Adapting LLM to Multi-lingual ESG Impact and Length Prediction using In-context Learning and Fine-Tuning with Rationale

Pawan Kumar Rajpoot, Ashvini Jindal, Ankur Parikh
SCB DataX Thailand, LinkedIn AI USA, UtilizeAI Research India

Abstract

The prediction of Environmental, Social, and Governance (ESG) impact and duration (length) of impact from company events, as reported in news articles, hold immense significance for investors, policymakers, and various stakeholders. In this paper, we describe solutions from our team "Upaya" to ESG impact and length prediction tasks on one such dataset ML-ESG-3. We employed two different paradigms to adapt Large Language Models (LLMs) to predict both ESG impact level and length of events. In the first approach, we leverage GPT-4 within the In-context learning (ICL) framework where a retriever identifies top K-relevant in-context learning examples for a given test example. The second approach involves instruction-tuning Mistral (7B) LLM to predict impact level and duration, supplemented with rationale generated using GPT-4. Our models secured second place in both French tasks where for one task fine-tuned Mistral model outperformed and for other task, GPT-4 with ICL outperformed. These results demonstrate the potential of different LLM-based paradigms for delivering valuable insights within the ESG investing landscape.

Keywords: ESG Impact, ESG Length, Large Language Models, FinNLP, Q-LoRA, In-Context Learning, Rationale Generation, Chain of Thoughts

1. Introduction

Environmental, Social, and Corporate Governance (ESG) factors have become pivotal in assessing the long-term sustainability and ethical impact of businesses, investments, and policy decisions. The integration of ESG criteria in investment strategies aims to mitigate risks, identify opportunities aligned with responsible practices, and foster positive change.

The advent of large language models (LLMs), exemplified by GPT-4 (Brown et al., 2020) (Thoppilan et al., 2022), marks a significant breakthrough in natural language processing (NLP). These models exhibit proficiency across various domains and can be readily applied to multiple NLP tasks. Traditionally, language models follow distinct pre-training and fine-tuning pipelines (Devlin et al., 2018) (Beltagy et al., 2019) (Raffel et al., 2020) (Lan et al., 2019) (Liu et al., 2021b), where fine-tuning occurs after pre-training on task-specific datasets in a fully-supervised manner.

A recent paradigm, In-context Learning (ICL) (Brown et al., 2020) (Thoppilan et al., 2022), reshapes NLP tasks, enabling LLMs to make predictions by learning from demonstrations presented within the context prompt. Under the ICL framework, LLMs achieve remarkable performance, rivaling fully-supervised methods, even with a limited number of demonstrations. The retrieval of contextually relevant examples plays a crucial role in overall performance, as LLMs benefit from examples similar to the "to be predicted" data point, reducing hallucination and improving performance.

This paper explores two approaches within the ML-ESG-3 dataset for English and French datasets: 1. Guiding GPT-4 under the ICL framework to predict ESG impact and event duration, using a learning-free dense retriever to identify top K relevant In-context learning examples. 2. Instruction-tuning the open-source LLM, Mistral, with 7B parameters to predict ESG impact and duration, incorporating rationale. Efficient fine-tuning is achieved through Parameter Efficient Fine Tuning (PEFT), specifically QLoRA 4-bits quantization.

2. Preliminary Background

2.1. Task Definition

As per the challenge "ESG Impact Level and Length Prediction" (Chen et al., 2024) is the task of automatically determining the ESG impact level - opportunity or risk and the duration (length) of the impact an event in the news article might have on the company". This shared task is a part of the Fifth Workshop on Knowledge Discovery from Unstructured Data in Financial Services, co-located with LREC-COLING 2024.

Let x denote the news article. Given a set of predefined impact level classes, Level=*Low*, *Medium*, *High* and a set of predefined impact length classes, Length=*Short-Term*, *Medium-Term*, *Long-Term*, the task aims to predict the class c_1 in level and c_2 in length for input x.

2.2. Data

The English dataset released with this task contains 545 train and 136 test (evaluation) instances. While the French dataset had 661 training examples and 146 test (evaluation) examples.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 274–278
20 May, 2024.

