

Analysing Modern Slavery Statements (MSS) Using Large Language Models (LLMs)

Ser-Huang Poon¹, Eghbal Rahimikia, Philip Jobi Vallavantra, Siliang Wei

Ser-Huang Poon (ser-huang.poon@manchester.ac.uk, ORCID: 0000-0002-7297-9401)
Eghbal Rahimikia (eghbala.rahimikia@manchester.ac.uk, ORCID: 0000-0001-6583-2332)
Philip Jobi Vallavantra (philipjobi.vallavantra@postgrad.manchester.ac.uk)
Siliang Wei (siliang.wei@manchester.ac.uk)

Abstract:

This chapter explores the application of Large Language Models (LLMs) to analyse Modern Slavery Statements (MSS). As businesses are increasingly required to produce MSS to comply with legislation, the volume of these documents presents challenges in monitoring and evaluation. LLMs offer a scalable solution for extracting, classifying, and assessing the content of MSS, enabling organisations and regulators to identify compliance issues and patterns of modern slavery risks. By leveraging LLMs, the analysis can go beyond simple keyword searches to understand the context and nuances within MSS, offering a more comprehensive insight into corporate responses to modern slavery. This chapter demonstrates the potential of LLMs to enhance transparency and accountability in corporate reporting on human rights issues, particularly in addressing modern slavery.

Keywords:

Modern Slavery Statements, Large Language Models, Corporate Transparency, Ethical AI, Corporate Accountability, Natural Language Processing

¹ Corresponding author. All four authors are affiliated with the Alliance Manchester Business School, University of Manchester, UK. We wish to express our gratitude to Lijing Fei, Shenglan Jin, Zhenru Liu, Kunxiang Peng, and Qinwen Zheng for their invaluable assistance in refining the LLM prompts, and to the Turing LLM Reading Group for their helpful suggestions. We are also grateful to Jürgen Rudolph (editor) and the two anonymous reviewers for their constructive comments. Additionally, we thank Matthew Pritchard for his work in collating modern slavery statements and testing the LLM prompts. We acknowledge with thanks the funding support provided by the Faculty of Humanities' Social Responsibility in the Curriculum initiative.

1. Introduction

Academic research and higher education can significantly contribute to advancing social good, corporate social responsibility (CSR), and sustainability (Rasche, 2010). This article, and the initiative it supports, is primarily focused on improving AI skills education and training within higher education. A key objective is to extend this training to include non-profit sectors and initiatives dedicated to social good. Specifically, the article demonstrates how we train postgraduate students to use AI tools for tasks such as analysing modern slavery statements (MSS) through textual analysis.

Modern slavery encompasses various forms of exploitation, including forced labour, human trafficking, and other severe violations of human rights (Bales, 2007; Kara, 2017). Siddharth Kara, in *Modern Slavery: A Global Perspective*, differentiates the overlapping categories of (modern) slavery, forced labour, human trafficking, and debt bondage/bonded labour, and estimates the “number of slaves at the end of the year 2016” to be “31.2 million” (Kara, 2017, p. 18). The UK Modern Slavery Act 2015 represents a significant legislative effort to address this issue within business operations and supply chains (Skrivankova, 2010). This Act mandates that all businesses with a global annual turnover exceeding £36 million must publish annual Modern Slavery Statements (MSS) (UK Home Office, 2015). These statements should detail the steps taken by these businesses to ensure that slavery and human trafficking are not occurring in their operations or supply chains.

The requirement for MSS aims to increase transparency and accountability among large corporations, encouraging them to implement and report on effective measures to prevent modern slavery (New, 2015; Crane et al., 2017). However, traditional methods of analysing MSS, which often involve manual reviews, can be time-consuming, inconsistent, and limited in scope (LeBaron and Rühmkorf, 2017). Thus, there is a need for more efficient and scalable approaches to evaluate these statements, ensuring they genuinely reflect corporate efforts to combat modern slavery.

Large Language Models (LLMs), such as those described by Radford et al. (2019), have emerged as powerful tools in the field of natural language processing (NLP). These models, trained on vast datasets, possess the ability to process and analyse extensive volumes of text data (Brown et al., 2020). Their capability to understand context, generate coherent text, and perform complex analyses makes them particularly suitable for evaluating MSS.

