory budget, I therefore limited fine-tuning to only the query and value adapters (*q_proj* and *v_proj*) while maintaining a rank of 8.

This decision is also supported by findings in the literature. Hu et al. (2021) suggests placing low-rank adapters in all four projections, namely, *q_proj*, *k_proj*, *v_proj*, and *o_proj*, with a rank of 8 and a scaling factor of 16, would replicate full fine-tuning. However, their ablation study (Hu et al., 2021, p. 10) shows that restricting adapters to just *q_proj* and *v_proj* still achieves over 95% of the performance gains.

Further support comes from Aghajanyan et al. (2021), who proposes that effective fine-tuning occurs within a low-dimensional intrinsic subspace, often involving just a few hundred parameters. This reinforces the choice to focus on the query and value projection, which have been shown to be the most information-dense. (Hu et al., 2021).

Therefore, to maximize efficiency and enable fair comparison, this research will fine-tune only those two primary adapters.

4.4.2 *Hyperparameter Optimization*

Hyperparameter optimization plays a critical role in maximizing the performance of LLMs. As with smaller neural networks, key hyperparameters such as learning rate, batch size, and the number of epochs significantly affect training dynamics, generalization ability, and computational cost. In large models, the impact of these choices is even greater due to their scale and resource demands. As shown in (Ding et al., 2024; Liao et al., 2022), even small changes in hyperparameters can lead to substantial differences in model performance and efficiency.

To tune ESG-LLaMA, Optuna (Akiba et al., 2019) was implemented, which is a widely adopted and efficient hyperparameter optimization framework. The Hugging Face `Trainer` class was combined with Optuna's `hyperparameter_search()` method to explore and evaluate different configurations. This setup uses the macro-averaged F1-score on the validation set as the objective, which is suited for our imbalanced multi-class classification.

The search space was defined for four key hyperparameters:

- Learning rate: sampled on a logarithmic scale between 1.5e-4 and 3e-4, as recommended by Bergstra et al. (2011) who show that log scale is important when testing values that can be very different in size.
- Batch Size: Categorical choice between 1 and 2, accounting for GPU memory limitations due to model size
- Epochs: Sampled as integers between 3 and 5. Values above 5 were avoided to limit overfitting and long training times.
- Gradient accumulation steps: Selected between 4 and 8 to simulate larger batch sizes under limited GPU memory. It means instead of updating the model every single batch, it runs 4 or 8 small batches, accumulates their gradients, and then updates the model. This technique lowers the memory usage while approximating large-batch training behavior (Gao et al., 2021).

Optuna was run for 10 trials. In each trial, a configuration was sampled, and the model was trained and evaluated on the validation set. Based on the macro F1-score, Optuna gradually learned which combinations were more promising. The use of the macro F1-score ensured that minority classes, such as 'Environment', were not ignored during evaluation. Each trial was logged to a CSV file for reproducibility. As shown in table 2, the best performance was achieved in trial 4.

Trial	Learning Rate	Batch Size	Epochs	Accum. Steps	Macro F1
0	0.000215	2	3	8	0.7987
4	0.000184	1	5	8	**0.8469**

Table 2: Trial 0 served as the starting point, while Trial 4 achieved the highest macro-averaged F1-score.

By implementing QLoRA, we were able to fine-tune a large language model, such as ESG-LLaMA, using a small training dataset and limited computational resources while maintaining strong performance.

4.4.3 *Prompting Strategy*

Initial experiments explored Chain-of-Thought-style prompting by appending the phrase *"Let's think step by step"* to the instruction. This approach aimed to encourage intermediate reasoning during classification. However, the model did not consistently produce reasoning steps, and performance remained weak. The outcome confirms the findings by Bsharat et al. (2024), who reported that Chain-of-Thought prompting does not generally enhance the performance of LLaMA-2 models, due to its limited reasoning capabilities.

Next, a simple instruction-style prompt was tested in two formats: a plain zero-shot version and a chat-style version using the `[INST]...[/INST]` format recommended for LLaMA 2. (AI, 2023) The chat-style format performed slightly better and was selected for further experiments. This format clearly separates user input from system-level context: the system message defines the task, followed by the specific input text, and the model is instructed to output exactly one label. The prompt is structured as follows:

```
<s>[INST] <<SYS>>
You are an ESG classification assistant.
Classify the text into exactly one category:
Environment, Social, or Governance.
<</SYS>>
Text: {example}
Answer ONLY with a single label:[/INST]
```

The generation and post-processing steps follow the same procedure used for ESG-LLaMA in the zero-shot setup: greedy decoding, prompt token slicing, and regular expression for label extraction following the original paper's technique (Xia et al., 2024). The final prediction is the matched ESG label from the model's response; otherwise, "None" is returned. However, as described in the results section, this fallback was never triggered.

4.5 *Few-Shot Learning*

A key challenge in ESG text classification is the scarcity of labeled data. So far, this research has explored two approaches to address this: zero-shot inference and parameter-efficient fine-tuning. In this section, a third strategy is introduced: few-shot prompting with a general-purpose large language model that has not been trained on ESG-specific data.

The goal is to see if a model can perform well on ESG classification when it's given only a few labeled examples. For this, the DeepSeek-LLM-7B-Chat model is used, which is a lightweight language model fine-tuned to follow instructions, developed by DeepSeek-AI (DeepSeek-AI, 2024).

4.5.1 *DeepSeek*

DeepSeek-LLM-7B is the base decoder-only transformer with 7 billion parameters, built on the same architectural principles as LLaMA (Touvron et al., 2023). As part of the DeepSeek family of models, it is designed to support advanced language understanding and reasoning in both English and Chinese (Bi et al., 2024). It consists of 30 transformer layers. Each layer can look at different parts of the input using 32 attention heads, which allow the model to focus on different words and their relationships. To help it understand the order of words in a sentence, the model uses a special method called Rotary Positional Embeddings (RoPE). It also uses an activation function called SwiGLU, which helps process the information more effectively. For stability during training, it applies RMSNorm, a technique that keeps the model's internal values balanced. Instead of making the model wider (which would require more memory), the developers chose to make it deeper by adding more layers, which improves its ability without needing too much extra computing power.

To help the model understand different languages, DeepSeek uses a tokenizer based on Byte-level Byte Pair Encoding (BBPE) (Bi et al., 2024). This tokenizer was trained on around 24 GB of text in both English and Chinese, allowing the model to handle multiple languages well. It has a vocabulary of about 102,400 unique tokens, including some special ones that can be used later if needed.