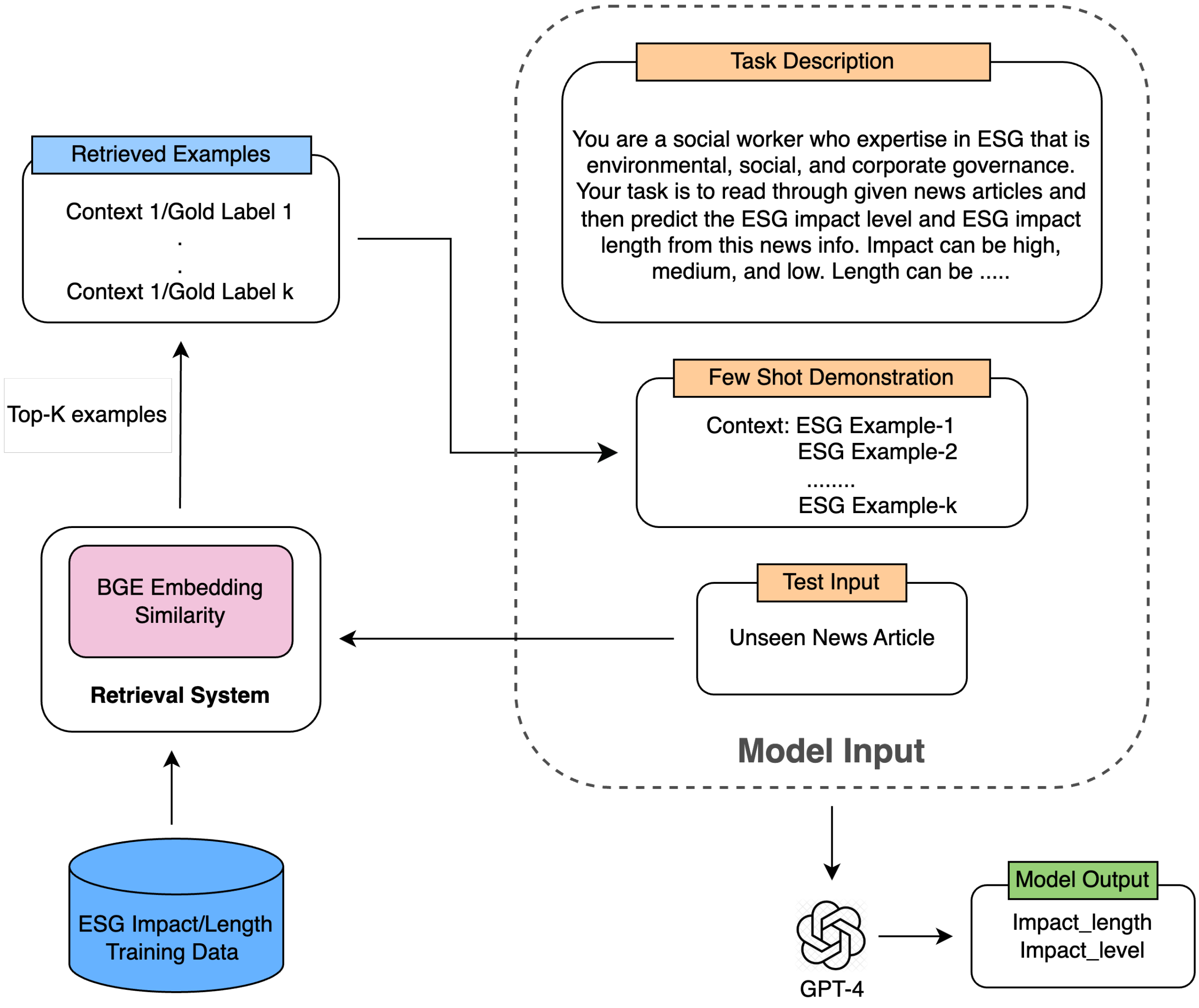
Task Description
You are a social worker who expertise in ESG that is environmental, social, and corporate governance. Your task is to read through given news articles and then predict the ESG impact level and ESG impact length from this news info. Impact can be high, medium, and low. Length can be .....
Retrieved Examples
Context 1/Gold Label 1
.
.
Context 1/Gold Label k
Top-K examples
Few Shot Demonstration
Context: ESG Example-1
ESG Example-2
....
ESG Example-k
BGE Embedding Similarity
Retrieval System
Test Input
Unseen News Article
Model Input
ESG Impact/Length Training Data
GPT-4
Model Output
Impact_length
Impact_level

Figure 1: Approach 1: In-context Learning

2.3. In-context Learning

In-context learning (ICL) is a key emergent ability of language models (Wei et al., 2023), allowing them to infer tasks from context. Unlike gradient-based 'in-weights learning' (which updates model parameters), ICL is gradient-free, adapting directly from the context ((Brown et al., 2020). Formally, each training instance is first linearized into an input text x and an output text y. Given a test input text x_{test}, in-context learning defines the generation of output y as $y_{test} \sim PLM\left(y_{test}|x_1, y_1, \ldots x_k, y_k, x_{test}\right)$, where k refers to number of in-context examples and $\sim$ refers to decoding strategies(e.g., greedy decoding and nuclear sampling (Li et al., 2022)), and each in-context example $e_i = (x_i, y_i)$ is sampled from a training set D. The generation procedure is especially attractive as it eliminates the need for updating the parameters of the language model when encountering a new task, which is often expensive and impractical. Notably, the performance of ICL on downstream tasks can vary from almost random to comparable with state-of-the-art systems, depending on the quality of the retrieved in-context examples (Rubin et al., 2021) (Liu et al., 2021a).