The versatility of LLMs lies in their unsupervised multitask learning abilities, allowing them to perform a wide range of linguistic tasks without requiring task-specific training (Devlin et al., 2019). This is achieved through their ability to learn patterns and structures in language from diverse and extensive datasets. However, fine-tuning these models on specific tasks or domains can lead to even better results by tailoring the model’s knowledge to the particular nuances and requirements of the task (Howard and Ruder, 2018). The model introduced by Radford et al. (2019), for instance, demonstrates these

capabilities, showcasing how LLMs, with or without fine-tuning, can be effectively applied to various tasks, including the analysis of MSS.

Analysing MSS is crucial for several reasons. Firstly, it helps ensure that companies are held accountable for their actions and commitments to combat modern slavery (Bales & Soodalter, 2009). By scrutinising these statements, stakeholders can identify gaps and areas for improvement, promoting higher standards of corporate behaviour (Crane, 2013). Secondly, a thorough analysis of MSS can highlight best practices and effective strategies that can be adopted by other companies, fostering a collaborative approach to eradicating modern slavery (Gold et al., 2015). Thirdly, evaluating MSS can provide valuable insights for policymakers, helping them understand the effectiveness of current regulations and identify areas where further legislative action may be needed (LeBaron and Rühmkorf, 2017).

The integration of LLMs into the analysis process offers significant advantages. Traditional methods of assessing MSS often involve manual reviews, which can be labour-intensive and prone to inconsistencies (Jackson et al., 2020). In contrast, LLMs can process large volumes of text quickly and consistently, applying the same criteria across all statements. This automation not only enhances efficiency but also ensures that the analysis is scalable, capable of handling the increasing number of MSS submissions each year (Brown et al., 2020).

The implementation of LLMs in MSS analysis involves several key steps. Firstly, researchers must curate a dataset of MSS from reliable sources such as the Modern Slavery Registry or the companies' public-facing websites (BHRRCC, 2023). This dataset provides the foundation for the analysis. Next, the text data is pre-processed to standardise and clean it, although MSS are typically well-written formal statements with few typographical errors, reducing the need for extensive data preparation.

Once the data is prepared, the LLM is deployed using carefully crafted prompts and a labelled subset of MSS, where the statements have been annotated according to a recognised methodology, such as the Business & Human Rights Resource Centre (BHRRCC) FTSE100 scoring framework (BHRRCC, 2023). This approach allows the LLM to adapt to the specific nuances of MSS, guiding the model to produce more accurate and relevant outputs by aligning its analysis with the established scoring criteria.

One of the primary advantages of using LLMs for MSS analysis is their ability to automate the evaluation process (Marcus and Davis, 2020). This ensures consistency and scalability across large volumes of data, allowing for a thorough and reliable assessment of MSS. Additionally, LLMs can provide detailed insights into the content of MSS, identifying patterns and trends, highlighting areas of compliance and non-compliance, and even generating summaries and recommendations. These insights can help researchers, policymakers, and other stakeholders better understand how companies are addressing modern slavery issues and where improvements may be needed (Marcus and Davis, 2020).

Given the significant challenges associated with traditional methods of evaluating MSS, this paper aims to explore the application of LLMs to automate and enhance the analysis process. By leveraging the advanced capabilities of LLMs, we seek to develop a robust framework for assessing the quality and comprehensiveness of MSS. This approach not only addresses scalability issues but also ensures a more consistent and objective evaluation, contributing to the overall effort to eradicate modern slavery (Radford et al., 2019).

The next section provides an overview of the BHRRC FTSE100 scoring methodology, a comprehensive framework used to evaluate MSS. Following this, we delve into the specifics of implementing LLMs in MSS analysis, including data preparation and preprocessing techniques. The paper then discusses designing effective prompts for LLM analysis, followed by the scoring and evaluation process. We also address the challenge of strategic ambiguities in MSS and conclude with a comparative analysis and benchmarking of LLM-generated scores against human evaluations, offering insights and recommendations for future research and practice.

2. BHRRC FTSE100 Scoring Methodology

The Business & Human Rights Resource Centre (BHRRC) published a comprehensive scoring methodology in 2018 and demonstrated its application using the MSS from FTSE100 constituent firms at that time. This methodology, based on the UK Modern Slavery Act 2015, is designed to evaluate the extent and quality of information provided in the MSS regarding companies' efforts to combat modern slavery and human trafficking within their business operations and supply chains. The BHRRC framework focuses on six key reporting areas: the organisation's structure, relevant policies, due diligence processes, risk assessment and management, measuring effectiveness, and training.