The model was trained on a huge dataset with 2 trillion tokens (Lu et al., 2024). This dataset was carefully cleaned to remove duplicates and low-quality data. It includes a mix of web data from Common Crawl, hand-picked high-quality texts, and content from different domains to make sure the model learns a wide range of language styles and topics.

After pretraining, it was improved through: (Bi et al., 2024):

1. Supervised Fine-Tuning (SFT): In this step, the model was trained using about 1.5 million example conversations, where each example includes a question and a good human-written answer. These examples came from various areas like language tasks, math problems, and coding questions, and were in both English and Chinese.

2. Direct Preference Optimization (DPO): In this second step model learns to give more helpful and safer answers by being rewarded for good behavior. It helps the model follow instructions more naturally and respond better to open-ended prompts.

The DeepSeek-LLM-7B-Chat model was selected for this part of the research due to its combination of accessibility, efficiency (Bi et al., 2024), and promising capabilities. As an open-source model, it is freely available and can be modified without restriction, which is ideal for academic use. With 7 billion parameters, it has a good balance between performance and memory requirements, making it suitable for experiments on limited hardware.

While DeepSeek-7B-Chat has not been directly evaluated on ESG tasks, models from the broader DeepSeek family have shown strong performance in instruction-following and reasoning tasks, particularly in reinforcement learning contexts (Wang et al., 2025). These characteristics make it a strong candidate for few-shot ESG classification using structured prompts and constrained outputs. The model can be found here.

4.5.2 *Few-Shot Prompting Strategy*

To adapt DeepSeek for ESG classification, we used a few-shot prompting strategy (Brown et al., 2020). Each input prompt included:

- A system message instructing the model to act as an ESG classifier and respond with a single word from the allowed categories.

- Six labeled example sentences (two per class) drawn from the training set.

- The target sentence to be classified.

This approach follows established prompt engineering practices for LLM-based classification (Liu et al., 2021). The system message is shown below:

```
system_msg = (
        "You are an expert ESG classifier. "
        "When given a piece of text,
        reply with exactly one word:"
        "Environment, Social, or Governance."
        )
```

Then each example follows this structure:

```
USER: Classify the ESG category of the following text:
[text]

ASSISTANT: [correct label]
```

The final prompt ends with a new `USER` message containing the target text, and the model is expected to respond as `ASSISTANT` with the predicted label.

To ensure that the model's output was always a valid ESG label, constrained decoding was implemented using the `force_words_ids` parameter in the HuggingFace Transformers library. This mechanism restricts the model's output to tokens corresponding to the three ESG categories, thereby preventing the generation of out-of-scope or ambiguous responses.

After generation, the predicted token was mapped to the corresponding ESG label using robust string matching. In rare cases where the output token did not exactly match any category, a logit-based fallback was applied. In this fallback, the model's output logits for the allowed tokens were compared, and the label with the highest probability was selected. This method of resolving ambiguous outputs by maximizing the model's confidence among valid classes is supported in recent literature as a best practice for forced-choice classification with LLMs (Boehnke et al., 2024; Mishra, 2023).

For each sentence, the model generated a single-token response. The extraction process included:

- Decoding the generated token.
- Mapping the decoded string to one of the three ESG categories via substring matching (e.g., "env" mapped to "Environment").
- If no match was found, selecting the category with the highest model confidence (logit value) among the allowed options.

In preliminary experiments without these constraints, the model occasionally produced "Unknown" outputs , which were outside the expected label set. The final constrained setup eliminated such cases, ensuring that every input received a valid ESG label.

4.6 *Evaluation Metrics*

Following Grandini et al. (2020), this research applies several evaluation metrics that are well-suited for multi-class classification under class imbalance. The main evaluation metric is the macro-averaged F1-score, which computes the F1-score for each class; balancing precision and recall through their harmonic mean, and then averages these scores equally across all classes. This ensures that minority classes, such as *Environment*, are not overshadowed by dominant ones and that model performance is evaluated fairly across all categories.

In addition to macro F1, balanced accuracy is also reported. This metric averages recall across classes and is particularly useful for understanding how well the model performs in identifying underrepresented categories. To provide a more detailed view, per-class precision and recall scores are included to highlight which ESG categories are more or less challenging to classify. Special attention is given to class-wise recall, since false negative instances where the model fails to identify the correct ESG topic may carry more practical consequences than occasional misclassifications.

All models are evaluated on the same held-out test set using macro F1-score, balanced accuracy, and per-class precision and recall, ensuring consistency and fairness in performance comparison.

5 RESULTS

This part of the research reports the results of evaluating three different strategies for ESG text classification: zero-shot inference, parameter-efficient fine-tuning, and few-shot prompting. All models were evaluated on the same held-out test set, and performance was measured using balanced accuracy, macro F1-score, and precision and recall for each class to ensure a fair and consistent comparison.

5.1 *Majority Class Baseline Performance*

To provide context for the models' performance, a basic baseline was established by always predicting the most frequent class in the training data. In this case, the majority class was *Social*, which appeared in 943 out of 2436 samples of our training set. This majority-class baseline achieved a balanced accuracy of 0.333 and a macro F1-score of 0.250. All tested models performed well above this baseline, demonstrating their effectiveness in distinguishing between the ESG categories.

5.2 *Zero-Shot Classification with Ensemble models and ESG-LLaMA*

Table 3 displays the precision, recall, and F1-score for each class obtained from two zero-shot classification approaches: the ESG-specific ensemble (composed of FinBERT-ESG, ESG-RoBERTa, and DeBERTa-v3-MNLI) and the ESG-LLaMA model. The ensemble consistently outperformed ESG-LLaMA in both precision and recall across all ESG categories. This suggests it made more accurate predictions and was better at detecting relevant examples, with particularly strong improvements in the Environment and Governance categories.

Model	Environment	Social	Governance
Precision Ensemble	0.602	0.786	0.697
Precision ESG-LLaMA	0.483	0.728	0.690
Recall Ensemble	0.941	0.759	0.507
Recall ESG-LLaMA	0.847	0.724	0.400
F1-score Ensemble	0.734	0.772	0.587
F1-score ESG-LLaMA	0.615	0.726	0.506

Table 3: Class-wise Precision, Recall, and F1-score for Ensemble and ESG-LLaMA

Since no model fine-tuning was applied, these results highlight the models' inherent ability to generalize to ESG-related texts. The ensemble's improved performance may be connected to the complementary strengths of its components, each capturing different aspects of ESG language. However, the relatively lower performance in the Governance category for both models suggests this area remains challenging, potentially due to its less distinct linguistic patterns.

