Luo, Y., Yang, T., Li, Q., Liu, Q. & Cui, X. (2024). Unmasking ESG exaggerations using generative artificial intelligence. In Li, E.Y. *et al.* (Eds.) *Proceedings of The International Conference on Electronic Business, Volume 23* (pp. 684-689). ICEB'24, Zhuhai, China, October 24-28, 2024

Unmasking ESG Exaggerations Using Generative Artificial Intelligence

Yunfang Luo [1,*]
Tao Yang [2]
Qingan Li [3]
Qiang Liu [4,*]
Xiling Cui [5]

*Corresponding author
[1] Master, Beijing Institute of Petrochemical Technology, Beijing, China, 2910470660@qq.com
[2] Master, Beijing Institute of Petrochemical Technology, Beijing, China, 1793795143@qq.com
[3] Master, Hong Kong Baptist University, Hong Kong, China, 23447214@life.hkbu.edu.hk
[4] Professor, Beijing Institute of Petrochemical Technology, Beijing, China, liuq@bipt.edu.cn
[5] Adjunct Professor, Beijing Institute of Petrochemical Technology, Beijing, China, cuixiling@gmail.com

ABSTRACT

Exaggeration is a major indicator of greenwashing, typified by excessively optimistic or idealistic portrayals for environmental protection. The purpose of this study is to identify exaggerated information in environmental, social and governance (ESG) reports by using generative artificial intelligence (GenAI). We analyze a collection of ESG reports using three prompt engineering strategies: few-shot, zero-shot, and chain of thought (COT). We also cross-validate our results using traditional text analytics and human intelligence. Using this strategy, we evaluate exaggeration in ESG reports in a novel way using GenAI. The use of GenAI creates a strong foundation for further study in these and related fields.

Keywords: GenAI, ESG, exaggerated information.

INTRODUCTION

With the advent of recent years, the financial sector has encountered a pervasive challenge known as greenwashing. Greenwashing is the practice of exaggerating or misrepresenting a company's environmental effect in order to give the impression that it is more environmentally conscious than it actually is (Huang & Chen, 2015). It occurs when formal public information like environmental, social and governance (ESG) reports, exacerbate information asymmetry and increasing the risk of market disruptions (Liu et al., 2024). Exaggeration, a common form of greenwashing, has been shown to hinder sustainable development goals by distorting their achievement and assessment (Cojoianu et al., 2020). This often manifests as overly positive language used to describe a company's environmental, social, or governance performance, without providing sufficiently supporting data or evidence. Therefore, detecting and addressing the exaggeration in ESG reports is critical.

The world is being shaped by artificial intelligence (AI), especially with the rapid development of generative AI (GenAI) like ChatGPT. Some studies in finance field have started to leverage AI to tackle the problem in ESG reports. For example, some research compares the traditional and AI-driven ESG ratings (Hughes et al., 2021). Some study investigates the AI's impact on greenwashing and sustainability reporting (Moodaley & Telukdarie, 2023). Yang et al. (2021) found that ESG disclosure reduces corporate bond credit spreads, lowering risk and boosting investor confidence, while Biju et al. (2023) link the sentiment scores on ESG to perceptions of greenwashing using MAXQDA software. These studies indicate the possibility and potential of applying AI in analyzing ESG reports. This study is thus inspired to make good use of AI, especially GenAI, for the evaluation of exaggeration in ESG reports (Jain et al., 2023).

Although prior studies have examined diverse techniques for scrutinizing and assessing ESG reports, the difficulty of identifying exaggerated assertions is still inadequately investigated, especially when utilizing state-of-the-art artificial intelligence technology. Furthermore, even while companies may use adjectives like "extreme," "complete," or "supreme" in their ESG reports, this does not always mean that they are overstating their accomplishments; in fact, some companies may excel in this area. Therefore, it is not possible to determine whether corporations are overstating their claims in a thorough manner by only relying on the terminology found in ESG reports. GenAI has the advantage of being able to discern whether there are reasonable reasons for suspicion regarding exaggeration by analyzing contextual details. This study aims to address this gap by leveraging GenAI to detect overstated descriptions in ESG reports, contributing to a more accurate and robust evaluation of corporate sustainability performance. Three different prompt engineering strategies, namely, zero-shot, few-shot, and chain of thought (COT), are applied to analyze a set of ESG reports. Furthermore, we also cross-validate the method with the traditional text analytics technique and human intelligence. This

