University College London

Department of Computer Science

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Emerging Digital Technologies, University College London

Leveraging GPT for ESG Extraction in Cryptocurrency: Addressing Data Scarcity through Prompt Engineering and Multi-Agent Collaboration

Author
Bing Chia

Academic Supervisor
Dr Silvia Bartolucci
Department of Computer Science
University College London

Industrial Supervisor
Mr Joshua De Vos
Research Lead
CCData

February 10, 2025

This dissertation is submitted as part of the requirement for the MSc Emerging Digital Technologies degree at UCL. It is substantially the result of my own work except where explicitly indicated in the text.

Abstract

This dissertation addresses the challenge of scarce datasets and the limitations of fine-tuning models in the evolving field of cryptocurrency, particularly regarding Environmental, Social, and Governance (ESG) issues. To overcome this, we leverage GPT-4 models without fine-tuning and explore various prompt engineering techniques, including Zero-Shot, Few-Shot, Chain-of-Thought, and Multi-Agent Debate. Our findings demonstrate that even without fine-tuning, GPT-4 effectively identifies relevant ESG content, making it a viable tool for complex, context-sensitive tasks.

In straightforward extraction tasks, Zero-Shot (ZS) outperformed Chain-of-Thought (COT) reasoning when provided with context. However, COT offered better generalizability and improved coherence in more complex scenarios. Few-Shot learning required careful example selection to avoid introducing noise, while the critique-refinement phase in Multi-Agent Debate reduced irrelevant content but risked losing true positives.

Building on these insights, we propose a Multi-Agent Iterative Debate framework, which refines responses over multiple iterations to improve accuracy. This framework significantly enhances the extraction of nuanced ESG content, providing a robust, scalable solution without requiring dataset-specific fine-tuning.

Acknowledgments

I would like to express my deepest gratitude to my academic supervisor, Dr. Silvia Bartolucci, for her invaluable guidance, support, and encouragement throughout the course of this research. Her insights were instrumental in shaping the direction of this dissertation. I also extend my thanks to my industrial supervisor, Joshua, for offering practical insights and continuous motivation during this project. Special thanks go to my family and friends for their unwavering support, patience, and understanding throughout this journey. Finally, I am grateful to the entire Department of Computer Science at University College London for providing a conducive environment for research and development.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Importance of ESG in Cryptocurrency

Environmental, Social, and Governance (ESG) considerations have gained significant prominence in the cryptocurrency space. As global awareness of sustainability and ethical practices increases, ESG is rapidly becoming a commonly recognized investment and grant application metric (1; 2). A sizable portion of the world's institutional money now has an ESG mandate, with global ESG-related assets under management (AUM) projected to reach $33.9 trillion by 2026, constituting 21.5% of total global AUM, according to a report from PricewaterhouseCoopers (3). Additionally, programs like the VeChain Development Grant and the Stellar Community Fund emphasize funding for projects that demonstrate strong ESG performance (2), further highlighting the importance of ESG considerations in securing investment and driving innovation within the cryptocurrency space.

In the context of cryptocurrencies, ESG issues such as energy consumption, waste management, regulatory compliance, community impact, and governance structures have become critical areas of concern (4; 5; 6). Recent reports and benchmarks, including those from CCData (7) and the Crypto Carbon Ratings Institute (CCRI) (8), have provided detailed analyses and data on these concerns. For instance, CCData, in collaboration with CCRI, published the inaugural ESG Benchmark, evaluating parameters such as carbon footprint, governance structure, and accessibility, with Ethereum, Solana, and Cardano receiving top ranks while Bitcoin was penalized due to its high energy usage from Proof of Work (PoW) consensus algorithm (9). Additionally, CCRI's MicA methodology offers a comprehensive way of estimating the carbon footprint of cryptocurrencies by calculating key metrics such as energy consumption, greenhouse gas emissions, waste production, and natural resource utilization (10).

While most recent studies and data companies focus on quantifiable ESG metrics, such as carbon emissions, and provide precise, well-founded methodologies through APIs, our research addresses a more complex NLP-related task: extracting ESG-related issues from unstructured text data. Unlike traditional industries where corporate social responsibility (CSR) reports are readily available, the cryptocurrency space lacks standardized reporting on ESG concerns. As a result, we turn to sources like Coindesk articles to gather insights. By developing a machine learning algorithm capable of automatically identifying these issues, we aim to significantly reduce the time and effort typically required for manual flagging and analysis. This approach streamlines the process of recognizing and addressing ESG concerns within the evolving and often less transparent crypto space.

1.2 Leveraging Large Language Models (LLMs)

With the advent of large language models (LLMs) like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), natural language processing (NLP) tasks have seen significant performance improvements. BERT achieved state-of-the-art results on benchmarks such as GLUE (General Language Understanding Evaluation) and SQuAD (Stanford Question Answering Dataset), showcasing its effectiveness in tasks like text classification, sentiment analysis, and question answering (11). Similarly, GPT has achieved state-of-the-art results in various NLP benchmarks, demonstrating superior capabilities in text generation, understanding, and comprehension across diverse tasks(12).

These models excel due to their transformer architecture and attention mechanism, which allow them to effectively capture contextual relationships in text. The transformer architecture relies on multi-head self-attention mechanisms that enable the model to attend to different parts of a sentence simultaneously, improving the model's ability to understand and generate text (13).

Previous efforts in using large language models (LLMs) for extracting ESG issues have primarily focused on pretraining and fine-tuning models, especially BERT, for data extraction tasks (14; 15; 16). While fine-tuned models often outperform baseline models, they typically require a significant amount of meticulously curated, human-annotated data, making the process labour-intensive and time-consuming. Moreover, the continually evolving ESG criteria require ongoing extensive labelling and model retraining to stay current with the latest trends, which can be both costly and environmentally unsustainable.

Additionally, fine-tuned models are known to have lower generalization performance on out-of-distribution data and can sometimes generate spurious outputs (17). The instability of the fine-tuning process, especially on small datasets, has been well-documented, with various factors such as random weight initialization and data order in stochastic optimization contributing significantly to performance variability (18).

1.3 Research Contributions

Given the challenges associated with fine-tuning and the absence of a labelled dataset, this study leverages GPT-4 to perform ESG issue extraction without the need for extensive fine-tuning. Previous research has shown that GPT models can perform effectively without such fine-tuning. For instance, (17) demonstrated GPT-3's robust performance across various tasks with minimal tuning, while (19) highlighted GPT's capability to classify issues almost as effectively as fine-tuned BERT models, proving its practical utility in real-world scenarios. Building on these findings, we aim to apply this approach to the ESG cryptocurrency space, demonstrating a list of prompt engineering methods, such as Chain-of-Thought reasoning (20), Multi-Agent Debate (21), Zero-Shot and Few-Shot learning (also known as in-context learning) (17), to achieve reliable results.

Additionally, we explored how multiple agents can collaborate more effectively in a framework we propose called Multi-Agent Iterative Debate. This framework is inspired by the original Multi-Agent Debate (21) and concepts from Reinforcement Learning (RL) with AI feedback (22). In this approach, agents iteratively refine their outputs based on feedback from other agents, much like RL agents adjust their strategies based on rewards or penalties. While our framework does not involve updating model weights directly, it leverages a form of learning through adaptation and refinement, aiming to improve extraction accuracy by considering diverse perspectives and enhancing collaborative decision-making.

To ensure the consistency and reliability of our results, we employ a comprehensive accuracy assessment framework that integrates multiple metrics. First, we create a small subset of labelled data, where we evaluate Precision, Recall, and Intersection over Union (IoU) to provide insights into the relevance and completeness of the extracted ESG issues. Precision and Recall measure the proportion of correctly identified issues relative to all identified and all relevant issues, respectively, while IoU quantifies the overlap between the extracted content and the ground truth, capturing the extent of information coverage.

For the unlabelled data, we employ semantic similarity analysis to maintain robust evaluation standards. This approach leverages the proximity of sentence embeddings to assess relevance, even when explicit labels are not available. By analyzing the semantic closeness between the model-generated sentences and those manually verified in the labelled set, we can infer the relevance of new sentences. This ensures that sentences not directly checked against the ground truth are still considered if they align closely with the verified content, thus enhancing the overall reliability of our evaluation.

The main contributions of this research are as follows:

- **Comprehensive Evaluation of Prompt Engineering Methods**: We systematically explore and compare various prompt engineering techniques for GPT-4, including Zero-Shot, Few-Shot, and Chain-of-Thought, Multi-Agent Debate approaches, providing insights into their effectiveness for ESG issue extraction.

- **Introduction of Multi-Agent Iterative Debate**: We propose a novel Multiagent Iterative Debate framework, inspired by Multi-Agent Debate and Reinforcement Learning with AI Feedback (RLAIF), which enhances the extraction and refinement of ESG-related content in complex and ambiguous contexts.

- **Addressing Model Limitations**: We identify and examine key limitations in using GPT-4 for ESG analysis, including issues such as hallucination, recency bias, and challenges in extracting nuanced content. By exploring these limitations, we provide valuable insights into the model's behaviour, offering practical guidance on how to configure and adapt GPT-4 to specific use cases.

- **Robust Accuracy Measures for Unlabelled Data**: In the absence of labelled datasets, we propose a comprehensive evaluation framework combining multiple metrics such as Precision, Recall, Intersection over Union (IoU), and semantic similarity.

Chapter 2

Literature Review

2.1 Domain-Specific BERT Variations for ESG Classification

FinBERT (23) is one of the pioneering BERT models specifically pre-trained on financial corpora, including the TRC2-financial corpus, to perform financial sentiment analysis. Its novel contribution lies in its ability to capture domain-specific nuances in financial text, making it exceptionally effective for sentiment analysis in the financial sector. FinBERT serves as a foundational model for many subsequent ESG BERT models, providing a robust benchmark for evaluating the performance of models designed for ESG-related text extraction and analysis. By leveraging the specialized financial knowledge ingrained during its pre-training, FinBERT facilitates significant improvements in accuracy, robustness, and applicability within the ESG domain. It enables researchers to effectively benchmark and enhance fine-tuned models, driving advancements in ESG data extraction and analysis.

"Mapping ESG Trends by Distant Supervision of Neural Language Model" (16) begins by manually labelling a dataset of Corporate Sustainability Reports (CSRs) with ESG Key Performance Indicators (KPIs) to establish relevance to ESG topics. These KPIs serve as a foundation for associating sentences with relevant ESG categories. For unlabelled data, the study uses cosine similarity to compare new sentences to these labelled KPIs, pre-assessing their relevance to ESG issues. Once this vector-space matching is done, the study fine-tunes a pre-trained BERT model to classify sentences as irrelevant, somewhat relevant, or highly relevant to ESG topics. By applying this method to earnings call transcripts from S&P 500 companies (2010-2018), the study offers valuable insights into trends in ESG discussions, particularly the growing emphasis on social issues. The model achieved an F1-score of 80.3% on a validation set, demonstrating the effectiveness of combining distant supervision with semantic knowledge for classifying ESG-related content. One notable limitation of the paper's methodology is the potential loss of granularity. By relying on cosine similarity for sentence classification, the approach ties relevance to surface-level associations between the text and predefined ESG Key Performance Indicators (KPIs). While this method effectively maps sentences based on keywords and phrases, it can overlook deeper contextual understanding. The semantic relationships identified by cosine similarity are often limited to surface-level textual matches, which might not fully capture the complexities or nuances present in ESG-related discussions.