Table 4 summarizes the models' overall performance, reporting balanced accuracy and the macro-averaged F1-score. The ensemble model attained superior results, with a balanced accuracy of 0.735 and a macro F1-score of 0.698, suggesting it performed more effectively in the zero-shot classification task.

Model	Balanced Accuracy	Macro F1-score
Ensemble	0.735	0.698
ESG-LLaMA	0.657	0.616

Table 4: Balanced Accuracy and Macro F1-score for Ensemble and ESG-LLaMA

Figure 4 displays the F1-scores for each ESG category, comparing the zero-shot performance of the ensemble model and ESG-LLaMA. The ensemble achieved higher scores across all three classes, which reinforces the earlier quantitative results. The largest performance difference appears in the Environment category, indicating the ensemble's stronger ability to capture environmentally relevant features without fine-tuning.

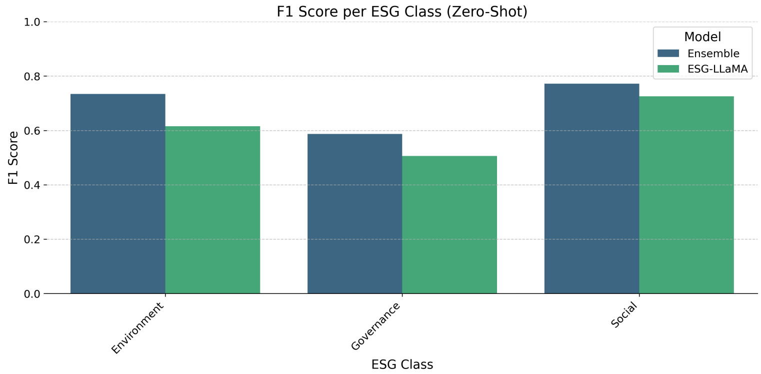

Figure 4: F1 score per ESG category for the Ensemble and ESG-LLaMA models (zero-shot setup).

To better understand model errors, confusion matrices were analyzed for both the ensemble model and the ESG-LLaMA model (see Figure 5). Both models demonstrated their strongest performance on the *Social* category. The ensemble correctly classified 154 instances, while ESG-LLaMA identified 147 accurately. This likely reflects the fact that *Social* was the majority category in the test dataset. The overall distribution of errors revealed noticeable differences between the two models. ESG-LLaMA showed a stronger tendency to confuse *Governance* with both *Environment* and *Social*, misclassifying 45 instances into each of these categories. In contrast, the ensemble model exhibited a more even distribution of errors across categories and achieved higher precision for *Environment*, correctly identifying 80 samples. This indicates that the ensemble was better at distinguishing environmentally focused texts with fewer misclassifications.

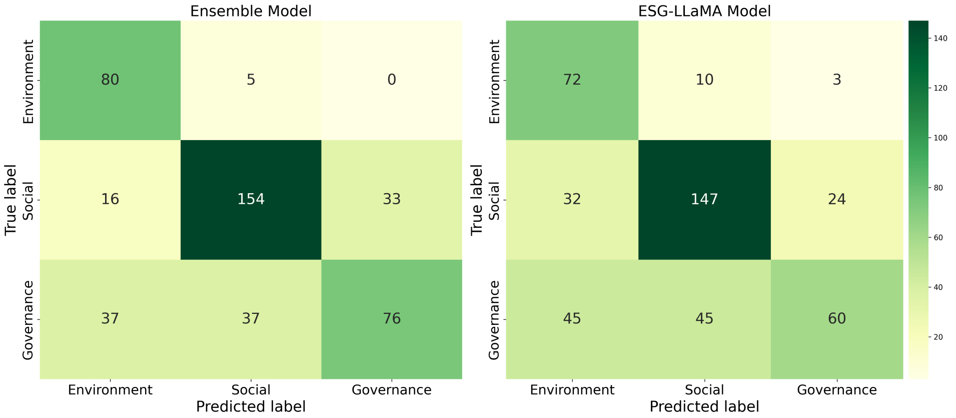

Figure 5: Confusion Matrices for Zero-Shot Models

5.2.1 *Detailed Error Analysis : Zero-Shot*

Environment:

- Ensemble: Incorrectly labeled 5 instances as *Social*, and none as *Governance*.
- ESG-LLaMA: Misclassified 10 samples as *Social* and 3 as *Governance*.

The ensemble demonstrates stronger precision and was more accurate overall in distinguishing Environmental texts, especially with minimal confusion involving *Governance*.

Social:

- Ensemble: Misidentified 16 samples as *Environment* and 33 as *Governance*.
- ESG-LLaMA: Showed higher confusion with *Environment* (32 misclassified) and fewer errors toward *Governance* (24 misclassified).

Both models had difficulties separating *Social* from the other categories. ESG-LLaMA showed a stronger bias toward confusing it with *Environment*, while the ensemble more often mistook it for *Governance*.

Governance:

- Ensemble: 37 samples misclassified as *Environment*, and 37 as *Social*.
- ESG-LLaMA: Displayed even greater confusion, with 45 instances each being misclassified as *Environment* and *Social*.

Governance remains the most challenging class for both models. ESG-LLaMA misclassified a total of 90 *Governance* samples (45 as *Environment*, 45 as *Social*), compared to 74 misclassifications by the ensemble model. This suggests the ensemble has better recall for *Governance*.

Since the focus of this research is the performance of the ensemble model, we do not consider the individual performance. However, a full classification report on each individual model in the ensemble is available in Appendix A on page 40.

5.3 *QLoRa Fine-Tuned Classification with ESG-LLaMA*

To assess the impact of task-specific training, ESG-LLaMA was fine-tuned using the ESG-labeled dataset. Table 5 presents the class-wise precision, recall, and F1-score for the Environment, Social, and Governance categories. The model achieved strong performance across all three classes, with F1-scores above 0.85. Particularly, the Social category showed the highest F1-score (0.901), while the Governance class (previously the most challenging) also saw a significant improvement.

Metric	Environment	Social	Governance
Precision	0.885	0.901	0.865
Recall	0.906	0.901	0.853
F1-score	0.895	0.901	0.859

Table 5: Class-wise Precision, Recall, and F1-score for fine-tuned ESG-LLaMA.