study pioneers the investigation of the exaggeration in ESG reports using GenAI. The cross-validation from other methods provides solid prove for the research findings. The application of GenAI also lays a good basis for further studies in this area and similar areas. Our research design contributes significantly to future studies in this subject and strengthens the persuasiveness of our conclusions. To sum up, our study presents a fresh angle on how to interpret and evaluate corporate ESG reports, offering a more thorough and detailed approach to evaluating the real environmental, social, and governance performance of a business.

LITERATURE REVIEW

GenAI in Finance Research

In the financial sector, the application of GenAI has become increasingly common. The following studies and systems showcase the latest advancements and methods in this field. For example, CHATREPORT has been developed as a sophisticated system to analyze corporate sustainability reports automatically. Its two primary functions are to integrate expert judgment to handle complicated situations and ensure response traceability to reduce communication errors. The accuracy and transparency of sustainability report analysis are improved by using this system (Ni et al., 2023).

Another study developed a revolutionary tool called ESGRevel to gather and analyze ESG data from business reports in a methodical manner. Report preprocessing, large language model (LLM) agent, and ESG metadata are its three primary parts that integrate LLMs with Retrieval-Augmented Generation (RAG) technology. The accuracy and effectiveness of examining ESG reports are improved by this design (Zou et al., 2023). Moyer (2023) examined the weaknesses and remedies found in Corporate Social Responsibility (CSR) reports pertaining to greenwashing. The effectiveness of CSR reports as ESG data sources was assessed by the study using BART large Multi-Genre Natural Language Inference (MNLI) models, which matched report themes with dimensions from well-known ESG grading systems such as Sustainalytics and Morgan Stanley Capital International (MSCI).

Additionally, Jain et al. (2023) studied the application of GPT-3.5 under ESG standards, highlighting its potential in the investment industry. The researchers created an ESG classifier module that can identify a company's ESG characteristics, helping investors make value-driven decisions. Kim et al. (2024) assessed GPT-3.5's capability to extract complex information from corporate disclosures. They found that while these models improve content by condensing information, they may also exaggerate information. Cao and Zhai (2023) investigated how well GPT-4 performed in analyzing company culture, sentiment, ESG, and Federal Reserve announcements. They found that GPT-4 performed exceptionally well in creating lists of pertinent keywords.

Furthermore, Yang et al. (2023) developed InvestLM, a novel large language model specifically for the financial sector. Comparative testing with GPT-4 demonstrated its superior performance in understanding financial texts. Guo et al. (2023) introduced the FinLMEval framework, which uses four specialized datasets to evaluate financial sentiment, ESG classification, forward-looking statements, and question-answering tasks. Td (2023) emphasized the advantages of generative large language models (GLLMs) in text data-driven accounting research, providing practical resources to help researchers maximize the use of these models. Föhr et al. (2023) explored the potential of integrating AI-based models, particularly GPT, into the ESG report auditing process. By utilizing GPT models to analyze report content and assess its consistency with the EU taxonomy, they enhance the efficiency and depth of the auditing process.

METHOD

This study utilizes GPT-4o to identify the exaggerations in ESG reports, thereby enhancing our capability to detect inflated information within them. To visually illustrate our research methodology more effectively, we provide a detailed depiction in Figure 1.