a) Organisation's Structure, Its Business, and Its Supply Chains

This section aims to understand how companies are organised and how their supply chains are structured. Companies are expected to provide a detailed description of their business model, including the sectors in which they operate and the geographic locations of their operations. Furthermore, they should disclose the key components of their supply chains, identifying the countries from which they source goods and services. This information is crucial as it helps to assess the complexity and reach of a company's supply chain, which can significantly impact the risk of modern slavery.

b) Policies Relevant to Slavery and Human Trafficking

The second area focuses on the policies that companies have implemented to address slavery and human trafficking. Effective policies are essential for setting the tone and expectations for corporate behaviour regarding modern slavery. Companies are evaluated on whether they have specific policies that prohibit slavery and human

trafficking, and how these policies are communicated to employees, suppliers, and other stakeholders. Additionally, the BHRRC methodology assesses the scope and comprehensiveness of these policies, including whether they cover all aspects of the company's operations and supply chain. The existence of a clear and robust policy framework is indicative of a company's commitment to tackling modern slavery.

c) Due Diligence Processes

Due diligence processes are critical for identifying, preventing, and mitigating the risks of modern slavery within a company's operations and supply chain. The BHRRC scoring methodology examines whether companies have established due diligence procedures to assess and address modern slavery risks. This includes evaluating the mechanisms in place for conducting risk assessments, supplier audits, and other investigative measures. Companies are also scored on the extent to which they engage with stakeholders, including workers and civil society organisations, in their due diligence processes. Effective due diligence is a proactive approach to managing modern slavery risks.

d) Risk Assessment and Management

Risk assessment and management concern how companies identify and manage the risks of modern slavery in their operations and supply chains. Companies are expected to provide a detailed description of their risk assessment processes, including the criteria and methods used to identify high-risk areas. The BHRRC methodology also looks at the steps companies take to mitigate these risks, such as implementing corrective action plans, improving supply chain transparency, and engaging in collaborative initiatives with other stakeholders. The effectiveness of a company's risk management strategy is a key indicator of its ability to address modern slavery issues.

e) Measuring Effectiveness

The focus here is on measuring the effectiveness of the actions taken to combat modern slavery. Companies are assessed on whether they have established performance indicators and metrics to evaluate the impact of their anti-slavery efforts. This includes tracking the number of audits conducted, the number of violations identified, and the remediation measures implemented. Additionally, the BHRRC methodology considers whether companies report on the outcomes of their initiatives, such as improvements in labour conditions and reductions in modern slavery incidents. Transparent and rigorous measurement of effectiveness is essential for demonstrating accountability and progress in combating modern slavery.

f) Training

The final area examines the training provided to employees, suppliers, and other stakeholders on modern slavery issues. Effective training programs are vital for raising awareness and building the capacity of individuals to identify and address modern slavery risks. Companies are evaluated on the scope and content of their training programs, including whether they cover relevant topics such as identifying signs of modern slavery, understanding company policies, and implementing due diligence procedures. The methodology also considers the frequency and reach of the training, as well as the mechanisms in place for evaluating its effectiveness. Comprehensive training is a key component of a company's strategy to prevent and combat modern slavery.

The BHRRC framework includes 54 main questions and 8 optional questions, each designed to evaluate specific aspects of the statements. Scores are assigned based on the comprehensiveness and transparency of the information provided. This rigorous scoring system helps to identify best practices and areas for improvement, ultimately driving higher standards of corporate accountability and transparency.

By adopting the BHRRC FTSE100 scoring methodology, researchers and stakeholders can systematically evaluate the quality and effectiveness of Modern Slavery Statements. This evaluation is crucial for holding companies accountable and ensuring that they take meaningful steps to eradicate modern slavery from their operations and supply chains.

3. Implementing LLMs in MSS Analysis

The advent of LLMs has revolutionised the field of NLP, providing unprecedented capabilities for understanding and generating human-like text (Radford et al., 2019). These models, trained on vast datasets, possess the ability to process and analyse extensive textual data, making them particularly suitable for evaluating MSS. By leveraging LLMs, researchers can automate the assessment process, ensuring consistency and scalability across large volumes of data.

LLMs, such as those developed by OpenAI, Google, Meta and other leading research institutions, are designed to perform a wide range of linguistic tasks without requiring task-specific training. This versatility is achieved through their ability to learn patterns and structures in language from diverse and extensive datasets. For example, the model described by Radford et al. (2019) in "Language Models are Unsupervised Multitask Learners" demonstrates the capability of LLMs to understand context, generate coherent text, and perform complex analyses based on unsupervised learning techniques.