"ESGBERT: Language Model to Help with Classification Tasks Related to Companies' Environmental, Social, and Governance Practices" (14) introduces ESGBERT, a domain-specific variation of BERT tailored for ESG text classification. The original BERT model undergoes further pre-training on an extensive ESG corpus from the Knowledge Hub of Accounting for Sustainability using Masked Language Modeling (MLM), capturing nuanced ESG-related language and context. This novel pre-training is followed by fine-tuning on ESG-related 10-Q filings to predict changes in environmental scores. ESGBERT's innovative dual-phase training approach allows it to achieve significant accuracy improvements over baseline models. Specifically, the model attained a training accuracy of 83.9% and a test accuracy of 67.09% for detecting changes in ESG scores. These results demonstrate ESGBERT's efficacy and robustness in handling ESG-specific classification tasks, contributing valuable insights and advancements to the field of ESG analysis. However, ESGBERT's specialization on a narrow ESG corpus limits its generalizability to other domains, such as cryptocurrency or technology, where ESG practices and challenges can vary significantly. Adapting the model to these different sectors would require additional fine-tuning data to adequately capture the unique ESG concerns and context-specific language of those industries, reducing its out-of-the-box applicability.

The paper "NLP for Responsible Finance: Fine-Tuning Transformer-Based Models for ESG" (15) introduces a robust method for assessing companies' ESG performance using textual data from annual reports. This study leverages ESG ratings from S&P Global and employs a novel pre-screening process using a BERT-based semantic similarity measure to identify ESG-relevant text segments. The authors fine-tuned three transformer-based models: BERT-Base, FinBERT, and RoBERTa-Large. Among these, the fine-tuned RoBERTa-Large model demonstrated exceptional performance, achieving the highest accuracy of 79% in predicting ESG ratings, which is 11 percentage points higher than traditional TF-IDF-based models. This work highlights the efficacy of combining semantic similarity measures with advanced transformer models to enhance the accuracy and reliability of ESG performance assessments, offering a significant improvement over conventional methods.While the paper presents a compelling method for assessing ESG performance, the reliance on pre-screening through semantic similarity measures could introduce certain limitations. The semantic similarity approach, while effective in filtering relevant text segments, may overlook deeper contextual nuances or subtle ESG indicators that do not directly align with predefined similarity metrics. This could potentially lead to the exclusion of critical information that is relevant but expressed in less conventional ways.

"Bridging the Gap in ESG Measurement" (24) presents an innovative approach to enhancing ESG measurement through the pre-training and fine-tuning of transformer models. The study pre-trains RoBERTa and DistilRoBERTa models on a substantial corpus of 13.8 million textual samples sourced from corporate news, annual reports, and sustainability reports. Following this, a keyword search identifies relevant text passages, which are then meticulously reviewed and annotated by experts to create specialized datasets for environmental, social, and governance categories. Distinct models are further pre-trained for each ESG domain, resulting in three separate models for E, S, and G. These datasets are used to fine-tune the pre-trained models, significantly boosting their ability to classify ESG-related content with high accuracy. The results are impressive, with the fine-tuned models achieving over 93% accuracy in the social and environmental domains and over 89% in the governance domain, demonstrating state-of-the-art performance and showcasing the effectiveness of this detailed, domain-specific approach. While the study offers an impressive approach to ESG measurement by fine-tuning models on distinct ESG categories, the reliance on keyword searches for initial passage identification presents a potential limitation. This method, while useful for capturing directly related content, may miss more subtle or implicit ESG issues that do not contain obvious keywords, leading to gaps in coverage. Additionally, the creation of separate models for each ESG domain, while improving domain-specific accuracy, could reduce the ability to identify interconnected ESG concerns that span multiple categories, such as how governance practices may influence environmental outcomes.

2.2 Approaches Using GPT for ESG Data Extraction

The paper "ESGReveal" (25) introduces an innovative approach for extracting and analyzing ESG data from corporate reports using Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) techniques. This methodology features an ESG Metadata Module, which establishes a query framework compliant with the ESG standards published by the Hong Kong Stock Exchange (HKEx), organizing ESG criteria into structured prompts. The Report Preprocessing Module utilizes advanced computer vision and NLP tools to convert ESG reports into structured knowledge bases, encompassing text, document outlines, and tables. The LLM Agent Module employs state-of-the-art LLMs like GPT-4 to extract relevant ESG data from the preprocessed reports, leveraging the metadata and structured knowledge bases to engineer precise and context-aware prompts for data retrieval. Notably, GPT-4 achieved an accuracy of 76.9% in data extraction and 83.7% in disclosure analysis. This work highlights the potential of combining LLMs with RAG techniques to enhance the accuracy and efficiency of ESG data extraction and analysis, setting a new standard in the field. While the preprocessing methods in "ESGReveal" are highly effective, they are relatively complex and may require significant computational resources. Additionally, the use of the HKEx guideline provides a strong framework for query extraction, but this approach may struggle to generalize effectively outside of the specific standards provided by the HKEx.

The paper "ESG Classification by Implicit Rule Learning via GPT-4" (26) explores the use of GPT-4 for ESG classification tasks without explicit training data. By employing innovative prompt engineering techniques to follow MSCI guidelines, chain-of-thought reasoning, and in-context learning with dynamically retrieved examples, the model effectively aligns with implicit ESG evaluation criteria. This approach is novel in its ability to leverage advanced language model capabilities without the need for extensive labelled datasets, showcasing the potential of GPT-4 in understanding and applying complex, domain-specific guidelines through implicit learning. The study utilizes a dataset of 800 Korean ESG news articles for training and 200 for testing, categorized by impact type and duration. While GPT-4 achieved an accuracy of 76.13% in Impact Type Classification, it faced challenges in predicting impact durations accurately, resulting in a lower accuracy of 43.98%, partly due to difficulties in handling long prompts and misclassification. This research highlights the potential of GPT-4 in automating ESG evaluations, presenting a significant step forward in applying LLMs to specialized, high-stakes domains without the need for explicit training data. The techniques used in the paper are robust and effectively leverage MSCI guidelines for prompt construction. However, they could potentially benefit from more complex or varied prompts to capture deeper contextual nuances and address the more intricate aspects of real-world ESG issues.

Chapter 3

Methodology

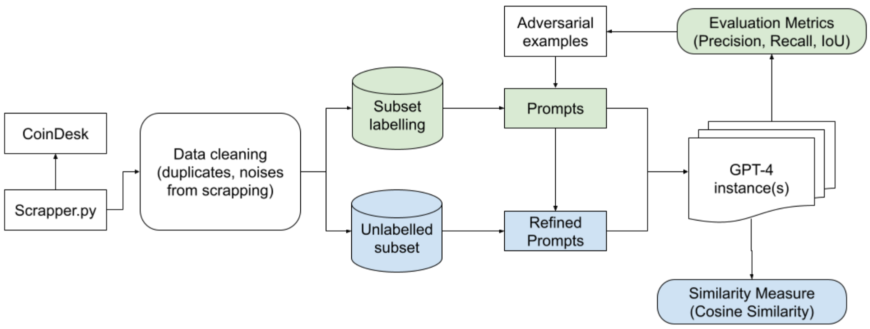

Figure 3.1: Project workflow

The project workflow, as illustrated in Figure 3.1, consists of the following steps:

1. **Manual Label and Review**:

   - A subset of 20 articles, retrieved as detailed in Section 3.1, is manually labelled. This involves creating a CSV file of ESG-related issues by extracting relevant sentences from the articles. These sentences are aligned with the topics defined in the system message (see Figure 3.4).

2. **Subset Verification**:

   - The manually labelled articles are processed using GPT-4 with various prompt configurations, as described in Section 3.3, to assess the model's performance.
   - Since the output sentences are generated in an array format, a regular expression is employed to efficiently extract the relevant sentences.
   - The quality of the extracted sentences is evaluated using key metrics such as *Precision*, *Recall*, *All Intersection over Union (IoU)*, and *Best IoU*, ensuring a comprehensive performance assessment, as detailed in Section 3.4.

3. **Iterative Prompt Refinement**:

- If certain sentences are not extracted accurately, the system message or prompt is iteratively refined. This process continues until the model consistently extracts the desired information from the labelled data.
- Adversarial examples (sentences that are frequently missed or incorrectly extracted—are identified) are integrated into the prompts, similar to Few-Shot learning, to enhance model performance by addressing challenging cases.

4. **Application to Unlabelled Data**:

   - Once the labelled subset achieves a consistent and reliable extraction result, it serves as a benchmark for evaluating the unlabelled data.
   - The refined prompt is applied to the entire unlabelled dataset, and the overall *cosine similarity* is calculated to assess how closely the new extractions match the benchmark, ensuring the quality of the results on unseen data.

3.1 Data Source

The dataset consists of 5800 articles scraped from CoinDesk (27), a leading source of news on cryptocurrency, blockchain, and the digital finance market. The collected data was cleaned to remove irrelevant or noisy content, such as copyright notices, privacy policies, and terms of use, which were inadvertently captured during the scraping process.

In order to preserve the integrity of the text for further analysis, no advanced natural language processing (NLP) techniques were employed. This decision was made to ensure that the original sentence structures were retained, as pre-processing steps like stemming or lemmatization could interfere with attention-based models used in subsequent stages.

To explore the relevance of environmental, social, and governance (ESG) topics within the cryptocurrency sector, a manual labelling process was applied to a subset of 20 selected articles. Sentences were labelled according to ESG criteria presented in an HSBC report (4), resulting in a total of 88 annotated ESG-related issues.

Figure 3.2 shows the distribution of article lengths on a logarithmic scale. The peak of the distribution occurs between a log value of 2.5 and 3.0, corresponding to articles approximately 300 to 1000 words in length, which represent the majority of the dataset. The wide variability in article length, with some articles exceeding 8000 words, highlights the challenge for models like GPT-4 in processing and extracting relevant ESG content consistently across varying document sizes. Shorter articles may contain less context, while longer ones present the risk of diluting key ESG issues with excessive detail.

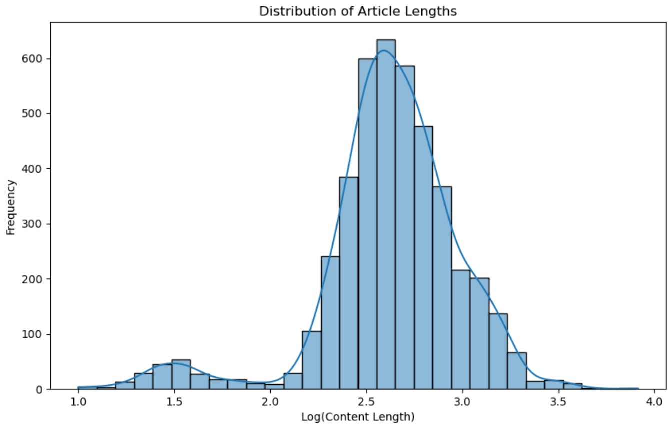

Figure 3.2: Distribution of CoinDesk Article Length.