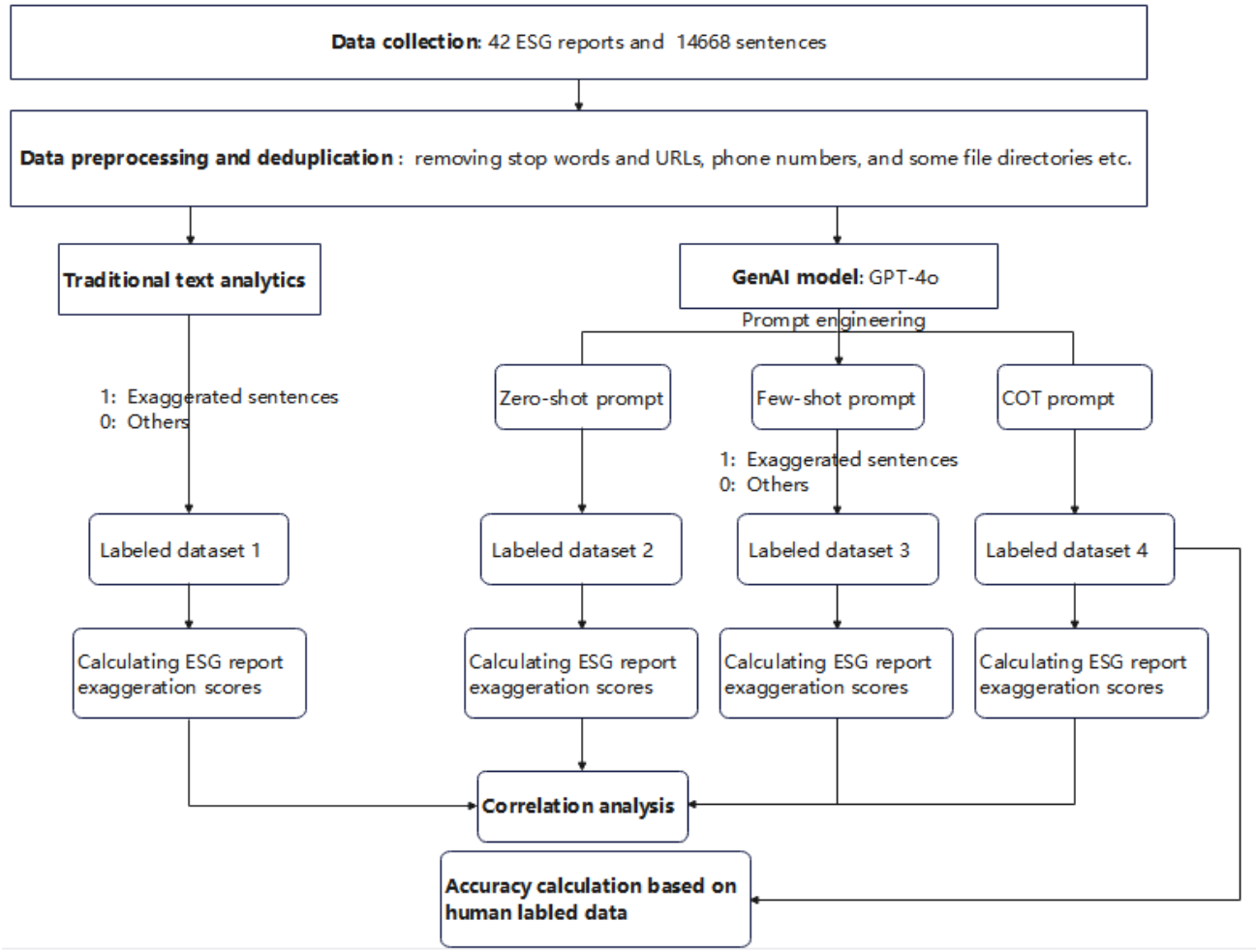

Figure 1: Research Methodology Flowchart

Data Collection

The data was collected by a Python script, crawling the ESG reports of all Chinese listed companies from CNINFO (http://www.cninfo.com.cn). A window of August 2023 to January 2024 was selected. Finally, we obtained 42 ESG reports with 14,668 sentences and they were converted from PDF to text format, stored into a text file for further analysis. To ensure data accuracy, we performed thorough data preprocessing and deduplication, such as removing stop words, deleting URLs, phone numbers, and some file directories etc. The cleaned data was then used as our research dataset.

Traditional Text Analytics

First, we constructed a lexicon of exaggeration words by integrating sentiment lexicons found in Chinese financial literature (Jiang et al., 2019), the CNKI dictionary, the Dalian University of Technology sentiment dictionary (DLUT), and the Chinese Financial Sentiment Dictionary authored by Bian Shibo from Shanghai University of Finance and Economics. This lexicon includes terms such as “extreme,” “complete,” and “supreme” and so on. We classified sentences with exaggeration words as exaggerated and those without as non-exaggerated using this exaggeration lexicon.