An early notable example of an LLM is BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2019). BERT utilises a deep bidirectional transformer architecture to pre-train on a vast corpus of text, enabling it to capture nuanced language representations. This model excels in various NLP tasks, including text

classification, question answering, and language inference, making it an invaluable tool for analysing MSS. The bidirectional nature of BERT allows it to consider the context from both preceding and succeeding words, leading to a more comprehensive understanding of the text.

Similarly, the transformer architecture, as described by Vaswani et al. (2017) in “Attention Is All You Need,” forms the backbone of many LLMs. This architecture employs a mechanism known as self-attention, which enables the model to weigh the importance of different words in a sentence relative to each other. By doing so, transformers can effectively capture long-range dependencies and intricate relationships within the text. This capability is crucial for analysing MSS, where understanding the interplay between various components of a statement can provide deeper insights into a company's compliance with modern slavery regulations.

Implementing LLMs in MSS analysis involves several key steps. First, researchers must curate a dataset of MSS from reliable sources, such as the Modern Slavery Registry or the company public facing websites. In classical NLP analysis, the step after data collection is to preprocess the text to standardise and clean the data. However, since MSS are well-written formal statements with a negligible number of typos, and most LLMs can directly process multiple file formats, there is significantly less need for extensive data preparation in this context.

After preprocessing, the LLM analysis is deployed for the specific task of MSS analysis. Although we do not have control over the LLM itself, we can improve the results by refining the prompt design. This fine-tuning involves enhancing the prompts using a labelled subset of MSS, where the statements have been annotated according to the BHRRC methodology. This process allows the prompt design to adapt to the specific nuances of MSS.

One of the primary advantages of using LLMs for MSS analysis is their ability to automate the evaluation process. Traditional methods of assessing MSS often involve manual reviews, which can be time-consuming and prone to inconsistencies. In contrast, LLMs can process large volumes of text quickly and consistently, applying the same criteria across all statements. This automation not only enhances efficiency but also ensures that the analysis is scalable, capable of handling the increasing number of MSS submissions each year.

Furthermore, LLMs can provide detailed insights into the content of MSS. By analysing the text, these models can identify patterns and trends, highlight areas of compliance and non-compliance, and even generate summaries and recommendations. This level of analysis can help researchers and policymakers better understand how companies are addressing modern slavery issues and where improvements may be needed.

In conclusion, implementing LLMs in MSS analysis offers significant advantages in terms of efficiency, consistency, and scalability. By leveraging the advanced capabilities of models like BERT and transformer-based architectures, researchers can automate the assessment of MSS, ensuring a thorough and reliable evaluation. The insights gained

from this analysis can inform policy decisions and drive improvements in corporate practices, ultimately contributing to the global effort to combat modern slavery.

4. Data Preparation and Preprocessing

This section outlines the processes involved in collecting, cleaning, and preparing data for a comprehensive examination of MSS.

The initial step in this endeavour is to collect MSS from credible and authoritative sources. One such reliable source is the Modern Slavery Registry, which aggregates statements from a multitude of companies required to comply with the UK modern slavery regulations. For larger companies, MSS can also be found and downloaded from their official websites, though this tends to cover only the most recent version of the statements.

In 2024 alone, 20,000 companies are expected to submit MSS. Given the potential for discrepancies and the varied ways companies may present their names, it is essential to match the reports with the company names in the government register accurately. To address this, we deploy Cleanco, an NLP tool designed to clean company names by removing legal designations such as “Ltd,” “Plc,” “Inc.,” and other suffixes that might otherwise introduce inconsistencies.²

After cleaning the company names, it is necessary to accurately match these names with their corresponding MSS. For this task, the Jaro-Winkler Similarity metric proves to be highly effective. The Jaro-Winkler metric is a string comparison algorithm that measures the similarity between two strings, allowing for the detection of typos and minor discrepancies. By applying this metric, we can ensure that each MSS is correctly associated with the intended company, thereby enhancing the integrity of the dataset.

Here, for illustration, we collected MSS for a sample of 13 companies from five sectors as shown below:

- Multiline retail, clothing & textile: MKS, Next, and Burberry
- Food & staples retailing, supermarkets & grocery: Sainsbury, Tesco, and Morrison
- Hotels, restaurants & leisure: InterContinental Hotels Group Plc
- Tobacco: British American Tobacco plc, and Imperial Brands Plc
- Food & beverage: Associated British Foods Plc, Coca-Cola, Diageo Plc, and Unilever Plc

This small but diverse sample aids in understanding sector-specific challenges and responses, thereby providing deeper insights into the effectiveness of a single set of LLM prompts for universal MSS analysis.