3. Adapting LLM for ESG Impact Level and Length Prediction

We employed two paradigms to adapt LLMs for the specific task. 1. In-context Learning and 2. Instruction Fine-Tuning.

3.1. In-context Learning

The formalization of the task under the ICL framework, using GPT-4 is shown in figure 1.

3.1.1. Prompt Construction

For each test example, a prompt is meticulously constructed and subsequently input to GPT-4. The prompt encompasses the following key components:

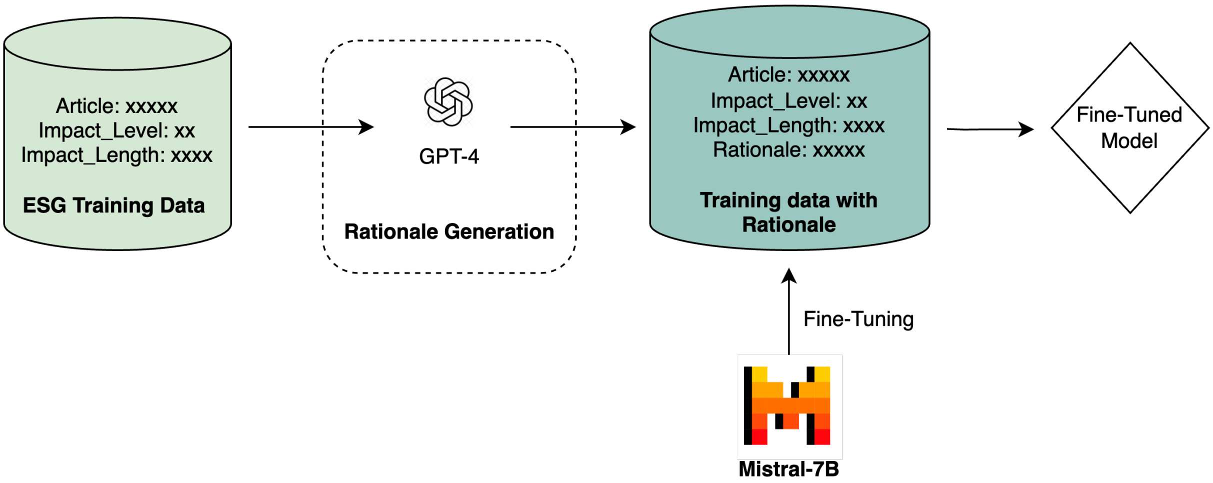

Figure 2: Approach 2: Instruction Fine-tuning

- **Task Description** - A concise overview is presented, outlining the task and the predefined classes for impact level and length prediction tasks.
- **K-shot Demonstrations** - The demonstration section involves the reformulation of each example, displaying the input x_{demo} and corresponding class labels $c1_{demo}$ and $c2_{demo}$ where c_1 refer to level and c_2 refers to length.
- **Test Input** - Test input x_{test} is provided, and GPT model is tasked with generating the corresponding class $c1_{test}$ and $c2_{test}$

3.1.2. Retrieval with Cosine Similarity over BGE Embedding

In-context learning (ICL) using demonstrations that are closer to the test sample within the embedding space tend to perform better than random selection. We used cosine similarity to find the most relevant examples from the training set. To represent these examples, we used BGE Embeddings (Xiao et al., 2023), and FAISS for fast similarity search (Johnson et al., 2019)

3.2. Instruction Fine-Tuning with Rationale

The formalization of the task under the Instruction Fine-Tuning with Rationale framework, using Mistral Base Model (7B) is shown in figure 2.

3.2.1. Rationale Generation

Since we are using relatively smaller model (7B) for fine-tuning, we employed Chain-of-Thought (CoT) based paradigm. Instead of directly generating only label, the LLM should generate both rationale and label. To generate both rationale and labels as output, the system needs to be fine-tuned with rationale as part of the output. Since the rationale behind labels wasn't available in the training dataset, we used GPT-4 to generate the rationale for each training data sample. Refer to section 8 for the description used in the prompt and generated rationale. For the English task in the dataset, this approach worked well. For the French language, we translated training data from French to English using GPT-4 and then used the same methodology to generate rationale.