3.2 GPT Configuration

In this project, the GPT-4 model is accessed using the g4f library (28) with the Chatgpt4o provider.

Previous work utilizing BERT models often processes texts at the sentence level due to BERT's lower token limit of 512 tokens. In our research, we do not split the sentences. Instead, we leverage GPT-4's significantly larger token limit of 8,192 tokens for the base model to process entire articles, enabling the model to capture the broader context and nuances in the full text. By maintaining the integrity of the entire article, GPT-4 can provide a more comprehensive understanding of the ESG issues within the cryptocurrency space, ensuring that important contextual information is not lost.

Several configurations and parameters are used to control GPT-4's behaviour and optimize its performance for the task at hand. Key aspects of the implementation are described below.

3.2.1 System Message and User Prompt

The distinction between the system message and the user prompt is crucial in guiding the model's responses. The system message provides essential background information and establishes the parameters for the model's behaviour throughout the interaction. This context directs the model to concentrate on relevant dimensions of the input data, which is especially beneficial in generating more accurate and contextually appropriate embeddings, enhancing overall coherency. Given that different stakeholders may prioritize various aspects of ESG, the system message effectively guides the model in focusing on pertinent issues.

Figure 3.3 shows how a general-purpose system message sets a broad context for the model's responses, aiming for conciseness and politeness. Figure 3.4 shows our system

```
You are ChatGPT, a large language model trained by OpenAI.
   Answer as concisely as possible while being polite and
   helpful.
```

Figure 3.3: Default System Message for OpenAI's ChatGPT

,

```
You are an expert in Environmental, Social, and Governance (
   ESG) topics, specifically within the cryptocurrency space.
Given an article, you will be asked to extract ESG issues
   from it.
Here are the key ESG issues that are particularly relevant in
    the context of cryptocurrencies:

Environmental (E): Energy Consumption, Carbon Emissions,
   Resource Management, Renewable Energy Usage, Electronic
   Waste Production.

Social (S): Labor Practice, Community Engagement and
   Inclusion, Security and User Protection, Entry Barrier and
    Accessibility, Market Instability, Illicit Activities,
   Influence of major financial institutions.

Governance (G): Decentralized Governance Models (off-chain
   and on-chain), Business Ethics and Transparency,
   Regulatory Compliance, Executive Compensation and
   Incentives, Tax Evasion, Geographical Differences and
   Regulatory Challenges
```

Figure 3.4: Example System Message for ESG Analysis in Cryptocurrencies

context where each aspect of ESG is taken from (4).

The user prompt consists of the specific query provided by the user at each interaction, directing the model to perform a particular task, such as extracting sentences related to a specific ESG category from a given article. While the system message helps set the context and focus within the embedding space, the user prompt is the immediate input that the model processes to generate a response. Engineering the prompt can have a more direct and noticeable impact on the output, making it a primary focus for achieving specific results, as shown in recent studies.

3.2.2 Temperature

The temperature parameter controls the randomness of the model's output. It ranges from 0 to 1, where a temperature of 0 produces deterministic results with no randomness, and a temperature of 1 introduces the highest level of randomness.

A lower temperature is preferred for tasks requiring precision and accuracy, such as extracting specific information from a text. By setting the temperature to 0, the model always selects the next word with the highest probability, producing deterministic results. If the temperature is not 0, the next tokens are selected using sampling methods, which introduce variability by randomly choosing from the top tokens based on their probabilities.

This study does not exclusively use a temperature of 0. For certain experiments, such as Chain-of-Thought Self-Consistency (COT-SC) and Multi-Agent Debate, the temperature is increased to produce variations in responses. Higher temperatures are beneficial in these scenarios to explore a broader range of potential answers and reduce confirmation bias by allowing the model to generate more diverse outputs.

3.2.3 Top-p vs Top-k sampling

Top-p and top-k are sampling strategies that work alongside temperature to control the model's output (29). Both methods aim to balance randomness and coherence by limiting the pool of candidate tokens from which the model can generate its next word.

Top-p (Nucleus Sampling): Top-p, or nucleus sampling, selects tokens from the smallest possible set whose cumulative probability is at least p. This method dynamically adjusts the number of tokens considered based on their probabilities, allowing for a more flexible and context-aware selection process. The cumulative probability is calculated as follows:

$$P_c = \sum_{i=1}^{k} P(w_i)$$

where $P(w_i)$ is the probability of the i-th token, and P_c is the cumulative probability. The smallest set of top tokens is chosen such that $P_c \geq$ top-p. In this study, top-p is set to 0.4, meaning that the model selects tokens from the top 40% of the probability mass.

Top-k: Top-k sampling, on the other hand, restricts the model from choosing from the top k most probable tokens. This approach provides a fixed number of candidates regardless of their cumulative probability. While top-k can be useful in certain scenarios, top-p is generally more effective for maintaining a balance between randomness and relevance, hence it is selected as the sampling method in our research.

3.2.4 Known Limitations

- **Token Limit:** The base GPT-4 model has a limit of 8,192 tokens, including both input and output tokens, as well as the system message, which poses a restriction. Although the average token lengths of our data fall within this limit, lengthy contexts can still result in truncation, potentially losing critical information. This token limit is particularly problematic for complex prompts such as Few-Shot learning and Multi-Agent Debate, where additional context and multiple interactions are necessary.

- **Context Length and Hallucinations:** Long contexts can lead to hallucinations, where the model generates content not grounded in the input data, affecting the reliability of the output (18).
- **Lack of Pretraining or Fine-tuning Capabilities:** At the time of writing, GPT-4 does not support additional pretraining or fine-tuning on datasets to adapt to particular tasks or domains. This limitation prevents us from conducting experiments that compare domain-specific adaptation performance.
- **Absence of Encoding Functions:** GPT-4 does not provide encoding functions directly, necessitating the use of separate pretrained transformer models to encode the output embeddings. These embeddings are subsequently used to calculate performance measures such as cosine similarity.

3.3 Prompt engineering

3.3.1 Zero-shot Learning

Zero-shot learning is a method where a model is applied to tasks or data it has not been explicitly trained on by leveraging its general knowledge base. This approach allows the model to generalize to new, unseen tasks without requiring additional task-specific training data (17).

This method is particularly valuable in domains like ESG analysis in cryptocurrencies, where labelled datasets are scarce or expensive to create. Zero-shot learning enables the model to generate insights and perform tasks with minimal human input, making it a cost-effective solution for emerging fields.

However, zero-shot learning also has its drawbacks. Without fine-tuning for specific tasks, the model lacks domain-specific understanding and may produce less precise or reliable results compared to models trained on labelled data. In emerging fields like ESG in cryptocurrency, the model's knowledge may be outdated or incomplete, leading to irrelevant or inaccurate outputs. Additionally, zero-shot learning struggles with tasks requiring nuanced, subjective interpretation, making it harder for the model to adapt to evolving concepts or subtle shifts in public perception within ESG-related content.

Findings from (30) indicate that prompt performance follows a U-shaped curve, with the highest performance when relevant information is at the beginning or end of the input context.

- **Task-First:** More effective for straightforward tasks where immediate clarity is crucial. Useful in structured tasks like code generation, simple Q&A, or data extraction.

- **Task-Last:** Better for complex tasks where the model benefits from having all the context first. Useful in summarization, context-driven generation, and nuanced Q&A.

In this research, we adopt the task-last approach. This strategy allows the model to fully comprehend the context and nuances of the articles before attempting the extraction task.

```
Article Title: {...}
Article Content: {...}
Task: Identify any sentences from the article that might
   involve ESG (Environmental, Social, Governance) topics.
Return your answer in a JSON array format with each
   identified sentence as a string.
```

Figure 3.5: Zero-shot Prompt

3.3.2 Few-shot Learning

Few-shot learning extends the capabilities of zero-shot learning by providing the model with a small set of examples to learn from before making predictions. While zero-shot learning relies solely on the model's pre-trained knowledge without task-specific input, few-shot learning enhances performance by introducing a limited set of annotated examples, offering critical context and guidance.

In our approach, we use four carefully selected adversarial examples—instances where the model previously struggled to extract relevant information or identified irrelevant content during zero-shot learning. For each example, we include the relevant sentence and its corresponding ESG analysis as shown in Figure 3.6. This structure ensures that the examples are concise while offering the model targeted learning opportunities that clarify how to distinguish between relevant and irrelevant content. By including justifications, we guide the model's reasoning process, helping it understand why specific content is classified as ESG-related.

These adversarial examples are particularly useful in resolving ambiguity and reshaping how the model organizes and interprets information within its embedding space. This adjustment improves the model's ability to differentiate between similar sentences or concepts, leading to more accurate future predictions. Additionally, we rephrase the adversarial examples to prevent the model from relying on memorized errors, ensuring a more authentic improvement in performance. This approach strengthens the model's accuracy and adaptability, making it better suited to handle complex and ambiguous ESG scenarios.

```
Article Title: {...}
Article Content: {...}
Examples: {
"""
Example 1:
Sentence: "At the same time, another poll indicated that
   Biden would win the popular vote, which simply reflects
   the total number of individual votes and has no bearing on
    the electoral college outcome."
ESG analysis: This sentence focuses solely on election
   results and the distribution of the popular vote, which is
    unrelated to cryptocurrencies or any ESG issues. While
   different presidential candidates may have varying stances
    on crypto, this sentence does not mention any direct
   connection between the election outcome and ESG concerns.
"""
...
}

Task: Identify any sentences from the article that might
   involve ESG (Environmental, Social, Governance) topics.
Return your answer in a JSON array format with each
   identified sentence as a string.
```

Figure 3.6: Few-shot Prompt

3.3.3 Chain-of-Thought (COT)

Chain-of-Thought (COT) is a reasoning method that decomposes complex problems into simpler sub-problems, thereby reducing the cognitive load on the model (20). This approach ensures that various aspects of the problem are considered sequentially rather than attempting to capture everything at once.

COT leverages the transformer architecture's left-to-right processing capability, where steps are sequentially processed and the model updates its embeddings with each step, incorporating the information processed so far. This progressive contextualization helps the model build a more nuanced understanding of the problem, leading to more accurate and coherent responses.

COT also reduces hallucinations by following a logical sequence and avoiding shallow or superficial responses. By adhering to a structured reasoning process, the model can provide more reliable and contextually appropriate outputs.