² The package Cleanco for cleaning company names is available on GitHub (<https://github.com/psolin/cleanco>) and can also be installed via PyPI (Python Package Index, <https://pypi.org/project/cleanco/>). One can include can install it using pip with the following command: pip install cleanco.

5. Designing Effective Prompts for LLM Analysis

Designing effective prompts for LLMs is essential to extract meaningful and accurate information from MSS in line with the BHRRC FTSE100 scoring methodology. This section provides detailed insights and examples to illustrate the best practices in prompt design.

Creating relevant prompts is the first step in ensuring that the LLM can generate comprehensive responses addressing all aspects of the questions posed. The BHRRC methodology includes 62 questions covering specific information about various dimensions of a company's operations, such as its structure, supply chains, policies, and risk assessments. For a detailed analysis, prompts can be designed to elicit precise and detailed information, which is crucial for evaluating the company's compliance with modern slavery regulations. However, for a large-scale automated process, we must split the prompts to return specific quantifiable or simple limited-range responses for machine evaluation. This is the approach we adopt.

Prompt engineering is the practice of designing and refining input prompts to guide an LLM in generating desired outputs. It is crucial because well-crafted prompts can significantly enhance the model's performance, ensuring that responses are relevant, accurate, and aligned with specific goals. Where appropriate, Chain-of-Thought (CoT) principles are applied in designing prompts to produce coherent and logically ordered responses (see Wang et al. 2023). These principles ensure that the LLM's outputs follow a structured approach, progressively building upon each piece of information according to the BHRRC framework. Gao et al. (2023) suggest that prompting for main points and explanations of relationships before requesting the LLM to conclude or make more complex inferences can yield better results. This approach mirrors the effectiveness of well-prepared in-context learning.

Using clear language is another vital aspect of effective prompt design. The language used in prompts should be straightforward and unambiguous to avoid any confusion. This clarity helps the LLM understand the exact information required, thereby improving the accuracy and relevance of the responses. This approach differs from using LLMs to detect strategic ambiguity, a practice employed to avoid accountability and narrow the scope of responsibility. The investigation of strategic ambiguity involves advanced NLP techniques, which will be further elaborated in Section 7.

Balanced detail is crucial in prompt design. While it is important to provide enough detail to guide the LLM's response, overloading the prompt with excessive information can lead to confusion and less effective outputs. The prompts should strike a balance, providing sufficient context and detail without overwhelming the LLM. This approach ensures that the responses are both comprehensive and manageable. Furthermore, Gao et al. (2023) suggest using close-ended rather than open-ended prompts, prompting for explicit rather than implicit causal relationships, and being specific with causal concepts in the prompts.

Through continuous refinement and evaluation, a set of effective prompts can be developed. These prompts should be capable of consistently extracting the required information from MSS, ensuring that the analysis aligns with the BHRRC FTSE100 scoring methodology. The iterative process of designing, testing, and refining prompts is crucial in achieving this goal. By following the principles of relevance, coherence, key elements, clear language, and balanced detail, researchers can develop a robust set of prompts that enhance the accuracy and reliability of LLM analysis in evaluating Modern Slavery Statements.

6. Scoring and Evaluation

The scoring and evaluation process is crucial for analysing MSS using LLMs. Our goal is to train LLMs to score MSS rather than conduct a qualitative assessment of reporting standards, which is beyond this paper's scope. This task begins with designing prompts that align with the BHRRC FTSE100 scoring methodology. Once calibrated, these prompts can be utilised in automated scoring algorithms. Here, we explain the process of developing a consistent set of prompts applicable to all MSS.

First, prompts are crafted to elicit detailed LLM responses, ensuring all aspects of the questions are addressed. For example, “Explain the company’s due diligence processes to prevent modern slavery, including how it monitors compliance and works with suppliers” directs the LLM to provide comprehensive information. Manual calibration, involving human annotators comparing LLM outputs with benchmarks, validates and refines these prompts. The BHRRC FTSE methodology seeks binary (Yes/No) responses for easier scoring. Hence, the question becomes, “Does the company’s due diligence process include monitoring compliance and working with suppliers?” which is easier to evaluate and automate.