3.2.2. Instruction Fine-Tuning

We fine-tuned the Mistral-7B base model using an English training set with GPT-4 rationale, plus a French set (translated to English) with GPT-4 rationale. Due to memory limits, we used 4-bit QLoRA (Dettmers et al., 2024) with rank 128 and alpha 256. Quantized LoRA was applied to self-attention Query, Key, Value matrices and Linear layers. We used gradient accumulation (steps=2), paged Adamw 32bit optimizer, cosine schedule (LR=2e-5), decay rate 0.01, and 5 warmup steps. Fine-tuning was done using axolotl [1]

4. Experiments and Results

Maximum of 3 submissions were allowed for each language subtask. We submitted for both English and French subtasks as shown in Table 1. Specifically, we submitted one entry with instruction-tuned Mistral model and two entries with ICL with different values of K (number of demonstrations retrieved).

[1] https://github.com/OpenAccess-AI-Collective/axolotl

Submission	Language	Approach
E1	English	Fine-tune
E2	English	ICL 10-shot
E3	English	ICL 20-shot
F1	French	Fine-tune
F2	French	ICL 10-shot
F3	French	ICL 20-shot

Table 1: Our Submission details

4.1. Impact Level

Table 2 shows results for Impact Level prediction task. For English language, our fine-tuned Mistral 7B based model outperformed GPT-4 with K-shot learning. For French language, the performance of fine-tuned Mistral and GPT-4 with 20-shot is comparable.

Submission	Micro-F1	Macro-F1
E1	54.41	48.40
E2	53.68	45.93
E3	51.47	46.09
F1	58.22	56.78
F2	58.22	56.69
F3	42.47	37.64

Table 2: Overall scores on Impact level prediction

4.2. Impact Length

Table 3 shows results for Impact Length prediction task. For English language, GPT-4 with 20-shot learning performs better than fine-tuned Mistral. However, for the French language, GPT-4 with 10-shot learning performs better than the fine-tuned Mistral model. In summary, for the Impact Level task, fine-tuned Mistral model outperformed GPT-4 with ICL. However, for Impact Length task, GPT-4 with ICL outperformed fine-tuned Mistral model.

Submission	Micro-F1	Macro-F1
E1	57.35	42.75
E2	51.47	38.55
E3	60.29	44.23
F1	46.58	42.86
F2	52.05	48.73
F3	41.10	32.09

Table 3: Overall scores on Impact length prediction

Overall, in the context of the shared-task, for the Impact Length prediction (French Language), our submission F2 got 2nd rank. For the Impact Level prediction (French Language), our submission F1 got 2nd rank. For the Impact Length prediction (English Language), our submission E3 got 7th rank and for the Impact Level prediction (English Language), our submission E1 got 17th rank.

5. Conclusion

This work explores the potential of GPT + ICL and Mistral (7B) + Fine-Tuning with Rationale on ESG Impact Level and Length Prediction task. For Impact Level prediction, the fine-tuned model performed better. Form Impact Length prediction, the GPT + ICL combination performed better. We achieved 2nd rank in both sub-tasks for French Langauge.

6. Acknowledgements

We thank the organizing committee of FinNLP-2024 along with the task-setting team of Shared Task ML-ESG-3 for giving us this opportunity to work on this problem. We also would like to thank Lambda Labs[2] for providing the compute resources required for this work.

7. Bibliographical References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Hanwool Lee, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multi-lingual esg impact duration inference. In *Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing and the 5th Knowledge Discovery from Unstructured Data in Financial Services*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

[2] https://lambdalabs.com/

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021b. A robustly optimized bert pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*, pages 471–484. Springer.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

8. Appendix

8.0.1. Rationale Generation

GPT4 Prompt for Rationale Generation

You are an annotator who has already annotated ESG IMPACT_LEVEL and IMPACT_LENGTH on news articles. Now you have to justify your annotations for each data point, why you gave the annotation? Give an answer in minimum words.

Article: High risk of strain on the power grid: an energy-by-energy status report The grid operator RTE has updated its forecasts for the winter. While these forecasts now predict a fall in electricity consumption, they also anticipate a drop in nuclear electricity generation. This should almost balance out. Despite everything, RTE warns that the risk of a power cut cannot be ruled out, particularly in January. Here is a status report on each energy source.
impact_level: high
impact_length: More than 5 years

GPT4 Rationale Generation Output

RATIOANLE_IMPACT_LEVEL: high - Risk of power cuts affects national energy security, economy, public safety, and has cascading effects on all sectors.
RATIOANLE_IMPACT_LENGTH: More than 5 years - Energy infrastructure and generation capacity issues are complex and typically take many years to address and improve.