In our approach, we first ask the model to locate and summarize the Environmental, Social, and Governance (ESG) aspects separately. This strategy ensures that the model understands the specific E, S, and G issues as distinct themes that persist throughout the article, thereby reducing the cognitive load associated with capturing all three aspects simultaneously. Once these summaries are generated, the model then extracts the original sentences from the article based on these summaries. This two-step process enhances the model's ability to provide contextual information about the article and justifies the reasoning behind the extracted sentences. This is unlike zero-shot extraction, where only sentences are extracted without the supporting contextual information.

```
Article Title: ...
Article Content: ...

Step 1: Identify and explain any Environmental (E) aspects
   mentioned in the article.
Environmental Aspects:

Step 2: Based on Step 1, extract the original sentences from
   the article that relate to the Environmental Aspects.
   Return the sentences in an array.
Environmental Array:

Step 3: Identify and explain any Social (S) aspects mentioned
    in the article.
Social Aspects:

Step 4: Based on Step 3, extract the original sentences from
   the article that relate to the Social Aspects. Return the
   sentences in an array.
Social Array:

Step 5: Identify and explain any Governance (G) aspects
   mentioned in the article.
Governance Aspects:

Step 6: Based on Step 5, extract the original sentences from
   the article that relate to the Governance Aspects. Return
   the sentences in an array.
Governance Array:
```

Figure 3.7: Chain-of-Thought (COT) Prompt

3.3.4 Chain-of-Thought Self-Consistency (COT-SC)

Self-consistency enhances Chain-of-Thought (COT) reasoning by leveraging multiple independent reasoning paths to produce more reliable and accurate results (31). In standard COT, the reasoning process can sometimes lead to inconsistent or inaccurate answers due to inherent variability in the model's generation process, even with a temperature setting of 0. This variability arises from the complex interactions between the model's internal weights and biases, as well as the context sensitivity and ambiguities present in the input data. To address this, self-consistency generates multiple reasoning paths and selects the most consistent answers across these paths.

n this study, we implement a variant of COT-SC that combines all sentences generated by independent reasoning paths, rather than selecting only the most consistent, overlapping sentences. This approach is motivated by the observation that relying solely on overlapping content can result in a limited and incomplete output. To introduce controlled variability and capture less obvious ESG-related content, we increase the temperature settings by using random values with a seed of 42. This method allows the model to explore a wider range of potential insights, enhancing the diversity of the extracted information without sacrificing coherence.

The prompt used in this variant is the same as in standard COT 3.7, ensuring consistency in the task setup. Given the expanded breadth of reasoning paths and varied sentence combinations, this approach maintains the robust reasoning framework of COT while exploring more nuanced and less obvious aspects of the articles.

3.3.5 Multi-Agent Debate

Inspired by the original Multi-Agent Debate process (21), which uses multiple agents working collaboratively to improve answers, we adapted this framework by allowing multiple GPT instances to work on the same set of responses. Similar to Chain-of-Thought Self-Consistency (COT-SC), the initial responses are generated using multiple reasoning paths, each with a randomly assigned temperature and a top-p value of 0.4, introducing controlled variability to capture a diverse set of answers.

Our approach follows a three-phase process:

1. **Generation of Initial Responses**: Multiple agents generate independent responses to the task, similar to COT-SC in Section 3.3.4.

2. **Critique Phase**: Agents independently assess these responses. During this phase, each response is broken down sentence by sentence, and agents evaluate the content based on three key aspects: completeness, relevance, and depth of reasoning (refer Figure 3.8).

3. **Refinement Phase**: Each agent votes to include or exclude each sentence based on all the critiques, and the majority decision determines whether a sentence is retained or eliminated, leading to a more refined and accurate set of responses.

This structured critique process, where each agent independently evaluates from its own perspective, introduces diverse viewpoints on what constitutes an ESG issue. The voting mechanism helps eliminate weaker, potentially noisy sentences by ensuring that only content agreed upon by the majority is retained. This Multi-Agent Debate framework not only embraces the increased variety of answers but also actively works to reduce the noise associated with it, enhancing both the diversity and accuracy of the final output.

To rigorously evaluate this method, we designed an experiment to assess how well the critique-refinement process filters out noise from the final set of answers. By increasing the number of GPT agents in the initial response generation, we aim to observe changes in both precision and recall. If the critique process successfully filters out irrelevant content, we expect an increase in precision. At the same time, the added variability from multiple agents is likely to raise the number of true positives, enhancing recall. This experiment will help determine whether the inclusion of more perspectives leads to better content extraction or diminishing returns due to an influx of noise.

```
"Given the article context and extracted ESG-relevant
   sentences, go through each sentence and critique their
   relevancy according to the key themes." \\
"Given the article context and extracted ESG-relevant
   sentences, evaluate if the sentence capture the full scope
    of the ESG issue, or is it missing key details?", \\
"Given the article context and extracted ESG-relevant
   sentences, analyze if there are underlying implications or
    connections to the key themes that may not be immediately
    obvious but could be relevant upon closer examination."

Sentence 1:
Critique:

...
```

Figure 3.8: Multi-Agent Debate Critiques

```
Article Title: {title}
Article Context: {content}
Extracted ESG Sentences: {sentences}

Task: Given the article context and extracted ESG-relevant
   sentences, review the critiques and vote if the sentences
   should be eliminated from the list.

Sentence 1:
Agent 1 Critique:
Agent 2 Critique:
...
```

Figure 3.9: Multi-Agent Debate Refinement Prompt

3.3.6 Multi-Agent Iterative Debate

The Multiagent Iterative Debate is an enhanced version of the Multi-Agent Debate framework, designed to improve the extraction of relevant ESG issues through a feedback loop

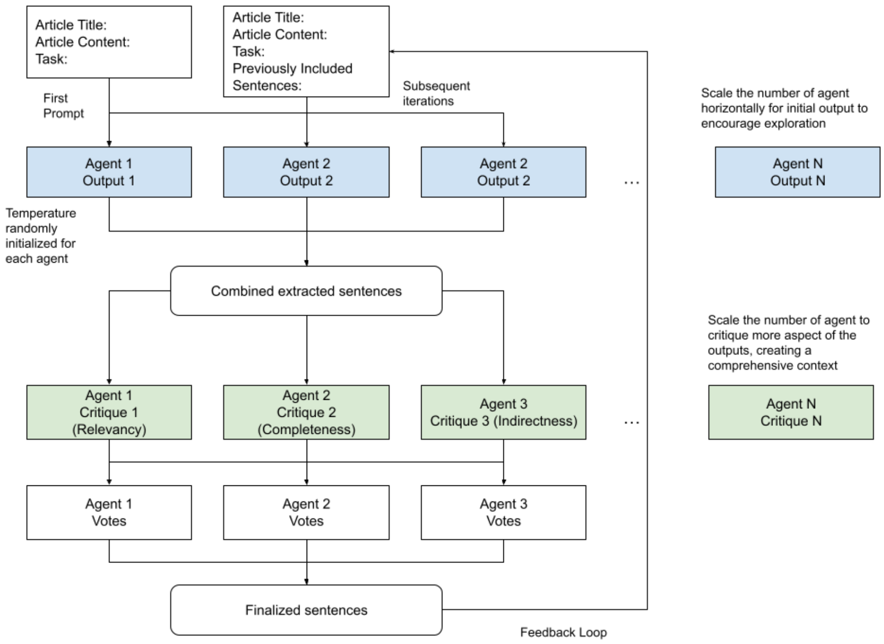

Figure 3.10: Multiagent Iterative Debate Workflow

mechanism inspired by Reinforcement Learning principles (22). This methodology aims to simulate a form of iterative learning and refinement, allowing agents to explore and identify new relevant answers or content through multiple rounds of debate.

The Multi-Agent Iterative Debate incorporates an iterative feedback loop that enables further exploration beyond the initial round of extractions. Unlike Reinforcement Learning, which relies on adjusting model weights and gradients based on feedback, our approach is constrained by the lack of access to underlying model parameters. Therefore, we implement a proxy feedback loop by using refined answers from one debate round as the starting point for the next iteration.

To execute this iterative process, the refined outputs i.e., the sentences extracted and agreed upon by the agents in the previous round—are fed into the subsequent iteration. During each new iteration, agents are tasked with extracting sentences that have not been identified in previous rounds. This step is crucial to ensure that the debate continues to evolve and explore new possible answers rather than reiterating the same content. We also incorporate previously discarded sentences back into the debate during subsequent iterations, allowing agents to re-evaluate and potentially reconsider them. This "re-thinking" process encourages a more thorough examination of the content, providing an opportunity to uncover insights that may have been overlooked or prematurely dismissed in earlier iterations.

A significant challenge in implementing this methodology is the tendency of language models to generate repetitive outputs, often providing the same sentences even when explicitly instructed not to. Despite clear guidance to produce new content, the model can still hallucinate and repeat previously extracted answers. To mitigate this, we manually remove the already-extracted sentences from the article before each new round of debate. This intervention helps reduce redundancy and forces the agents to focus on uncovering additional relevant information that has not yet been considered, enhancing the diversity and quality of the final outputs.

Despite removing extracted sentences for the extraction phase, the agents are still provided with the full context of the article during the critique round. This approach ensures that while the agents are searching for new, unexplored content, they remain informed by the complete context, allowing for a more holistic critique and enabling them to identify potential gaps or overlooked aspects in the extracted content. By iteratively refining the output and critiquing the results with the full context in mind, the Multi-Agent Iterative Debate aims to maximize the thoroughness and relevance of ESG issue extraction across multiple rounds of analysis.

3.4 Evaluation Method

3.4.1 Encoding

Encoding is the process of converting text into numerical vectors that can be processed by machine learning models. This step is crucial for calculating subsequent evaluation metrics such as precision, recall, Intersection over Union (IoU), and cosine similarity, as it allows the model to quantitatively compare and analyze the content based on its numerical representation. Traditional methods of encoding include Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and GloVe (Global Vectors for Word Representation). These methods fall short of capturing the full semantic meaning of sentences, especially when compared to large pretrained transformer models. Transformers produce contextually rich embeddings through their self-attention mechanism, which captures the relationships between words in a sentence, and their extensive pre-training on large text corpora, allowing them to understand and adapt to complex linguistic features. This results in more accurate and context-aware representations compared to static methods like GloVe and TF-IDF.

As mentioned in Section 3.2.4, GPT-4 doesn't provide an encoding function directly, hence an alternative SentenceTransformer('paraphrase-MiniLM-L6-v2') model is chosen due to its efficient and robust performance. It is fine-tuned for paraphrase identification, making it particularly well-suited for tasks like recognizing different expressions of the same underlying concept. In terms of ESG issues extraction, this capability is crucial as it ensures the model can reliably detect and interpret various ways in which the same ESG issue is expressed. This allows us to assess how well the model captures the relevant ESG concerns, leading to a more accurate and comprehensive extraction of pertinent information.

3.4.2 Intersection over Union (IoU)

Intersection over Union (IoU) is a metric commonly used in computer vision, particularly for object detection tasks, to measure the overlap between predicted and ground truth bounding boxes. In our context, IoU is employed to assess the quality of sentence extractions by the model, specifically targeting issues like the model's tendency to truncate sentences, leading to incomplete information extraction.

In the realm of ESG issue extraction, some concerns are articulated over multiple sentences. The model might only extract a portion of these sentences, resulting in an incomplete representation of the issue. This partial extraction poses challenges when calculating traditional metrics like Precision and Recall, as these metrics require exact matches. IoU addresses this challenge by providing a percentage of token overlap between the extracted and reference sentences, thus offering a more granular evaluation of the extraction's completeness.