In evaluating LLM capability in assessing modern slavery statement disclosures, we use LLM to score companies' responses to 54 operational questions across six reporting areas using a 3-point scale (0, 0.5, 1). We then compare LLM-generated scores with human scores to assess grading accuracy. Since correctly matching the “1” (good measure) cases and the “0” (lack of consideration) cases are equally important, we use Accuracy and Cohen’s kappa to evaluate LLM performance. These metrics are calculated for each question, reporting area, and the entire statement, helping us identify and target areas where LLM prompts may need revision.

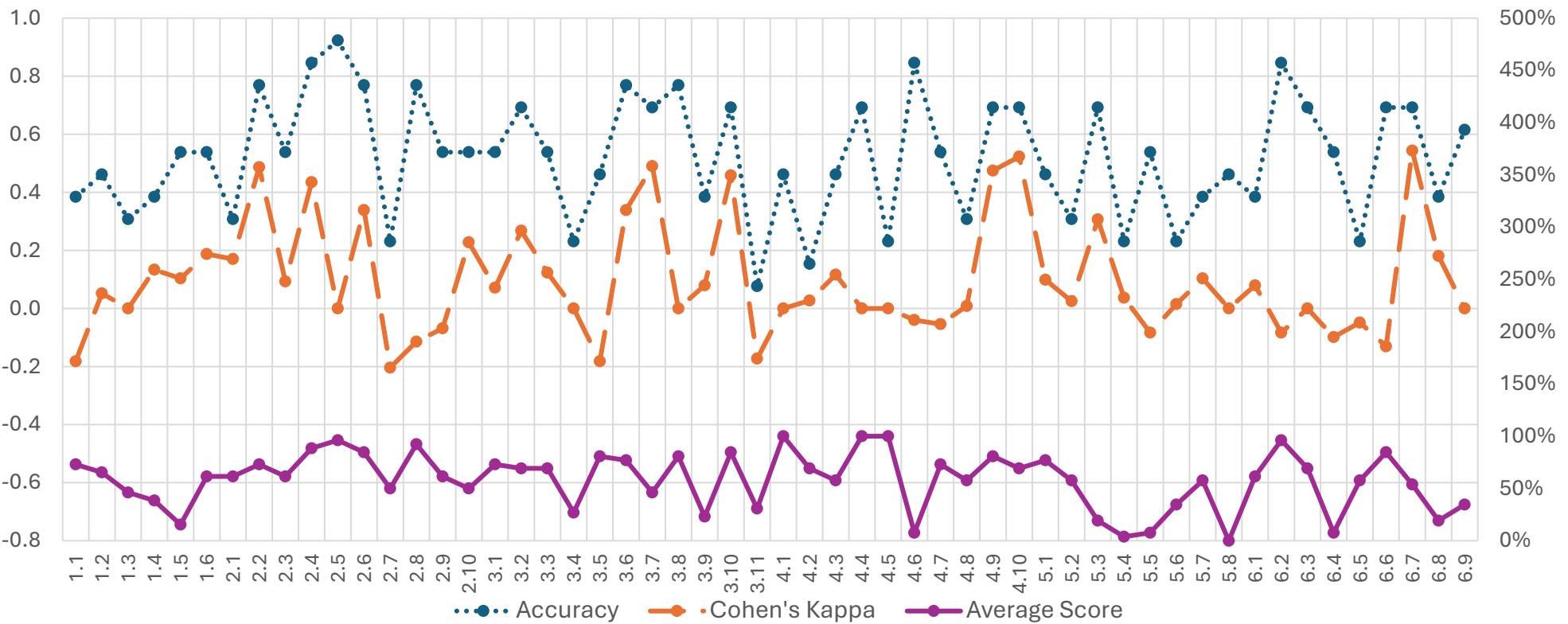
Figure 1 shows the results from a pilot study involving 13 companies (used for student training in Section 8(d)). The figure displays accuracy as a dotted line, Cohen’s kappa as a dashed line, and, on a separate scale, the human scores for each question averaged across the 13 companies. While the two error statistics generally follow the same trend, there are instances where they diverge (e.g., Q2.5 Supplier Policies and Q6.2 Tailored Training). A closer analysis revealed that most responses to Q2.5 and Q6.2 were “1,” and high LLM accuracy. However, these questions also have a very high likelihood of receiving a “1,” so the high accuracy is mainly due to this expected probability. When this factor is

accounted for, Cohen's kappa—which measures agreement beyond chance—shows a lower value.