The overall evaluation matrices are summarised in Table 6. The QLoRA fine-tuned ESG-LLaMA reached a macro-averaged F1-score of 0.885 and a balanced accuracy of 0.887. Compared to its zero-shot performance, this shows a significant improvement in both predictive performance and class balance. These results confirm that fine-tuning on ESG-specific data helps the model to better align with domain-specific language patterns, leading to more reliable classification.

Metric	Fine-tuned ESG-LLaMA
Macro F1-score	0.885
Balanced Accuracy	0.887

Table 6: Overall macro F1-score and balanced accuracy for fine-tuned ESG-LLaMA.

The confusion matrix for the fine-tuned ESG-LLaMA model (See Figure 6) provides further insights into its improved classification behavior. Compared to the earlier zero-shot version, the fine-tuned ESG-LLaMA makes fewer mistakes and shows stronger overall performance across all ESG categories.

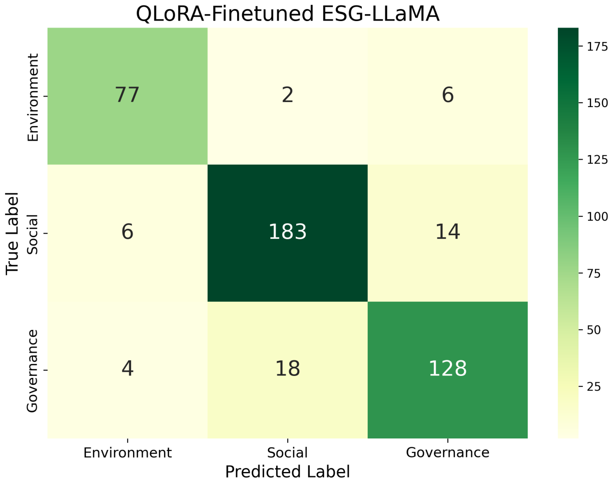

Figure 6: Confusion Matrix for Fine-Tuned ESG-LLaMA

5.3.1 *Detailed Error Analysis: QLoRA Fine-Tuned ESG-LLaMA*

- *Environment:* The model correctly predicted 77 out of the total Environment texts, with only 2 misclassified as Social and 6 as Governance. This means it now understands environmental language more clearly and shows less confusion with other categories, especially *Social*, which used to be a common source of error.

- *Social:* The *Social* category was the most accurately classified, with 183 true positives. Only 6 instances were confused with Environment, and 14 were misclassified as Governance. Although *Social* is the most common class, the model didn't overpredict it and could still correctly identify most examples. Fine-tuning helped the model learn which words and patterns belong specifically to social topics.

- *Governance:* Predictions of the Governance class also benefited greatly from fine-tuning. The model correctly predicted 128 instances, with 18 confused as Social and just 4 as Environment. These numbers show a considerable drop in misclassification compared to earlier evaluations.

Fine-tuning with QLoRA clearly reduced confusion between different ESG categories. Most importantly:

- Environmental texts are now much less likely to be mistaken for Social issues, showing clearer boundaries between categories.
- Social texts, which often mix language from the other two categories, still cause some confusion but are classified with more confidence than before.
- Governance, which used to be the hardest to identify correctly, now shows high precision and recall, with very few errors related to Environment.

The error pattern is more balanced, and the model now focuses more accurately on the correct category. Overall, fine-tuning reduced bias toward the majority class and made the model more sensitive to the small differences in context that are key for accurate ESG classification.

5.4 *Few-Shot Classification with DeepSeek*

Table 7 presents the class-wise precision, recall, and F1-score for the few-shot DeepSeek model. The model achieves high precision for the Environment (0.857) and Social (0.846) classes, while precision for Governance is notably lower (0.540). Recall is highest for the Governance class (0.860), indicating that the model is sensitive to Governance examples but also frequently predicts this class, as reflected in its lower precision. The F1-scores, which balance precision and recall, are highest for Environment (0.730), followed by Governance (0.663) and Social (0.678).

Metric	Environment	Social	Governance
Precision	0.857	0.846	0.540
Recall	0.635	0.567	0.860
F1-score	0.730	0.678	0.663

Table 7: Class-wise Precision, Recall, and F1-score for the few-shot DeepSeek model.

Table 8 reports the overall macro F1-score and balanced accuracy. The few-shot DeepSeek model achieves a macro F1-score of 0.690 and a balanced accuracy of 0.687, substantially outperforming the majority-class baseline. This demonstrates the model's ability to capture relevant features for all ESG categories, rather than simply learning to predict the most frequent class.

Metric	Few-Shot DeepSeek
Macro F1-score	0.690
Balanced Accuracy	0.687

Table 8: Overall macro F1-score and balanced accuracy for the few-shot DeepSeek model.

5.4.1 *Detailed Error Analysis: DeepSeek*

To gain deeper insight into the performance of the few-shot DeepSeek model, we analyzed its errors using the confusion matrix presented in Figure 7.

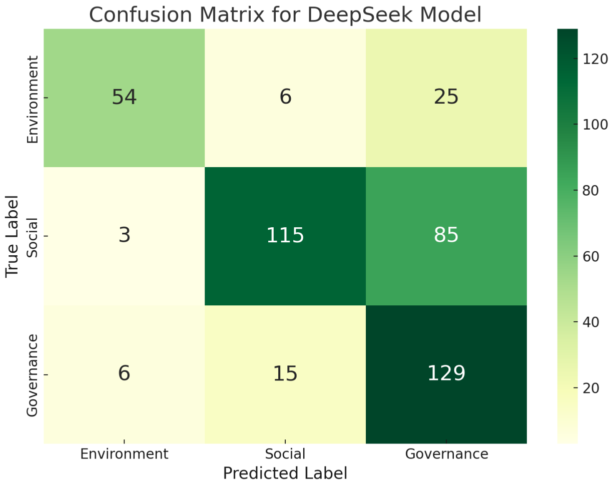

Figure 7: Confusion Matrix for DeepSeek Model

Environment: 54 sentences were correctly classified as Environment. 25 Environment sentences were incorrectly classified as Governance and 6 were misclassified as Social. The model tends to confuse Environment with Governance more than with Social. This may indicate overlapping vocabulary or thematic similarities between Environment and Governance texts in the dataset.