IoU is calculated as the ratio of the intersection (common tokens) over the union (total unique tokens) between the extracted sentence and the ground truth:

$$\text{IoU} = \frac{\text{Number of common tokens between extracted sentence and ground truth}}{\text{Total number of unique tokens in both extracted sentence and ground truth}}$$

We utilize a threshold of 0.5 for IoU to determine whether an extraction is considered adequately representative of the ground truth.

To further refine our evaluation, we use two specific variations of IoU:

- **All IoU**: This metric calculates the IoU by considering the union of all extracted sentences with the ground truth. It provides a comprehensive measure of how well the model captures the entirety of the relevant content, including the consideration of noise. All IoU is akin to precision but offers a broader view by assessing token overlap rather than exact matches, thus accounting for partial correctness and completeness of extractions.

- **Best IoU**: This metric calculates IoU for only the best-matching extracted sentence for each ground truth sentence. It focuses on the completeness of the result while discarding noise, similar to recall. Best IoU better indicates the quality of extractions by highlighting how well the most relevant sentence aligns with the ground truth, irrespective of any irrelevant sentences extracted.

By using these metrics alongside precision and recall, we ensure a more nuanced assessment of the model's ability to capture ESG issues comprehensively and accurately.

3.4.3 Cosine Similarity

GPT encodes sentences into a high-dimensional vector space, where each sentence is represented as a vector capturing its semantic and contextual nuances. The system message plays a crucial role in guiding the model to generate outputs that align with specific ESG themes, ensuring relevance and coherence. In this context, cosine similarity is a key tool for evaluating how well the generated sentences align with the labelled ones by quantifying their semantic similarity.

Cosine similarity is widely used in natural language processing (NLP) tasks as it provides a robust way to measure the similarity between text data. It compares the direction of vectors in an inner product space, which is particularly effective for text comparison because it focuses on the angle between vectors, rather than their magnitude. This makes it resilient to variations in sentence length and scale, allowing for an accurate comparison of the underlying semantic meaning. By applying cosine similarity, we can assess the quality of the model's output, ensuring that the generated sentences are contextually aligned with the reference labels and minimizing the inclusion of irrelevant or noisy information.

Cosine similarity is defined as:

$$\text{cosine\_similarity} = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|}$$

where $\vec{A}$ and $\vec{B}$ represent the vectors of the two sentences. The dot product $\vec{A} \cdot \vec{B}$ measures their alignment, and $\|\vec{A}\|$ and $\|\vec{B}\|$ are the magnitudes of the vectors.

3.4.4 Pairwise Cosine Similarity

When comparing multiple sentences, we calculate pairwise cosine similarity to measure the overall similarity between each pair of sentences across the dataset. Given N sentences, $\frac{N(N-1)}{2}$ unique pairs are generated, ensuring a comprehensive comparison of all sentence pairs, even across different articles.

Chapter 4

Results and Discussion

4.1 Inherent Problems

Before examining the detailed outcomes of our experiments, it's important to consider some fundamental challenges associated with using large language models like GPT-4 for ESG issue extraction. These models, while powerful, are prone to behaviours that can impact the accuracy and reliability of their outputs such as generating information that isn't explicitly present in the input, often influenced by broader patterns in the training data. Additionally, the model's interpretations can be affected by ambiguity and varying levels of context, potentially leading to diverse interpretations of ESG-related content. These factors must be considered when evaluating the performance and reliability of the model, as they play a significant role in shaping the outputs.

4.1.1 Insufficient information

> Bitcoin got switched on, BTC started getting mined and stuff started happening – small stuff at first. Pizzas <u>were purchased</u>. A website morphed from a place to swap Magic: The Gathering cards to a giant crypto exchange–and then got <u>massively hacked</u>.

Figure 4.1: Example of Insufficient Information

Figure 4.1 illustrates an example where the underlined text links to another article, causing a lack of context. The website in question is Mt. Gox, originally a platform for trading Magic: The Gathering cards, which later became the largest Bitcoin exchange. In 2014, Mt. Gox was hacked, leading to the loss of 750,000 BTC. Since Mt. Gox controlled the private keys of its users, this hack caused users to lose access to their BTC, underscoring critical security vulnerabilities in the Bitcoin ecosystem and highlighting the importance of personal key management.

Understanding the full context of this event clearly classifies it as a Social and Governance issue. However, without access to external articles or detailed background information, the model may struggle to confidently identify these concerns. The lack of comprehensive context limits the model's ability to fully assess the governance failures and security implications of the hack.

To improve the model's ability to assess such events, it would need to tap into broader, pre-existing knowledge about the implications of security breaches within the ESG framework. One method to achieve this is through Chain-of-Thought (COT) reasoning, where the model breaks down its thought process into smaller steps. This structured approach could direct the model's focus toward critical themes in system messages such as "Security and User Protection," helping the model pay more attention to these key topics during the summarization phase. Although in some cases Zero-Shot (ZS) without COT outperforms the COT approach, COT may still be beneficial in situations where deeper contextual awareness is required, as it reduces cognitive load and directs the model's focus to the most relevant aspects of the event.

Alternatively, a multi-agent debate approach can address the issue of missing context by having multiple agents provide diverse interpretations of the input. Each agent independently analyzes the event and engages in a critique phase, where they evaluate the ESG implications from different angles. This collaborative process ensures that various aspects of the hack are thoroughly explored, helping to identify overlooked concerns. By refining each other's outputs, the agents collectively fill in information gaps, ultimately improving the accuracy and reliability of the final ESG classification.

4.1.2 Ambiguity

Market volatility is a significant social concern in cryptocurrency due to the general lack of risk disclosure. In one of the articles titled *"Bitcoin and Crypto Closing Out Lame Quarter and One Analyst Believes More Pain Could Be in Store,"* there is a paragraph discussing the market instability observed during the quarter.

> Entering the second quarter, bitcoin was trading just shy of the $71,000 level, and at press time (about 60 hours before Q2 officially closes) was changing hands at $60,800, a decline of more than 14%.

> Dragged down by even larger declines in many altcoins, the broader CoinDesk Index fell more than 21% during the last three months. Among the movers, Solana (SOL) tumbled 30%, Ripple's (XRP) dropped 23%, and Dogecoin (DOGE) plunged 42%. The index's best performer was the above-mentioned ether with its 5% slide.

Figure 4.2: Examples of ambiguity

The sentences above highlight significant price drops in BTC and various altcoins. Despite the system message being designed to capture market instability, not all prompts identify sentences in Figure 4.2 as ESG-related concerns. Upon further analysis of market movements over the past 12 quarters (3 years), BTC, SOL, and DOGE experienced similar price drops in 2 out of the 12 quarters (14%, 30%, and 42%, respectively), while XRP (23%) faced such declines in 3 quarters. These findings suggest that such substantial declines are not unprecedented, underscoring recurring market volatility. However, it remains ambiguous whether this should be considered normal market fluctuations or indicative of deeper market instability relevant to ESG concerns.

This ambiguity suggests that a systematic approach may be more effective in these cases than relying solely on a language model. A systematic approach would involve manually calculating price movements across different periods to determine if observed changes are within expected volatility ranges or if they signal deeper instability. By establishing a quantitative threshold for what constitutes "instability," this method can help clarify whether significant market movements should be classified as ESG concerns.

In terms of labelling, these ambiguities are still considered ESG concerns and are counted towards the result if the model fails to extract them.

4.1.3 Hallucination

Hallucination in generative AI refers to the model's tendency to produce plausible-sounding but incorrect or fabricated content. This issue can manifest in various ways, such as the repetition of phrases or the introduction of public perceptions that are not explicitly discussed in the original text. For instance, the model often introduces discussions about environmental footprints even when they are not directly relevant, sometimes exaggerating or misrepresenting the situation.

Figure 4.3 illustrates how the model incorrectly linked "miner sales" to mining activities, despite the context not supporting this connection. This reflects a heavy bias in the training data. When a model is trained on data that disproportionately emphasizes certain themes, ideas, or perspectives, it may develop a skewed understanding of what is relevant or important. This bias can cause the model to generate outputs that reflect these biases, even when they are not substantiated by the input data.

To mitigate hallucination, fine-tuning the model on a domain-specific dataset containing verified information is crucial. This process helps anchor the model in accurate and relevant knowledge, thereby reducing the likelihood of generating incorrect outputs. Additionally, employing a multi-agent debate approach or further prompting the model to verify facts can improve content reliability by encouraging consistency and ensuring that the responses are grounded in facts rather than assumptions or biases.

Step 1: Identify and explain any Environmental (E) aspects mentioned in the article.
Environmental Aspects:
The article does not explicitly mention any environmental issues related to energy consumption, carbon emissions, resource management, renewable energy usage, or electronic waste production. However, the mention of "miner sales" could imply energy consumption and carbon emissions indirectly, as mining activities are known to consume significant energy and can contribute to carbon emissions.

Step 2: Based on Step 1, extract the original sentences from the article that relate to the Environmental Aspects.
Environmental Array:
```json
The recent price drop was caused by miner sales, some pressure from German-seized BTC, and, of course, the imminent transfer of Mt. Gox coins expected in early July."

```

Figure 4.3: Example of hallucination caused by heavy bias

4.2 Experimental Results

4.2.1 System Message

	ZS System Message	COT System Message	ZS User Prompt	COT User Prompt	ZS No Prompt	COT No Prompt
TP	51	48	46	42	4	38
FP	40	39	47	41	4	45
FN	36	39	41	45	84	49
Precision	0.56	0.55	0.49	0.51	0.50	0.46
Recall	0.59	0.55	0.53	0.48	0.05	0.44
All IoU	0.47	0.45	0.44	0.44	0.05	0.40
Best IoU	0.63	0.60	0.60	0.53	0.04	0.46

Table 4.1: Comparison of Different System Message Placement

We first explore whether the system message effectively guides the model's response towards the desired key ESG issues. To assess this, we tested the Zero-Shot (ZS) and Chain-of-Thought (COT) approaches across three different settings:

1. Using the system message as outlined in Figure 3.4, which provides explicit ESG themes in the model's system message.

2. Embedding the system message directly within the user prompt, instead of using a separate system message.

3. Not using any system message or embedded guidance at all.

The results in Table 4.1 show that Setting 3, which lacks any form of ESG guidance, yielded the lowest performance, with only 4 out of 8 ESG issues correctly identified. In this setting, the model relies solely on its pre-trained knowledge of ESG topics to extract relevant content, but this knowledge may be incomplete or outdated, especially in emerging fields like cryptocurrency. The absence of explicit guidance causes the model to struggle in pinpointing relevant ESG concerns, leading to a reduced ability to accurately extract ESG content.

In contrast, Settings 1 and 2, which include ESG themes in different forms, show significantly improved performance. Both settings offer explicit guidance that helps the model focus on the key ESG issues, leading to higher accuracy in ESG extraction. Setting 1, which places the ESG guidance in the system message, slightly outperforms Setting 2. This difference can be attributed to the model's ability to handle system messages with less cognitive load compared to embedding the guidance within the prompt itself.