Figure 1 also shows that Cohen's kappa is notably low for Q1.1 Products and Customers, Q2.7 Policy Dissemination, Q3.5 Supplier Mentoring, and Q3.11 Corrective Actions. A detailed analysis reveals that for these questions, the LLM tends to be more generous in its scoring, searching for relevant information across the entire statement. In contrast, the human scorer focuses more narrowly on specific, targeted areas of the statement. As a result, the LLM's scores are often 0.5 points higher than the human scores. In these cases, it may be appropriate to consider increasing the human scores if a broader interpretation of the statement is justified.

Finally, the human scores show very low results for Q1.5 Names and Countries of Tier 1 Suppliers, Q4.6 Consultation with Rights Holders, Q6.4 Encouraging Supplier Training, and the entire section on Effectiveness. In these cases, although accuracy is high, Cohen's kappa is low because the answers are consistently low. This indicates that the expected value of these scores is low, with a high probability of receiving low scores, which explains the lower Cohen's kappa.

Figure 1. LLM Accuracy and Cohen Kappa: Human Scores as Benchmark



7. Addressing Strategic Ambiguities

Stevenson and Cole (2018) analysed 101 MSS from various industries, including retail, manufacturing, agriculture, and construction, covering reporting years from 2015 to 2017. They identified significant shortcomings in the detection practices of modern slavery within these supply chains, including limited scope and frequency of audits, inconsistent inspections, and a lack of direct worker engagement, which can obscure deeper issues of exploitation. Enhancing audit practices, incorporating worker feedback, and integrating advanced technologies can significantly improve the detection of modern slavery. The study also found that companies often implement vague and ineffective remediation plans, fail to address the root causes of exploitation, and provide insufficient support for victims. Developing clear and actionable remediation plans, focusing on long-term solutions, and offering comprehensive support systems for victims can lead to more effective and sustainable remediation efforts. For disclosure practices, Stevenson and Cole identify issues such as minimal transparency, inconsistent reporting standards, and an overemphasis on positive aspects while downplaying ongoing challenges. Adopting standardised reporting frameworks, encouraging detailed and honest reporting, and engaging independent third-party verification can enhance transparency and credibility, providing a more accurate picture of companies' efforts to combat modern slavery.

More recently, Meehan and Pinnington (2021) analysed 150 Modern Slavery Statements from various industries, including retail, manufacturing, and logistics, covering reporting years 2016 to 2019. They found that companies use strategic ambiguity to obscure accountability, shift responsibility, and present an illusion of compliance, which undermines genuine efforts to address modern slavery in supply chains. These ambiguities often manifest as vague language, defensive reassurances, transferring responsibility, and scope reduction, ultimately undermining the effectiveness of corporate accountability. Addressing these ambiguities is essential to ensure that companies are held accountable for their actions and commitments.

Strategic ambiguities in Modern Slavery Statements pose a significant challenge to genuine efforts in combating modern slavery. This section outlines the process of identifying, categorising, and enhancing the detection of strategic ambiguities using LLMs.

Strategic ambiguities in MSS present a significant challenge to genuine efforts in combating modern slavery. These ambiguities often manifest as vague language, defensive reassurances, transferring responsibility, and scope reduction, ultimately undermining the effectiveness of corporate accountability. Addressing these ambiguities is crucial to ensure that companies are held accountable for their actions and commitments. This section outlines the process of identifying, categorising, and enhancing the detection of strategic ambiguities using LLMs, with insights from Zhang et al. (2023) on the efficacy of active learning in improving text classification.

Zhang et al. (2023) find that active learning significantly improves large-scale text classification by selectively annotating the most informative samples. Their comparative study demonstrates that active learning strategies outperform traditional methods in efficiency and accuracy, enhancing performance in large-scale data environments. Applying this approach to the detection of strategic ambiguities in MSS can significantly enhance the ability of LLMs to identify and categorise vague and misleading statements.

a) Creating Synthetic Examples of Strategic Ambiguities

To train LLMs effectively, it is useful to create synthetic examples of the three main types of strategic ambiguities: defensive reassurance, transferring responsibility, and scope reduction.