Social: 115 sentences were correctly classified as Social. 85 Social sentences were misclassified as Governance, a substantial proportion, and only 3 Social sentences were misclassified as Environment. The largest source of error for Social is confusion with Governance. This suggests that the model has difficulty distinguishing between Social and Governance topics, possibly due to ambiguous or multi-faceted content that touches on both areas.

Governance: 129 sentences were correctly classified as Governance. 15 Governance sentences were misclassified as Social and 6 Governance sentences were misclassified as Environment. The model more often mislabels Social sentences as Governance than the other way around. This imbalance implies that, when it's unsure, the model gravitates toward the Governance label, which is evident in the class's high recall paired with lower precision.

The model frequently predicts Governance when uncertain, as evidenced by the high number of Social and Environment sentences misclassified as Governance. This is consistent with the lower precision but higher recall for the Governance class reported in Table 7.

Environment sentences are less likely to be confused with Social, and most misclassifications are with Governance. This suggests that the linguistic features of Environment are more distinct from Social, but still share some overlap with Governance.

The frequent confusion between Social and Governance, especially the 85 Social examples mislabeled as Governance, highlights the model's difficulty in distinguishing these two overlapping categories. This confusion could arise from overlapping themes within the sentences or from a lack of clearly defined examples.

Overall, the confusion matrix reveals that while the model performs well on Environment and Governance individually, there is substantial confusion between Social and Governance categories.

5.5 *Overall Comparison*

Figure 8 illustrates the F1-scores achieved by each model for the three ESG categories: Environment, Social, and Governance. Across all classes, the QLoRA ESG-LLaMA model consistently outperforms the other approaches, achieving F1-scores above 0.85 for both Environment and Social, and above 0.80 for Governance. This indicates that parameter-efficient fine-tuning with QLoRA achieves the most robust and balanced performance across ESG topics.

The Few-Shot DeepSeek and Zero-Shot Ensemble models demonstrate comparable performance, with F1-scores in the range of 0.65 to 0.75 for most categories. Notably, both models perform best on the Environment and Social classes, but show a relative decline in F1-score for Governance, suggesting that distinguishing Governance-related content remains a greater challenge for models that are not trained or fine-tuned on a specific task.

The Zero-Shot ESG-LLaMA model, while outperforming a naive baseline, lags behind the other approaches, particularly for the Governance

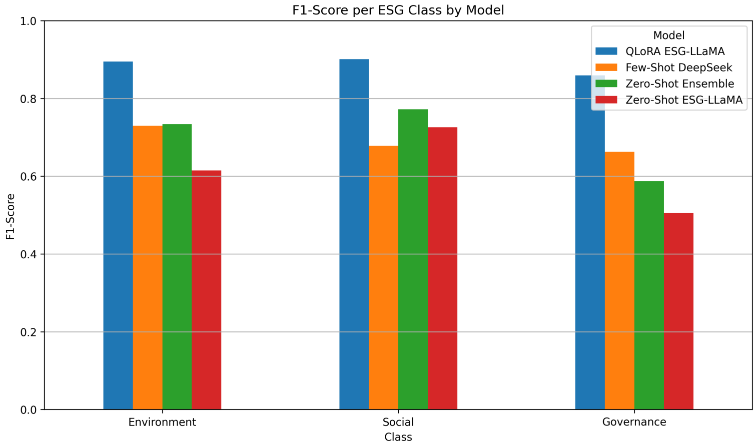

Figure 8: Overview of Each Model's F1-Score per Category.

class where its F1-score drops below 0.60. This result highlights the limitations of zero-shot inference for nuanced, domain-specific classification tasks such as ESG topic detection.

Overall, the results demonstrate that while all advanced models substantially surpass the majority-class baseline (macro F1-score = 0.250, not shown), parameter-efficient fine-tuning (QLoRA) provides the greatest gains in both precision and recall across all ESG categories. These findings underscore the importance of targeted adaptation for achieving high-quality text classification in specialized domains.

6 DISCUSSION

The primary goal of this research was to identify the most effective low-resource strategy for ESG text classification when only limited labeled data is available, specifically evaluating QLoRA fine-tuning, few-shot prompting, and zero-shot ensembles. To explicitly revisit the research questions: (1) How does QLoRA fine-tuning ESG-LLaMA perform compared to its zero-shot variant and a zero-shot ensemble of FinBERT-ESG, ESG-RoBERTa, and DeBERTa-v3-MNLI? (2) How effective is few-shot prompting with an instruction-tuned LLM (DeepSeek) compared to zero-shot and QLoRA-fine-tuned models?

Across the held-out test set, QLoRA fine-tuned ESG-LLaMA delivered the highest macro F1-score (0.885) and balanced accuracy (0.887), clearly outperforming both its zero-shot variant (macro F1 = 0.616) and the zero-shot ensemble (macro F1 = 0.698). This finding supports the Dettmers et al. (2023) evaluation that 4-bit LoRA fine-tuning can approximate the effectiveness of full fine-tuning, as our fine-tuned model closely matched supervised benchmark performance (macro F1 ≈ 0.91) reported for ESG-RoBERTa trained on significantly larger datasets (Xia et al., 2024), despite updating only less than 1% of the model's weights.

The zero-shot ensemble also achieved strong results, surpassing ESG-LLaMA in zero-shot mode and significantly outperforming the majority-class baseline (macro F1 = 0.250). The ensemble's advantage is largely attributed to model diversity, as each component captures different linguistic patterns, resulting in improved overall classification accuracy. This finding directly aligns with prior ensemble research (H. Lee et al., 2024; Zhou, 2012), reinforcing that heterogeneous ensembles enhance generalization.

Few-shot prompting with DeepSeek-LLM-7B-Chat achieved a macro F1-score of 0.690, matching the performance of the zero-shot ensemble. However, it showed the largest variation across classes and tended to over-predict the Governance category.

What's especially interesting is that giving DeepSeek only two example prompts per category was enough to reach the same F1-macro score as a much more complex ensemble of both specialized and general models. This demonstrates that even smaller language models can quickly adapt to new classification tasks with just a few in-context examples.

The result supports key findings from previous research. The original GPT-3 paper (Brown et al., 2020) demonstrated that language models can handle a broad variety of tasks effectively. Similarly, in the context of ESG news classification, Tian and Chen (2024) showed that increasing the number of few-shot examples led to higher macro F1-scores for GPT-4, reinforcing the effectiveness of few-shot learning as a general strategy.