GenAI Methods

The GenAI used in this study is GPT-4o, one of the most advanced large language models recently. The GPT-4o model was configured with a temperature setting of 0.01, a top-p value of 0.01, and a max tokens limit of 4095. Three prompt engineering strategies are selected, namely, zero-shot, few-shot and chain-of-thought (COT) because they are the most commonly used prompts. In addition, each of them offers distinct advantages in leveraging the model’s capabilities (Wei et al., 2022):

- Zero-shot prompting refers to performing a task without providing the model with any examples. It can demonstrate the model’s ability to generalize from its training data to new, unseen tasks without additional examples, showcasing its inherent understanding and flexibility.
- Few-shot prompting involves providing the model with some examples to help it better understand and execute the task. It can enhance the model’s performance by providing a small number of examples, which helps it better grasp the specific nuances of the task and improves its contextual comprehension.
- Chain-of-thought (COT) prompting is a method that guides the model to think through the problem step-by-step. It facilitates a structured reasoning process, allowing the model to tackle more complex problems by breaking them down into manageable steps, leading to more accurate and coherent outputs.

Although not as sophisticated as COT, recent research has shown that zero-shot and few-shot prompting can effectively extract clinical information (Sivarajkumar et al., 2024). Therefore, we keep all the three prompts engineering methods in the experiment. The specific details are shown in Table 1. GPT-4o labels the exaggerated sentences as "1" and others as "0".

Table 1: Prompt Engineering Methods

Zero-shot prompting	
The model predicts the answer given only a description of the task.	
You are tasked with detecting exaggerated sentences.	Task Description
Here is the ESG reports you need to analyze:	Prompt
{"role": "user", "context": "{ESG Document Context}"}	
Few-shot prompting	
In addition to the task description, we add some sample input/output pair for model to do the in sample learning.	
You are tasked with detecting exaggerated sentences.	Task Description
Here are some samples you can refer:	Sample
{"role": "user", "context": "{Sample Input 1}"}	
{"role": "user", "context": "{Sample Return 1}"}	
{"role": "user", "context": "{Sample Input 2}"}	
{"role": "user", "context": "{Sample Return 2}"}	
…… (Additional Sample Pair)	
Here is the ESG reports you need to analyze:	Prompt
{"role": "user", "context": "{ESG Document Context}"}	
Zero-shot-COT prompting	
The model is guided to generate intermediate reasoning steps before arriving at the final answer.	
You are tasked with detecting exaggerated sentences.	Task Description
To complete this task, follow these steps:	Chain of Thought
1. Carefully read through the entire ESG report.	
2. As you read, look for sentences that may be exaggerated.	
3. …… (Remain Steps Describe)	
Here is the ESG reports you need to analyze:	Prompt
{"role": "user", "context": "{ESG Document Context}"}	

Calculation of Exaggeration Scores

The calculation of exaggeration scores has two steps. First, we calculated the exaggeration score of each sentence identified by the above four methods, respectively. We use the custom-built lexicon of exaggeration words we built to quantify the degree of exaggeration using the formula: Sentence exaggeration score = Number of exaggeration words in the sentence / Sentence length. Then, we calculated the exaggeration score of each report based the four solutions separately by adding the exaggerating scores of the selected sentences generated by the four methods in that report.

Correlation Analysis and Accuracy Calculation

We employed SPSS tools to perform a correlation analysis on the exaggeration scores for all the EGS reports. We examined the relationships between the scores generated by the following techniques: traditional text analytics, and GenAI (including zero-shot prompting, few-shot prompting, and COT prompting). The direction and intensity of the correlations between these methods were ascertained by calculating the Pearson correlation coefficients. To further check the accuracy of the each prompting strategy of GenAI, we used a human labeled dataset in the training set as the benchmark. The corresponding text analytics and GPT-4o labels were introduced in this dataset and compared with the human labels. The accuracy was calculated as the number of the same results as the human labels divided by the number of all the sentences.