When using a system message (Setting 1), the model treats the guidance as a separate, overarching instruction, allowing it to maintain focus on the key themes throughout the processing of the text. In contrast, when the guidance is embedded directly within the user prompt (Setting 2), the model may initially focus on the ESG themes but gradually shift its attention away as it processes more of the input. This gradual drift in attention could lower precision, especially as more text is introduced since the model treats the embedded guide as just another part of the input rather than a persistent instruction.

Overall, the results indicate that explicit guidance—whether via a system message or embedded within the prompt—greatly enhances the model's performance in extracting ESG content. However, placing the guidance in the system message (Setting 1) leads to slightly better results by reducing cognitive load and maintaining focus on the key ESG issues throughout the process.

4.2.2 Zero-Shot (ZS) vs Chain-of-Thought (COT)

When comparing the Chain-of-Thought (COT) and Zero-Shot (ZS) approaches across different settings, distinct patterns in performance emerge. Under Setting 3, where no key ESG themes are provided in the system message or user prompt, ZS demonstrates the lowest performance, likely due to its reliance on a straightforward, left-to-right extraction method. This approach typically identifies sentences based on surface-level matches to its learned embedding space, which causes it to miss more nuanced ESG issues that require deeper contextual understanding.

In contrast, COT significantly outperforms ZS under the same Setting 3, correctly identifying 38 ESG issues compared to the fewer issues captured by ZS. COT's structured reasoning process allows the model to adopt a more holistic approach, considering the broader context and developing an understanding of ESG-related content beyond immediate sentence-level extraction. By breaking down the reasoning process step-by-step, COT enables the model to detect less obvious ESG themes, even when no explicit guidance is provided, as it is able to connect various elements of the article and deduce relevant ESG concerns.

Our quantitative analysis reveals that the ZS approach struggles to capture issues that span multiple sentences, as depicted in Figure 4.4. ZS tends to focus on isolated matches, which limits its ability to detect broader issues that require an understanding of multiple interconnected sentences. On the other hand, COT, with its iterative extraction process, shows the capability to recognize relationships between sentences, allowing it to extract content that aligns with ESG themes in a more comprehensive manner.

However, there are trade-offs. While COT often extracts sentences based on shorter intermediate summaries, this can sometimes lead to overlooking additional sentences that contribute to the same ESG issue, as shown in Figure 4.5. In this case, while COT identified the social aspect of community engagement and extracted key sentences, ZS captured a larger number of relevant sentences within the same theme, extracting more True Positives. This demonstrates that although COT offers a more structured reasoning approach, it may occasionally miss multiple relevant sentences in favour of focusing on core themes.

Overall, while COT tends to outperform ZS in recognizing nuanced ESG issues, the ZS approach still offers value in extracting a higher number of True Positives in some cases. These findings suggest that the choice between ZS and COT should depend on the specific nature of the task—COT is more suited for tasks requiring deeper context and reasoning, while ZS may be preferable for high-level extraction of ESG issues where immediate relevance is more important than structured reasoning.

> Wang, a deputy managing editor, provided the third CoinDesk scoop honored by the Polk Awards: A story revealing that Bankman-Fried and nine co-workers lived together in a luxury Bahamas condominium and at times dated each other while running his companies, including the fact that Bankman-Fried and Alameda CEO Caroline Ellison had once been a couple.
> The piece raised concerns about nepotism, secrecy and conflicts of interest, previewing the tone of a scathing report issued soon after on FTX's extremely lax administrative procedures.

Figure 4.4: COT extracted both sentences, while ZS only captured the latter.

4.2.3 Few-shot

In our experiments, both the FS (Few-Shot) and FS-COT (Few-Shot Chain-of-Thought) approaches underperformed compared to their Zero-Shot (ZS) and COT counterparts. This outcome was unexpected, as few-shot learning typically provides examples that offer additional context and guidance, which should, in theory, improve model performance.

Step 3: Identify and explain any Social (S) aspects mentioned in the article.
Social Aspects:
1. **Market Instability:** The article notes that while some crypto companies have retreated from the U.K., Strike is expanding, which indicates a response to market conditions.
2. **Community Engagement and Inclusion:** Strike's expansion into the U.K. and its commitment to further bitcoin adoption suggests an effort to engage with the community and promote inclusion in the cryptocurrency space.

Step 4: Based on Step 3, extract the original sentences from the article that relate to the Social Aspects.
Social Array:
"'json ["While some crypto companies have retreated from the U.K., Strike said it was expanding its presence and doubling down on its commitment to further bitcoin adoption globally.",
"The company will serve clients cross-border from its European base."
] "'

"Strike, the payments application that uses the Bitcoin blockchain, has started a U.K. service, the company said in a blog post on Tuesday, expanding its global reach just months after rolling out in Europe and Africa.",
"The payments firm has been expanding aggressively and now operates in more than 100 countries and territories around the world.",
"While some crypto companies have retreated from the U.K., Strike said it was expanding its presence and doubling down on its commitment to further bitcoin adoption globally.",
"The company will serve clients cross-border from its European base.",

Figure 4.5: COT extracted 2 sentences related to Community Engagement, while ZS captured more True Positives for the same aspect.

	ZS	ZS-COT	FS	FS-COT	COT-SC	Multiagent Debate
TP	51	48	45	43	58	52
FP	40	39	36	51	65	44
FN	36	39	42	44	29	35
Precision	0.56	0.55	0.56	0.46	0.57	0.54
Recall	0.59	0.55	0.52	0.49	0.65	0.60
All IoU	0.47	0.45	0.46	0.44	0.50	0.47
Best IoU	0.63	0.60	0.55	0.53	0.64	0.64
Cosine Similarity	0.87	0.90	0.86	0.91	0.92	0.92

Table 4.2: Comparison of Different Prompt Engineering Approaches

However, several factors likely contributed to the lower effectiveness of these approaches in extracting ESG issues.

One major factor is the increased cognitive load imposed by the inclusion of few-shot examples. While these examples are designed to aid the model by providing reference points, they can also introduce complexity, particularly when the examples deviate from the context of the system message or the article itself. This discrepancy can cause the model to experience confusion, as it may struggle to reconcile the differences between the few-shot examples and the actual content it is processing. The model's focus may become divided between the provided examples and the task at hand, which could lead to less accurate extractions.

Another contributing factor is recency bias, where the model gives disproportionate weight to the most recent information presented in the prompt. As noted by (30), this bias can cause the model to prioritize the examples that appear closer to the end of the prompt, potentially overlooking earlier instructions or context provided in the system message. This phenomenon may have occurred in our few-shot settings, where the model placed too much emphasis on the examples rather than extracting relevant ESG content from the article itself.

In summary, the few-shot approach did not improve ESG extraction as expected. Instead, it introduced additional complexity and biases that reduced the model's effectiveness. These findings suggest that while few-shot learning can be beneficial in many scenarios, it may not always be the best approach for tasks requiring a nuanced understanding of broad, context-sensitive issues like ESG in cryptocurrency. Careful selection and alignment of examples with the system message are crucial to ensuring that few-shot learning enhances rather than hinders model performance in these cases.

4.2.4 COT-SC

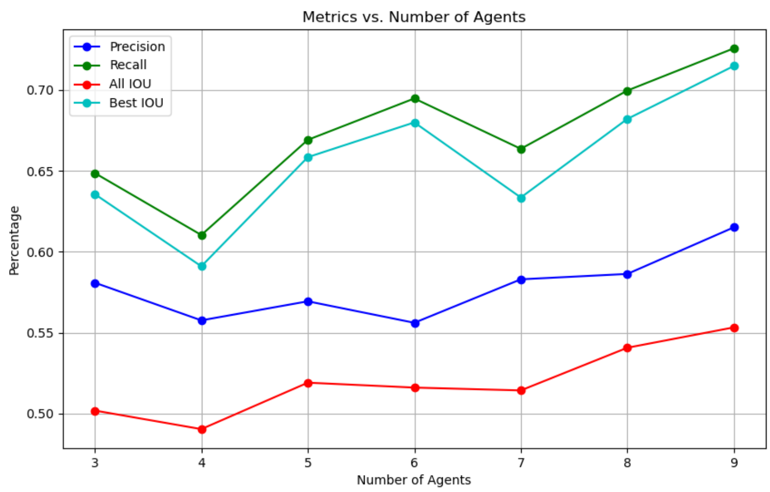

Figure 4.6: COT-SC with different number of agents

We experimented with COT-SC (Chain-of-Thought Self-Consistency) using different numbers of agents and generally observed an increase in recall as more agents were added, as shown in Figure 4.6. However, we also observed drops in recall specifically when the number of agents was set to 4 and 7. This decrease in recall may be attributed to overlapping efforts among agents, where multiple agents extract similar, obvious sentences rather than identifying unique, relevant content. The random initialization of temperature settings for each agent likely contributes to this overlap, as agents may focus on the same easily identifiable ESG-related sentences, which limits the diversity of the output and results in lower recall.

Despite these fluctuations, the overall recall for COT-SC consistently outperformed other methods, such as ZS, COT, and FS, with recall rates reaching up to 65%. This suggests that, while using multiple agents introduces some noise and redundancy, the randomness and increased number of responses ensure that the most relevant sentences are still captured. The diversity in agent responses, though not perfectly optimal, allows for a broader extraction of ESG content, contributing to higher recall.

Interestingly, precision remains relatively stable at 57%, indicating that although multiple agents extract overlapping sentences, this redundancy does not significantly inflate the number of False Positives (FP). The repetition of irrelevant content among agents tends to cluster around the same sentences, meaning that irrelevant information is not overly scattered throughout the results. This balance between recall and precision demonstrates that COT-SC is effective at filtering out noise while still capturing the majority of relevant ESG-related content.

Upon further manual inspection, we found that simply increasing the number of agents does not necessarily lead to a more diverse set of extracted sentences. While recall improves with the addition of agents, the model still tends to generate a limited range of distinct answers, as agents with similar initialization often focus on the same sections of text. This highlights a key limitation: scaling the number of agents alone is not sufficient to achieve true diversity in the outputs or to uncover novel ESG insights. To improve the variety of extracted content, it may be necessary to introduce more varied contextual setups or provide agents with distinct perspectives and prompts, ensuring that each agent focuses on different aspects of the article.

4.2.5 Multiagent Debate

The critique prompts play a critical role in shaping the accuracy and relevance of the system's ability to extract ESG issues. As shown in Table 4.2, the Multi-Agent Debate method effectively reduces the number of false positives (FPs) by 21 compared to COT-SC. However, this reduction comes with a slight trade-off, as true positives (TPs) decrease by 6, leading to an overall increase in precision but a drop in recall. This indicates that while the system becomes more precise, some relevant ESG content is missed.

Through experimentation with different critique prompt settings, we observed a clear trade-off in the system's performance. When the critique prompts are overly strict—for example, instructing agents to rigorously assess the relevance of each sentence and retain only those deemed highly important—the model tends to eliminate both TPs and FPs. While this strict approach helps to filter out irrelevant noise, it also risks discarding valuable ESG-related information. As a result, the system may miss critical content, leading to a lower recall of true ESG issues.