The confusion matrices provided valuable insights into the effectiveness of each approach. Fine-tuning notably reduced confusion between Environment and Social categories and significantly boosted Governance precision (from 0.506 to 0.865). DeepSeek, even with few-shot examples, still struggled to differentiate between Social and Governance, highlighting the inherent complexities in ESG text classification that remain challenging without domain-specific tuning.

6.1 *Scientific and Societal implications*

This research addresses an important gap, which is finding reliable ESG sentence classifiers with limited annotated data. It shows that two complementary low-resource approaches can effectively bridge this gap:

Parameter-efficient fine-tuning (QLoRA): By updating under 1% of model parameters, QLoRA achieves macro-F1 scores comparable to full supervised training. This contributes to NLP methodology by demonstrating that high-accuracy domain adaptation is possible without access to large-scale hardware or extensive labeled data, thus lowering the barrier for researchers and small- to medium-sized enterprises (SMEs).

Zero-shot ensemble of pre-trained models: In settings where training is not possible, a diverse ensemble of pretrained models is an accessible solution when training resources are unavailable.

Together, these methods provide a practical toolkit: QLoRA delivers high accuracy with minimal resources, while the ensemble ensures baseline performance in zero-resource scenarios. On a social level, both approaches enhance access to automated ESG screening, empowering stakeholders, especially in low-resource contexts, to detect greenwashing and social-washing in real time and create greater corporate transparency.

6.2 *Limitations*

Despite promising results, limitations must be acknowledged. The dataset was small and imbalanced, particularly for the Environment class; and data augmentation may have introduced paraphrasing artifacts that biased the model. Moreover, due to memory constraints, QLoRA fine-tuning was limited to two adapter modules (query and value projections), possibly limiting performance. Moreover, while multiple prompt templates were tried and the best-performing version was retained, the exploration was not systematic; a broader prompt search or automatic prompt optimisation could reveal further gains. Nevertheless, given the rigorous methodology and consistent evaluation metrics employed, these comparative findings remain valid and actionable despite the limitations.

6.3 *Future Work*

Future work should expand the dataset to a larger, more balanced, multilingual corpus at the paragraph or document level and evaluate models on full sustainability reports. Richer PEFT configurations (e.g., QLoRA with additional adapters or prefix-tuning) could be explored on larger models if resources allow. Data augmentation can be enhanced by combining paraphrasing with techniques like back-translation and masked language modeling. Finally, evaluating models on full sustainability reports and non-English texts will better reflect their practical utility.

7 CONCLUSION

This research addressed the challenge of ESG text classification in low-resource settings by evaluating three approaches: QLoRA fine-tuning, few-shot prompting, and zero-shot model ensembling. The results showed that QLoRA-fine-tuned ESG-LLaMA achieved the highest macro-F1 score, while a simple ensemble of off-the-shelf models also delivered strong performance with less computational demands than ESG-LLaMA and zero labeled dataset. These findings demonstrate that effective ESG classification is possible even with limited labeled data and resources, lowering the barrier for small and resource-constrained organizations to implement ESG screening tools. This contributes to the broader goal of making ESG analysis more accessible, scalable, and inclusive within the field of sustainable finance.

REFERENCES

Aghajanyan, A., Gupta, S., & Zettlemoyer, L. (2021, August). Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 7319–7328). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.568

Ahmad, H., Yaqub, M., & Lee, S. H. (2023). Environmental-, social-, and governance-related factors for business investment and sustainability: A scientometric review of global trends. *Environment, Development and Sustainability*, *26*, 2965–2987. https://doi.org/10.1007/s10668-023-02921-x

AI, M. (2023). Meta llama 2: Model cards and prompt formats [Accessed: 2025-05-07]. https://www.llama.com/docs/model-cards-and-prompt-formats/meta-llama-2/

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. https://arxiv.org/abs/1907.10902

Alliance, G. S. I. (2018). Global sustainable investment review.

Allingham, J. U., Wenzel, F., Mariet, Z. E., Mustafa, B., Puigcerver, J., Houlsby, N., Jerfel, G., Fortuin, V., Lakshminarayanan, B., Snoek, J., Tran, D., Ruiz, C. R., & Jenatton, R. (2023). Sparse moes meet efficient ensembles. https://arxiv.org/abs/2110.03360

Almatrafi, O., & Johri, A. (2025). Leveraging generative ai for course learning outcome categorization using bloom's taxonomy. *Computers and*

Education: Artificial Intelligence, *8*, 100404. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100404

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. https://arxiv.org/abs/1908.10063

Arshad, M. A., Jubery, T. Z., Roy, T., Nassiri, R., Singh, A. K., Singh, A., Hegde, C., Ganapathysubramanian, B., Balu, A., Krishnamurthy, A., & Sarkar, S. (2025). Leveraging vision language models for specialized agricultural tasks. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6320–6329. https://doi.org/10.1109/WACV61041.2025.00616

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 24). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., Gao, H., Gao, K., Gao, W., Ge, R., Guan, K., Guo, D., Guo, J., Hao, G., Hao, Z., . . . Zou, Y. (2024). Deepseek llm: Scaling open-source language models with longtermism. https://arxiv.org/abs/2401.02954

Bird, J. J., Ekárt, A., & Faria, D. R. (2020). Chatbot interaction with artificial intelligence: Human data augmentation with t5 and language transformer ensemble for text classification. https://arxiv.org/abs/2010.05990

Birti, M., Osborne, F., & Maurino, A. (2025). Optimizing large language models for esg activity detection in financial texts. https://arxiv.org/abs/2502.21112

Boehnke, J., Pontikes, E., & Bhargava, H. K. (2024). *Decoding unstructured text: Enhancing llm classification accuracy with redundancy and confidence* (tech. rep.). University of California, Davis. https://d30i16bbj53pdg.cloudfront.net/wp-content/uploads/2024/06/Decoding-Unstructured-Text-Enhancing-LLM-Classification.pdf

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. https://arxiv.org/abs/2005.14165

Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2024). Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. https://arxiv.org/abs/2312.16171

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. *Proceedings of the Twenty-First International Conference on Machine Learning*, 18. https://doi.org/10.1145/1015330.1015432

Chen, T., Xu, B., Zhang, C., & Guestrin, C. (2016). Training deep nets with sublinear memory cost. https://arxiv.org/abs/1604.06174

da Costa-Luis, C., Wolf, T., et al. (2023). tqdm: A Fast, Extensible Progress Bar for Python and CLI [Version X.Y.Z].

de Freitas Netto, S. V., Sobral, M. F. F., Ribeiro, A. R. B., & Soares, G. R. d. L. (2020). Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, *32*(1), 1–12. https://doi.org/10.1186/s12302-020-0300-3

DeepSeek-AI. (2024). Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*. https://github.com/deepseek-ai/DeepSeek-LLM

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. https://arxiv.org/abs/2305.14314

Dettmers, T., & Zettlemoyer, L. (2023). The case for 4-bit precision: K-bit inference scaling laws. *International Conference on Machine Learning*, 7750–7774.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems*, 1–15.