RESULTS

As shown in the Table 2, all methods exhibit significant positive correlations with each other. The correlations between traditional text analytics scores and the three types of GPT-4o prompting are all high, with the correlation coefficients 0.599, 0.668 and 0.579 respectively, indicating a high degree of agreement between the GPT-4o techniques and conventional text analytics methods in terms of recognizing exaggerated sentences. The correlation between the three prompting methods of GPT-4o are also significant, showing the consistency of the three strategies.

Table 2: The Correlation Coefficients of Exaggerated Scores

	Accuracy	Traditional text analytics scores	Zero-shot scores	Few-shot scores	COT scores
Traditional text analytics scores	41.61%	1			
Zero-shot scores	70.77%	0.599**	1		
Few-shot scores	60.56%	0.668**	0.763**	1	
COT scores	77.26%	0.579**	0.506**	0.532**	1

(Note: * * indicates at the $p<0.01$ level)

We also calculate the accuracy of each methods using a human labeled data set. The results were 41.61%, 70.77%, 60.56%, and 77.26%, respectively. According to our research, the traditional text analytics has low accuracy on identifying exaggerated sentences. Among the GPT-4o strategies, COT prompt technique shows better consistency with human-labeled data and is more accurate at spotting exaggerated statements. This highlights the important benefit of COT prompt in these kinds of semantic analysis tasks.

FUTURE WORK AND EXPECTED CONTRIBUTIONS

Current Limitation and Future Work

This work-in-progress is a valuable starting point, but further research is needed to address several limitations and refine the methodology.

First, the current dataset lacks diversity with only ESG reports from Chinese listed companies. To enhance the study's representativeness and universality, future research will consider enriching the dataset by including ESG reports from businesses across different industries and geographical areas of the globe. The enlargement of the research scope will make the research findings more generalizable.

Second, the current prompt engineering methods may not capture all nuances of the data. The prompts we used can be further optimized in future studies, by combining some of the prompt strategies. For example, researchers have found that zero-shot chain-of-thought (zero-shot COT) that combines the principles of zero-shot prompting and COT can achieve superior results (Kojima et al., 2022). By refining the prompt engineering techniques, we can improve the accuracy and reliability of identifying exaggerated statements, providing more effective tools to the practice.

Finally, a package will be developed based on the optimal prompt strategy, ensuring the feasibility of our research findings. In recent years, there is a call for user-friendly AI tools for the practice (Virvou, 2023). This study will be a response to such a call to meet the needs from the industry.

Contributions

This work-in-progress pioneers the investigation of applying GenAI to identifying the exaggeration in ESG reports. All the three prompt engineering strategies, i.e., zero-shot, few-shot and COT in GPT-4o, show promise for enhancing the identification of overstated information, providing adaptable and efficient solutions to the conventional techniques.

Moreover, this study highlights significant differences and advantages over traditional methods. Traditional ESG reporting analysis methods often rely on predefined rules or statistical models, which may have limitations in the face of complex and diverse textual data. The GenAI method proposed in this study demonstrates flexibility and adaptability when processing large-scale text data with high uncertainty. This study proposes a novel prompt engineering strategy, enriching the analytical landscape and offering a new theoretical framework for combining AI technology with traditional methods.

Furthermore, this study is expected to open a new line of study in the analysis of ESG reports. It can also contribute to the GenAI application research by using the real data from the field. Therefore, the study will be further developed and performed to get more research findings to make more contributions to literature and practice.

ACKNOWLEDGMENT

This work is partially supported by the Beijing Municipal Education Commission (Grant No. 22019821001) in China.

REFERENCES

Biju, A. V. N., Kodiyatt, S. J., Krishna, P. N., & Sreelekshmi, G. (2023). ESG sentiments and divergent ESG scores: Suggesting a framework for ESG rating. *SN Business & Economics*, *3*(12), 209. https://doi.org/10.1007/s43546-023-00592-4

Cao, Y., & Zhai, J. (2023). Bridging the gap–The impact of ChatGPT on financial research. *Journal of Chinese Economic and Business Studies*, *21*(2), 177-191. https://doi.org/10.1080/14765284.2023.2212434

Cojoianu, T., Hoepner, A. G., Ifrim, G., & Lin, Y. (2020). GreenWatch-shing: Using AI to detect greenwashing. *AccountancyPlus-CPA Ireland*. Retrieved from https://ssrn.com/abstract=3627157