On the other hand, when the critique prompts are too lenient—allowing agents to accept sentences that broadly mention ESG topics—the model captures more TPs but also includes a higher number of FPs. This lenient approach expands the range of content detected by the system and increases recall, but at the cost of precision, as irrelevant information is more likely to be included in the final output.

Striking the right balance between strictness and leniency in critique prompts is essential for optimizing the system's performance. A well-calibrated set of prompts ensures that the agents can distinguish between significant ESG content and irrelevant details. By fine-tuning the critique prompts to be neither too strict nor too lenient, the system can maintain a high level of precision while maximizing the capture of relevant ESG issues. This balance minimizes noise and enhances the accuracy of the extractions, leading to a more reliable identification of ESG concerns in the analyzed articles.

In future work, further refinement of the critique prompts may involve adaptive strategies, where the strictness of the prompts dynamically adjusts based on the complexity of the input. This could help maintain a consistent balance between recall and precision across different types of articles, ensuring that both broad and nuanced ESG issues are effectively captured.

4.2.6 Multiagent Iterative Debate

When analyzing the performance of the Multi-Agent Iterative Debate in Figure 4.7, a general trend of increasing recall and Best IoU, accompanied by decreasing precision and All IoU, becomes evident. This trade-off highlights the system's growing ability to capture more ESG-related sentences over successive iterations (as reflected by increasing recall), while precision declines as more marginally relevant or irrelevant content is included. By the second iteration, recall reaches 77%, continuing to rise beyond 90% in later iterations, but at the expense of precision, which drops below 50% from the second iteration due to the system's expanding coverage of less relevant content. This trend underscores the inherent tension between broadening recall to capture additional true positives and maintaining precision by avoiding false positives.

Best IoU improves early on, showing that the model can capture more relevant content initially, but later declines, indicating that high-quality extractions occur primarily in the first few rounds. This suggests that while the system successfully broadens its exploration, it struggles to maintain the same level of relevance in its extractions as iterations increase.

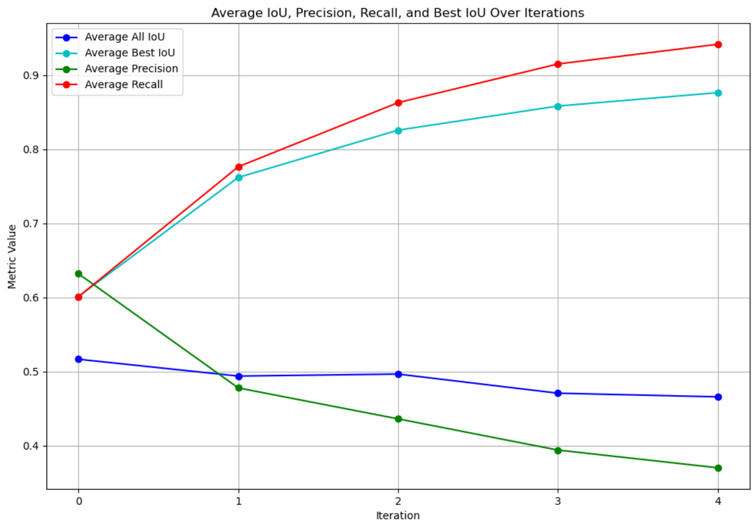

Figure 4.7: Multiagent Iterative Debate performance across different iterations

In Figure 4.8, we observe that the system's improvements begin to plateau after the third iteration. This suggests that most of the salient ESG content is extracted early, leaving fewer opportunities for the model to identify meaningful new content. The significant drop in Intersection over Union (IoU) after the third iteration reflects the introduction of more noise, as the system struggles to find additional relevant sentences. This plateau indicates diminishing returns in subsequent iterations, where the model's effort to capture new content inadvertently increases noise rather than discovering new true positives.

To address these challenges, adaptive critique strategies could be employed to dynamically adjust the strictness of prompts in later iterations. As relevant information tends to be extracted earlier in the process, increasing the strictness of the critique prompts in later rounds would help avoid the substantial decline in Intersection over Union (IoU) observed after the second iteration (Figure 4.8). By making the system more selective over time, this approach can help maintain precision and prevent unnecessary noise from entering the final output.

Moreover, implementing an early termination mechanism could further optimize the system by halting iterations when agents consistently reject or eliminate sentences, signalling that the system has reached its convergence point. This would prevent unnecessary iterations that mainly capture noise, improving both computational efficiency and precision.

Despite some challenges, the Multi-Agent Iterative Debate method stands out as the only approach to achieve over 90% recall with only a slight drop in precision. This demonstrates its effectiveness in broadening recall without significantly compromising accuracy, making it a strong candidate for comprehensive content extraction. However, to maintain

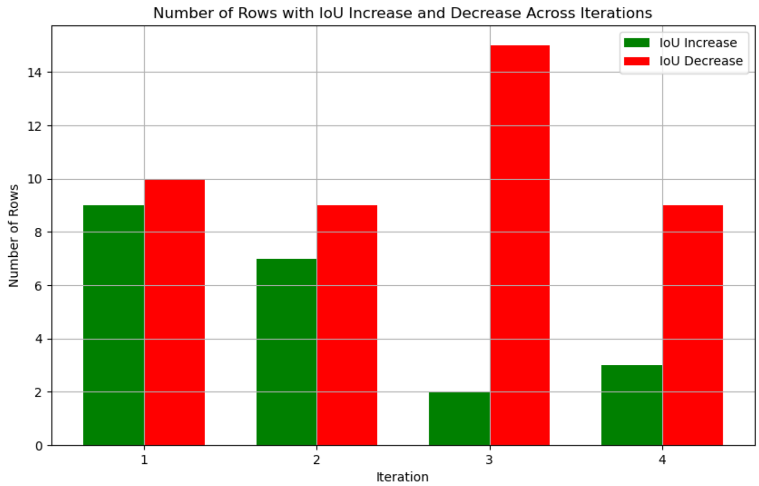

Figure 4.8: Total Increase/Decrease across different iterations

precision across multiple iterations, refining the critique prompts and introducing an early termination mechanism could help reduce noise and mitigate diminishing returns. By improving the critique process, the system could strike a better balance between exploration and precision, offering a more robust solution for extracting nuanced ESG issues.

4.2.7 Cosine Similarity

The results show consistently high cosine similarity for prompts using the Chain of Thought (COT) approach, indicating that COT produces more semantically coherent and consistent outputs. By following a structured reasoning process, COT helps the model maintain focus, reducing fluctuations in its responses and ensuring that the extracted content is more relevant to the task.

This high similarity suggests that COT enables the model to concentrate on the most important aspects of ESG issues, leading to outputs that are more uniform and aligned in terms of content. In contrast, Zero-Shot (ZS) methods, which lack this structured reasoning, tend to produce more variable outputs, with greater fluctuation in relevance and precision. For this reason, COT is often preferable to ZS when semantic consistency and a focused extraction process are critical.

However, while COT ensures coherence, it may limit the model's ability to capture a broader range of ESG issues, particularly those that are less obvious. To balance consistency with diversity, COT can be complemented with strategies like Multi-Agent Debate or adaptive critique refinement, which encourage the model to explore more varied content. This allows for greater coverage of nuanced ESG issues, while still maintaining coherence in the final output. By combining COT with these broader exploration strategies, the system can ensure both consistency and diversity in its extractions, maximizing the quality and comprehensiveness of the results.

4.3 Threats of Validity

Labelling in ESG analysis can significantly impact the performance and evaluation of NLP models due to its subjective nature. Different stakeholders may have varying interpretations of what constitutes an ESG issue, leading to biased labelling where sentences are tagged as ESG-related based on individual perspectives rather than explicit content. For instance, some may consider price fluctuations as indicative of market instability under ESG criteria, while others may not, especially if the fluctuations are typical market behaviour. This kind of assessment might be better handled with more objective, quantitative scripts rather than relying solely on LLMs for interpretation. Even with guidelines like those from HSBC (4), ambiguities can cause inconsistent labelling, skewing the model's performance metrics, and making it challenging to ensure objective evaluation in areas where the relevance of content to ESG criteria is not always clear-cut.

The study's findings are also constrained by the limited dataset, which reduces the confidence intervals around the results and limits the generalizability of the conclusions. Furthermore, the use of closed-source GPT models adds an additional layer of unpredictability. These models may be updated or modified by their developers without notice, leading to changes in behaviour and performance that are difficult to account for in research. This makes reproducibility and consistency in results an ongoing challenge, complicating the evaluation of model performance across different iterations of the same task.

4.4 Other Considerations

When evaluating different approaches for ESG sentence extraction using LLM models, several practical considerations come into play, particularly around token usage, cost, and computational time. More sophisticated models like the Multi-Agent Debate model tend to consume significantly more tokens and therefore increase costs when using proprietary APIs like OpenAI's chatGPT. In terms of computational complexity, methods can be ranked from most to least demanding as Multiagent Iterative Debate > Multi-Agent Debate > Chain of Thought Self-Consistency (COT-SC) > Chain of Thought (COT) > Few-Shot (FS) > Zero-Shot (ZS). Multi-agent systems, which involve several model instances interacting, inherently require more computational resources and time, leading to higher operational costs.

The effectiveness of these approaches also heavily relies on the base model's ability to comprehend what is relevant to key ESG themes in the system message. This understanding directs the model to focus on the correct embedding space, which is crucial for accurately identifying sentences pertinent to ESG criteria. If the base model lacks this foundational understanding, even advanced techniques like multi-agent debate may struggle to yield optimal results. Thus, selecting the appropriate model and method involves balancing performance gains against costs and computational feasibility.

Chapter 5

Conclusion and Future Work

In this study, we explored various prompt engineering techniques and multi-agent collaboration frameworks for extracting Environmental, Social, and Governance (ESG) issues using GPT-4. The results demonstrated that even without fine-tuning, GPT-4 can achieve substantial performance in identifying relevant ESG content, proving that fine-tuning is not always necessary for complex, context-sensitive tasks like ESG analysis.

Among the methods explored, Zero-Shot (ZS) outperformed Chain-of-Thought (COT) in straightforward extraction tasks where context was provided. ZS leveraged the model's pre-existing knowledge to efficiently extract relevant content. However, COT proved more effective in tasks that required deeper reasoning, as its structured process reduced cognitive load and enhanced consistency in more complex scenarios. We also found that Few-Shot (FS) learning, while typically useful, requires careful selection of examples. Misaligned examples can introduce noise and confusion, reducing the model's effectiveness rather than improving it.

The critique phase in the Multi-Agent Debate framework proved useful in reducing noise, but we found that critique prompts need to be meticulously designed. If the prompts are too strict, they may lower true positives; too lenient, and they may increase irrelevant content. Additionally, while scaling the number of agents improved recall by capturing more relevant content, it became clear that ensuring diversity in agents' outputs is crucial for capturing nuanced ESG issues that might otherwise be overlooked.