Ding, T., Chen, T., Zhu, H., Jiang, J., Zhong, Y., Zhou, J., Wang, G., Zhu, Z., Zharkov, I., & Liang, L. (2024). The efficiency spectrum of large language models: An algorithmic survey. https://arxiv.org/abs/2312.00678

Gao, L., Zhang, Y., Han, J., & Callan, J. (2021). Scaling deep contrastive learning batch size under memory limited setup. https://arxiv.org/abs/2101.06983

Gema, A. P., Minervini, P., Daines, L., Hope, T., & Alex, B. (2024). Parameter-efficient fine-tuning of llama for the clinical domain. https://arxiv.org/abs/2307.03042

Goswami, K., Islam, M. K. S., & Evers, W. (2023). *A case study on the blended reporting phenomenon: A comparative analysis of voluntary reporting frameworks and standards—gri, ir, sasb, and cdp* (tech. rep.). ACSDRI. https://acsdri.com/wp-content/uploads/2023/09/2023-A-case-study-on-the-blended-reporting-phenomenon-GRI-IR-SASB-CDP-Goswami-Islam-Evers_compressed.pdf

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: An overview. https://arxiv.org/abs/2008.05756

Hansen, L., & Salamon, P. (1990). Neural network ensembles. *I E E E Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001. https://doi.org/10.1109/34.58871

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., et al. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

He, P., Gao, J., & Chen, W. (2023). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. https://arxiv.org/abs/2111.09543

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. https://arxiv.org/abs/2106.09685

Huang, A. H., Wang, H., & Yang, Y. (2022). Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*. https://doi.org/10.1111/1911-3846.12832

Junprung, E. (2023). Sasb-aligned esg sentences [Accessed: 2025-05-03]. https://www.kaggle.com/datasets/edwardjunprung/sasb-aligned-esg-sentences

Khan, M., Serafeim, G., & Yoon, A. (2016). Corporate sustainability: First evidence on materiality. *The Accounting Review*, *91*(6), 1697–1724. https://doi.org/10.2308/accr-51383

Lee, H., Lee, S. H., Park, H., Kim, J. H., & Jung, H. S. (2024). Esg2preem: Automated esg grade assessment framework using pretrained ensemble models. *Heliyon*, *10*(4), e26404. https://doi.org/10.1016/j.heliyon.2024.e26404

Lee, J., & Kim, M. (2023). Esg information extraction with cross-sectoral and multi-source adaptation based on domain-tuned language models. *Expert Systems with Applications*, *221*, 119726. https://doi.org/10.1016/j.eswa.2023.119726

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., ... Wolf, T. (2021, November). Datasets: A community library for natural language processing. In H. Adel & S. Shi (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing: System demonstrations* (pp. 175–184).

Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-demo.21

Liao, L., Li, H., Shang, W., & Ma, L. (2022). An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks. *ACM Trans. Softw. Eng. Methodol.*, *31*(3). https://doi.org/10.1145/3506695

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. https://arxiv.org/abs/2107.13586

Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Yang, H., Sun, Y., Deng, C., Xu, H., Xie, Z., & Ruan, C. (2024). Deepseek-vl: Towards real-world vision-language understanding. https://arxiv.org/abs/2403.05525

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., & Bossan, B. (2022). Peft: State-of-the-art parameter-efficient fine-tuning methods.

Mehra, S., Louka, R., & Zhang, Y. (2022). Esgbert: Language model to help with classification tasks related to companies' environmental, social, and governance practices. *Embedded Systems and Applications*, 183–190. https://doi.org/10.5121/csit.2022.120616

Mishra, S. (2023, November). ESG impact type classification: Leveraging strategic prompt engineering and LLM fine-tuning. In C.-C. Chen, H.-H. Huang, H. Takamura, H.-H. Chen, H. Sakaji, & K. Izumi (Eds.), *Proceedings of the sixth workshop on financial technology and natural language processing* (pp. 72–78). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.finnlp-2.11

Moreira, A., Rodrigues, A. C., & Ferreira, M. R. (2025). Where is human resource management in sustainability reporting? ESG and GRI perspectives. *Sustainability*, *17*(7), 3033. https://doi.org/10.3390/su17073033

Organisation for Economic Co-operation and Development. (2025, February). *Behind esg ratings: Unpacking sustainability metrics*. OECD Publishing. Paris. https://doi.org/10.1787/3f055f0c-en

Palaniappan M, M., Vedhamani, A., & K B, S. (2023). Zero-shot learning for text classification: Extending classifiability beyond conventional techniques. *2023 IEEE Region 10 Symposium (TENSYMP)*, 1–6. https://doi.org/10.1109/TENSYMP55890.2023.10223610

pandas development team, T. (2020, February). *Pandas-dev/pandas: Pandas* (Version latest). Zenodo. https://doi.org/10.5281/zenodo.3509134

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems (NeurIPS)*, *32*, 8024–8035.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*.

Prottasha, N. J., Mahmud, A., Sobuj, M. S. I., Bhat, P., Kowsher, M., Yousefi, N., & Garibay, O. O. (2024, October). *Parameter-efficient fine-tuning of large language models using semantic knowledge tuning*. arXiv: 2410.08598 [cs.CL]. https://arxiv.org/abs/2410.08598

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67. http://jmlr.org/papers/v21/20-074.html

Rane, N., Choudhary, S. P., & Rane, J. (2024). Ensemble deep learning and machine learning: Applications, opportunities, challenges, and future directions. *Studies in Medical and Health Sciences*, *1*(2), 18–41. https://doi.org/10.48185/smhs.v1i2.1225

Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., & Leippold, M. (2024). Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. *Finance Research Letters*, *61*, 104979. https://doi.org/10.1016/j.frl.2024.104979

Sokolov, A., Caverly, K., Mostovoy, J., Fahoum, T., & Seco, L. (2021). Weak supervision and black-litterman for automated esg portfolio construction. *The Journal of Financial Data Science*, *3*(3), 129–138.