Föhr, T. L., Schreyer, M., Juppe, T. A., & Marten, K. U. (2023). Assuring sustainable futures: Auditing sustainability reports using AI foundation models. *Available at SSRN 4502549*. https://doi.org/10.2139/ssrn.4502549

Guo, Y., Xu, Z., & Yang, Y. (2023). Is ChatGPT a financial expert? Evaluating language models on financial natural language processing. *arXiv preprint arXiv:2310.12664*. https://doi.org/10.48550/arXiv.2310.12664

Huang, R., & Chen, D. (2015). Does environmental information disclosure benefit waste discharge reduction? Evidence from China. *Journal of Business Ethics*, *129*, 535-552. https://doi.org/10.1007/s10551-014-2173-0

Hughes, A., Urban, M. A., & Wójcik, D. (2021). Alternative ESG ratings: How technological innovation is reshaping sustainable investment. *Sustainability*, *13*(6), 3551. https://doi.org/10.3390/su13063551

Jain, Y., Gupta, S., Yalciner, S., Joglekar, Y. N., Khetan, P., & Zhang, T. (2023). Overcoming complexity in ESG investing: The role of generative AI integration in identifying contextual ESG factors. *Available at SSRN*. http://dx.doi.org/10.2139/ssrn.4495647

Jiang, F., Lee, J., Martin, X., & Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, *132*(1), 126-149. https://doi:10.1016/j.jfineco.2018.10.001

Kim, A., Muhn, M., & Nikolaev, V. V. (2024). Bloated disclosures: Can ChatGPT help investors process information? *Chicago Booth Research Paper*, (23-07), 2023-59. http://dx.doi.org/10.2139/ssrn.4425527

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199-22213. https://doi.org/10.48550/arXiv.2205.11916

Liu, G., Qian, H., Shi, Y., Zhang, Y., & Wu, F. (2024). Does ESG report greenwashing increase stock price crash risk?. *China Journal of Accounting Studies*, 1-25. https://doi.org/10.1080/21697213.2024.2303070

Moodaley, W., & Telukdarie, A. (2023). Greenwashing, sustainability reporting, and artificial intelligence: A systematic literature review. *Sustainability*, *15*(2), 1481. https://doi.org/10.3390/su15021481

Moyer, J. (2023). ESG metric variance and methodology standardization. *Honors Theses andCapstones*, 754. Retrieved from https://scholars.unh.edu/honors/754/

Ni, J., Bingler, J., Colesanti-Senni, C., Kraus, M., Gostlow, G., Schimanski, T., ... & Leippold, M. (2023). ChatReport: Democratizing sustainability disclosure analysis through LLM-based tools. *arXiv preprint arXiv:2307.15770*. https://doi.org/10.48550/arXiv.2307.15770

Sivarajkumar, S., Kelley, M., Samolyk-Mazzanti, A., Visweswaran, S., & Wang, Y. (2024). An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, *12*, e55318. https://doi.org/10.2196/55318

Td, K. (2023). Generative LLMs and Textual Analysis in Accounting:(Chat) GPT as Research Assistant? https://doi.org/10.2139/ssrn.4429658

Virvou, M. (2023). Artificial Intelligence and User Experience in reciprocity: Contributions and state of the art. *Intelligent Decision Technologies*, *17*(1), 73-125. https://doi.org/10.3233/IDT-230092

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837. https://doi.org/10.48550/arXiv.2201.11903

Yang, Y., Du, Z., Zhang, Z., Tong, G., & Zhou, R. (2021). Does ESG disclosure affect corporate-bond credit spreads? Evidence from China. *Sustainability*, *13*(15), 8500. https://doi.org/10.3390/su13158500

Yang, Y., Tang, Y., & Tam, K. Y. (2023). InvestLM: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*. https://doi.org/10.48550/arXiv.2309.13064

Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang, S., Tong, H., Xiao, L., & Zhou, W. (2023). ESG Reveal: An LLM-based approach for extracting structured data from ESG reports. *arXiv preprint arXiv:2312.17264*. https://doi.org/10.48550/arXiv.2312.17264