Finally, the Multi-Agent Iterative Debate method outperformed other approaches in its ability to learn and refine outputs across iterations. By allowing agents to learn from previous results, this iterative process enhanced recall, but it also highlighted the need for an early termination mechanism. When the answer converges, further iterations tend to introduce noise rather than new relevant content.

Additionally, the methods proposed in this study, including prompt engineering and multi-agent collaboration, are not limited to ESG extraction. With appropriate modifications, such as adapting the system message and addressing domain-specific challenges, these techniques can be extended to other emerging fields that face similar issues of scarce labelled data and complex content extraction. Future research can build on these results to further enhance model precision, adapt critique systems dynamically, and refine iterative processes to handle complex, nuanced content more effectively.

Our research made several key contributions:

- **Exploration of Prompt Engineering**: We systematically compared the strengths and weaknesses of various prompt engineering methods for GPT-4, providing a comprehensive understanding of how different approaches impact the model's performance. This exploration offers a systematic approach to achieving reliable results, allowing practitioners to choose the most appropriate method for their specific needs.

- **Advanced Multi-Agent Collaboration Setting**: We explored advanced multi-agent collaboration settings, showcasing how multiple instances of GPT can work together to refine outputs and improve the quality of ESG issue extraction. This collaborative framework highlights the potential for multi-agent systems to enhance the capabilities of language models in complex analytical tasks.

- **Addressing Current Limitations**: Our study also addressed some of the current limitations in using GPT for ESG analysis. By identifying areas where the model struggles, such as dealing with ambiguous or nuanced ESG content, we provide a clear path for future improvements and adaptations. This helps in understanding the boundaries of GPT's capabilities and informs strategies for overcoming these challenges.

- **Accuracy Measures without Labelled Data**: We demonstrated how manual verification and semantic similarity can be used to measure accuracy in the absence of labelled data. Alongside cosine similarity, we used Intersection over Union (IoU), precision, and recall to evaluate the model's performance, providing a comprehensive and rigorous assessment of its ability to extract relevant ESG content.

By systematically comparing prompt engineering methods, exploring multi-agent collaboration, addressing current limitations, and demonstrating different evaluation metrics, this study contributes valuable insights into the use of GPT for ESG analysis. It lays the groundwork for future research and development, pointing to several promising directions for enhancing the capabilities of language models in this important area.

5.1 Future Work

While our findings are promising, there are several avenues for future research and development to enhance the effectiveness and accuracy of GPT models in extracting ESG issues:

- **Domain-Specific Adaptation through Unsupervised Pre-training**: To enhance the model's ability to understand nuanced ESG terminology and context, a promising direction for future work is to leverage unsupervised pre-training. This technique involves further pre-training the GPT model on a domain-specific corpus, such as ESG reports, CSR documents, and financial disclosures, without requiring any labelled dataset. By employing objectives like Masked Language Modeling (MLM) or Next Sentence Prediction (NSP), the model's embedding space is updated to better reflect the language and concepts specific to the ESG domain. This approach aligns with our strategy of avoiding fine-tuning, as it does not necessitate labelled data and retains the model's general versatility while enhancing its performance in specialized tasks. Through this adaptation, the model could become more adept at recognizing ESG-specific terminology and better equipped to extract relevant issues, thus improving its utility and accuracy in identifying ESG-related content.

- **Reinforcement Learning with AI Feedback**: Building on the multi-agent collaboration framework, a potential extension involves implementing Reinforcement Learning (RL) to further refine the model's performance. Once we have access to the model's internal layers or gradients, we could employ RL techniques to train the model more effectively. In this approach, AI critiques would serve as a form of feedback or reward signal that guides the model towards more accurate extractions. By penalizing the model for generating irrelevant or noisy outputs (negative rewards) and rewarding it for correctly identifying ESG issues (positive rewards), we can create a reinforcement learning loop where the model iteratively learns to minimize errors and reduce hallucinations. Over time, this policy optimization process would enable the model to improve its decision-making capabilities and accuracy in identifying ESG content by reinforcing desirable behaviours and discouraging undesired ones.

- **Systematic Investigation of Prompt Phrasing and Keyword Scope**: A focused exploration on understanding the scope and influence of specific keywords within the system message could significantly enhance prompt engineering strategies. For example, determining what the term "market instability" encompasses and how it is interpreted by the model can provide critical insights into its contextual understanding and application range. By systematically testing and analyzing how different phrasings and keywords affect the model's outputs, we can map the semantic boundaries and nuances that each term covers. This approach will help refine prompt design, ensuring more comprehensive and precise extraction of ESG issues, and ultimately improving the model's utility in various real-world scenarios.

- **Combining Retrieval-Augmented Generation (RAG) with Named En-**

tity Recognition (NER): Future research could explore integrating Named Entity Recognition (NER) with Retrieval-Augmented Generation (RAG) to further improve the model's performance. By first applying NER to extract relevant entities from the text, such as companies, locations, and key individuals, we can focus the retrieval process on content that directly pertains to these entities. The extracted information can then be stored in a database organized by embedding space, enabling efficient and context-aware information retrieval. When querying the system, this approach allows for more precise retrieval of ESG-related issues by asking entity-specific questions, such as "What are the relevant ESG issues of X entity?" This combination of NER and RAG leverages the strengths of both techniques to provide more targeted and relevant responses, enhancing the model's accuracy and contextual relevance in extracting ESG data.

By pursuing these future directions, we can further advance the capabilities of language models like GPT in the domain of ESG analysis, making them more robust, accurate, and applicable across various industries and contexts.

Bibliography

[1] W. Rigali, "Esg and cryptocurrency: Considerations for market participants," 2021. [Online]. Available: https://www.theracetothebottom.org/rttb/2021/8/31/esg-and-cryptocurrency-considerations-for-market-participants

[2] Rocknblock.io, "Blockchain ecosystem grants list," 2023. [Online]. Available: https://rocknblock.io/blog/blockchain-ecosystem-grants-list

[3] M. Hochstein, N. Baker, J. Hamilton, R. Perper, S. Reynolds, C. Ligon, F. Munawa, C. Thompson, A. Baydakova, and D. Nelson, "Finding a way forward for crypto, web3 and digital assets: Consensus report 2023," 2023. [Online]. Available: https://www.coindesk.com/layer2/2022/03/09/people-will-get-burned-matt-odell-on-the-long-road-to-bitcoin-privacy/

[4] C. Sarmiento, A. Tyler, and H. G. Research, "The esg of cryptocurrencies: Risks and opportunities," May 2022. [Online]. Available: https://www.research.hsbc.com

[5] M. Alqudah, L. Ferruz, E. Martín, H. Qudah, and F. Hamdan, "The sustainability of investing in cryptocurrencies: A bibliometric analysis of research trends," *International Journal of Financial Studies*, vol. 11, no. 3, p. 93, 2023. [Online]. Available: https://doi.org/10.3390/ijfs11030093

[6] Candriam, "Esg in cryptocurrencies: Assessing the impact of digital assets on sustainability," 2022. [Online]. Available: https://www.candriam.com/

[7] CCData, "Esg rankings," 2023. [Online]. Available: https://ccdata.io/research/esg-rankings

[8] Crypto Carbon Ratings Institute, "Ccri indices," https://indices.carbon-ratings.com/, accessed: 2024-07-11.

[9] CCData and C. C. R. I. (CCRI), "Esg benchmark report," 2023, accessed: 2024-07-11. [Online]. Available: https://ccdata.io/research/esg-rankings

[10] C. C. R. I. (CCRI), "Methodologies to calculate the proposed mandatory sustainability indicators required by the eu markets in crypto-assets (mica)

regulation," 2024, accessed: 2024-07-11. [Online]. Available: https://carbon-ratings.com/

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems.* NIPS, 2017, pp. 6000–6010.

[14] S. Mehra, R. Louka, and Y. Zhang, "Esgbert: Language model to help with classification tasks related to companies' environmental, social, and governance practices," in *Proceedings of the EMSA, SEA, AIFU, NLCAI, BDML, BIoT, NCOM, CLOUD, CCSEA, SIPRO 2022*, 2022, pp. 183–190.

[15] S. Pasch and D. Ehnes, "Nlp for responsible finance: Fine-tuning transformer-based models for esg," in *2022 IEEE International Conference on Big Data (Big Data).* IEEE, 2022, pp. 3532–3536.

[16] N. Raman, G. Bang, and A. Nourbakhsh, "Mapping esg trends by distant supervision of neural language models," *Machine Learning and Knowledge Extraction*, vol. 2, no. 4, pp. 453–468, 2020. [Online]. Available: http://www.mdpi.com/journal/make

[17] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[18] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi, "Revisiting few-sample bert fine-tuning," in *International Conference on Learning Representations (ICLR)*, 2021.

[19] G. Colavito, F. Lanubile, N. Novielli, and L. Quaranta, "Leveraging gpt-like llms to automate issue labeling," in *21st International Conference on Mining Software Repositories (MSR '24).* ACM, 2024, p. 12.

[20] J. Wei *et al.*, "Chain of thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, 2022.

[21] Y. Du *et al.*, "Improving factuality and reasoning in language models through multiagent debate," *MIT CSAIL*, 2023.

[22] Y. Bai, A. Chen, A. Askell, N. DasSarma, A. Jones, M. Mazeika, S. Kundu, E. Perez, T. Lattimore, U. Mini, S. Ringer, Z. Hatfield-Dodds, S. Kadavath, T. Cai, R. Greenblatt, L. Reynolds, J. Muller, J. Uesato, S. Bowman, O. Evans, and D. Hendrycks, "Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback," 2024.

[23] D. Araci, "Finbert: A pretrained language model for financial communications," *arXiv preprint arXiv:1908.10063*, 2019.

[24] T. Schimanski, A. Reding, N. Reding, J. Bingler, M. Kraus, and M. Leippold, "Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication," *Finance Research Letters*, vol. 61, p. 104979, 2024.

[25] Y. Zou, M. Shi, Z. Chen, Z. Deng, Z. Lei, Z. Zeng, S. Yang, H. Tong, L. Xiao, and W. Zhou, "Esgreveal: An llm-based approach for extracting structured data from esg reports," *arXiv preprint arXiv:2312.17264*, 2023.

[26] H. J. Yun, C. Kim, M. Hahm, K. Kim, and G. Son, "Esg classification by implicit rule learning via gpt-4," *arXiv preprint arXiv:2403.15040*, 2024. [Online]. Available: https://arxiv.org/abs/2403.15040

[27] CoinDesk, "Coindesk - leader in cryptocurrency, blockchain, and digital finance news," https://www.coindesk.com/, accessed: 2024-09-06.

[28] Anirudh, "G4f - an openai free api wrapper," https://github.com/techwithanirudh/g4f, accessed: 2024-09-06.

[29] IBM, "Model parameters and prompting - ibm documentation," https://www.ibm.com/docs/en/watsonx/saas?topic=lab-model-parameters-prompting, 2024, accessed: 2024-09-06.

[30] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *arXiv preprint arXiv:2307.03172*, 2023.

[31] X. Wang *et al.*, "Self-consistency improves chain of thought reasoning in language models," *Proceedings of the International Conference on Learning Representations*, 2023.