Sokolov, A., Mostovoy, J., Ding, J., & Seco, L. (2021). Building machine learning systems for automated esg scoring. *The Journal of Impact and ESG Investing*, *1*, 39–50. https://doi.org/10.3905/jesg.2021.1.010

Sustainability Accounting Standards Board (SASB). (2017). *Esg integration insights: 2017 omnibus edition* (Accessed 2025-04-29). Sustainability Accounting Standards Board. https://www.sasb.org

Tan, C., Yin, K., Wu, H., & Zhou, P. (2025). Analysts' esg attention and stock pricing efficiency: Evidence from machine learning and text analysis. *Journal of Accounting Literature*. https://doi.org/10.1108/jal-06-2024-0128

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Alpaca: A strong, replicable instruction-following model [Retrieved November 6, 2023].

Tian, K., & Chen, H. (2024, May). ESG-GPT:GPT4-based few-shot prompt learning for multi-lingual ESG news text classification. In C.-C. Chen, X. Liu, U. Hahn, A. Nourbakhsh, Z. Ma, C. Smiley, V. Hoste, S. R. Das, M. Li, M. Ghassemi, H.-H. Huang, H. Takamura, &

H.-H. Chen (Eds.), *Proceedings of the joint workshop of the 7th financial technology and natural language processing, the 5th knowledge discovery from unstructured data in financial services, and the 4th workshop on economics and natural language processing* (pp. 279–282). Association for Computational Linguistics. https://aclanthology.org/2024.finnlp-1.31/

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., & Lample, G. (2023). Llama: Open and efficient foundation language models. https://arxiv.org/abs/2302.13971

United Nations Global Compact. (2004). *Who cares wins: Connecting financial markets to a changing world* (tech. rep.). United Nations. https://www.unglobalcompact.org/docs/issues_doc/Financial_markets/who_cares_who_wins.pdf

Wang, S., Zhang, S., Zhang, J., Hu, R., Li, X., Zhang, T., Li, J., Wu, F., Wang, G., & Hovy, E. (2025). Reinforcement learning enhanced llms: A survey. https://arxiv.org/abs/2412.10400

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., . . . Rush, A. M. (2020). Huggingface's transformers: State-of-the-art natural language processing. https://arxiv.org/abs/1910.03771

Xia, L., Yang, M., & Liu, Q. (2024, March). Using pre-trained language model for accurate ESG prediction. In C.-C. Chen, T. Ishigaki, H. Takamura, A. Murai, S. Nishino, H.-H. Huang, & H.-H. Chen (Eds.), *Proceedings of the eighth financial technology and natural language processing and the 1st agent ai for scenario planning* (pp. 1–22). -. https://aclanthology.org/2024.finnlp-2.1/

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *Proceedings of EMNLP-IJCNLP 2019*, 3914–3923. https://doi.org/10.18653/v1/D19-1404

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms* (1st). Chapman; Hall/CRC. https://doi.org/10.1201/b12207

APPENDIX A

Below are the complete classification reports for each individual model in the ensemble.

```
Classification Report for FINBERT-ESG:

              precision    recall  f1-score   support

 Environment      0.704     0.894     0.788        85
  Governance      0.724     0.420     0.532       150
      Social      0.700     0.837     0.762       203

    accuracy                          0.705       438
   macro avg      0.709     0.717     0.694       438
weighted avg      0.709     0.705     0.688       438
```

```
Classification Report for ESG-RoBERTa:

              precision    recall  f1-score   support

 Environment      0.630     0.882     0.735        85
  Governance      0.451     0.640     0.529       150
      Social      0.811     0.424     0.557       203

    accuracy                          0.587       438
   macro avg      0.631     0.649     0.607       438
weighted avg      0.653     0.587     0.582       438
```

```
Classification Report for DeBERTa_v3_MNLI:

              precision    recall  f1-score   support

 Environment      0.581     0.929     0.715        85
  Governance      0.497     0.500     0.498       150
      Social      0.715     0.532     0.610       203

    accuracy                          0.598       438
   macro avg      0.598     0.654     0.608       438
weighted avg      0.614     0.598     0.592       438
```

APPENDIX B

The following prompt was initially tested for zero-shot ESG-LLaMA . However, it did not produce reliable results.

The full prompt and classification report are shown below:

```
    "You are an expert in classifying ESG data.
    You will start your response with 'Label:'.\n"
    "Classify the following text into one of
    the four ESG categories, choose an answer from "
    "{Environment, Social, Governance}\n"
    f"Text: {text}\n"
    "Label: So, the answer is"
```

```
              precision    recall  f1-score   support

 Environment       0.45      0.06      0.10        85
      Social       0.49      0.90      0.64       203
  Governance       0.55      0.21      0.31       150

    accuracy                           0.50       438
   macro avg       0.50      0.39      0.35       438
weighted avg       0.51      0.50      0.42       438
```

The following prompt was initially tested for few-shot with DeepSeek 7B-chat model. However, due to the poor performance, it was not used in the main setup. Below are the exact prompt and classification report.

```
    Classify the ESG category of the following text.
    Text: [example]
    Options: Environment, Social, Governance
    Answer: [label]
```

```
Macro F1-Score: 0.5296
Micro F1-Score: 0.5884
Weighted F1-Score: 0.5474


              precision    recall  f1-score   support

 Environment       0.43      0.91      0.58        85
      Social       0.74      0.74      0.74       203
  Governance       0.68      0.17      0.27       150

   micro avg       0.60      0.58      0.59       438
   macro avg       0.61      0.61      0.53       438
weighted avg       0.66      0.58      0.55       438
```

APPENDIX C

- Hugging Face Transformers 4.51.3(Wolf et al., 2020)

- Datasets 3.5.0 (Lhoest et al., 2021)
- Optuna 4.1.0 (Akiba et al., 2019)
- PEFT (LoRA and QLoRA fine-tuning) 0.15.2 (Mangrulkar et al., 2022)
- PyTorch 2.6.0 (Paszke et al., 2019)
- Scikit-learn 1.5.1 (Pedregosa et al., 2011)
- NumPy 1.26.4 (Harris et al., 2020)
- Pandas 2.2.2 (pandas development team, 2020)
- Seaborn 0.13.2 (Waskom, 2021)
- tqdm 4.66.5 (da Costa-Luis, Wolf, et al., 2023)
- bitsandbytes (Dettmers et al., 2022)
- python 3.12