| Type | Example | Explanation |
|------------------------------------|---|--|
| Defensive Reassurance | We are committed to maintaining ethical standards. | This statement is vague and does not specify any actions taken or planned. |
| | Our company prioritises human rights in all operations. | A general statement without concrete examples or evidence of prioritisation. |
| | We have a zero-tolerance policy for modern slavery. | Lacks details on how this policy is enforced or monitored. |
| Transferring Responsibility | We rely on our suppliers to enforce anti-slavery practices. | Shifts the enforcement burden to suppliers without explaining how the company verifies compliance. |
| | Our third-party auditors assess our supply chains for modern slavery risks. | Responsibility is given to auditors, without detailing how the company follows up on their findings. |
| | We expect all our business partners to adhere to our ethical guidelines. | Expects partners to comply but doesn't describe monitoring or enforcement mechanisms. |
| Scope Reduction | We focus our anti-slavery efforts on high-risk countries. | Limits efforts to specific countries, ignoring potential risks elsewhere. |
| | Our primary concern is modern slavery in the manufacturing sector. | Focuses only on manufacturing, excluding other sectors. |

| Type | Example | Explanation |
|------|---|---|
| | We ensure compliance with anti-slavery laws in our direct operations. | Only addresses direct operations, not the wider supply chain. |

b) Identifying Ambiguities

The first step in addressing strategic ambiguities is to train LLMs to detect them within MSS. This involves using targeted prompts that guide the LLMs to identify vague or misleading language. For instance, a prompt might be: "Identify sentences like 'We are committed to maintaining ethical standards' that do not specify actions taken or planned." By using such specific prompts, LLMs can learn to recognise patterns and phrases that indicate strategic ambiguity.

Training the LLMs involves exposing them to numerous examples of MSS with known instances of strategic ambiguity. This allows the models to learn and generalise from these patterns, improving their ability to detect similar ambiguities in new texts.

c) Categorising Ambiguities

Once LLMs are guided to identify strategic ambiguities, the next step is to categorise them. This involves using both synthetic cases and real-world examples to create a comprehensive categorisation system for ambiguous language. By evaluating how well the LLM can discern between genuine efforts and superficial compliance, researchers can refine their understanding of different types of ambiguities.

For example, defensive reassurance might be categorised by phrases like "we strive to" or "we are committed to" without specifying actions. Transferring responsibility could be indicated by statements like "we expect our suppliers to" or "our partners are responsible for." Scope reduction might involve narrowing the focus to certain aspects of the supply chain while ignoring others, indicated by language such as "in our direct operations" or "within our first-tier suppliers."

By systematically categorising these ambiguities, LLMs can be prompted to not only detect vague language but also classify it according to the specific tactic being employed. This categorisation aids in understanding the extent and nature of strategic ambiguities across different MSS.

d) Enhancing Detection

The detection and categorisation of strategic ambiguities are iterative processes that require continuous refinement. Enhancing the LLM's ability to detect and categorise ambiguities involves regular updates to the training data and prompt design based on feedback from human annotators. Human annotators play a crucial role in verifying the

outputs of the LLM and providing insights into areas where the model may misinterpret or overlook ambiguities.

This process includes several practical steps:

- Regularly updating the dataset with new examples, including synthetically created or human-labelled MSS that capture diverse instances of strategic ambiguity.
- Iterating on prompt design to ensure that the LLM receives clear and precise instructions for detecting and categorising ambiguities.
- Involving human annotators to review the LLM's outputs, identify errors, and provide feedback to refine the prompts and improve the model's performance.

For instance, if the LLM consistently misses certain types of ambiguous language, human reviewers can highlight these instances and adjust the training prompts accordingly. This iterative feedback loop helps in gradually improving the model's accuracy and reliability in detecting and categorising strategic ambiguities.

By systematically identifying, categorising, and refining the detection of strategic ambiguities, organisations can better ensure that their MSS are genuine reflections of their efforts to combat modern slavery. This approach not only enhances corporate accountability but also contributes to the broader goal of eradicating modern slavery from supply chains.

8. Comparative Analysis and Benchmarking

Comparative analysis and benchmarking are essential components in evaluating the effectiveness and reliability of LLMs in assessing MSS. By comparing LLM-generated scores with human evaluations, ensuring consistency in responses, and conducting longitudinal analysis, organisations can enhance their understanding of corporate disclosures and identify areas for improvement.

a) Benchmarking Against Human Scores

The first step in comparative analysis involves benchmarking the scores generated by LLMs against those provided by human evaluators using the BHRRC framework. This process allows for a direct comparison of the LLM's performance with that of human experts. The BHRRC framework, which includes a comprehensive set of criteria for assessing MSS, serves as a robust baseline for this evaluation.

To conduct this benchmarking, a sample of MSS is scored by both human evaluators and the LLM. These scores are then compared to identify areas where the LLM performs well and areas needing improvement. For instance, if the LLM consistently aligns with human scores in identifying well-documented policies but struggles with detecting vague

language or strategic ambiguities, this discrepancy highlights specific areas for further refinement (LeBaron & Lister, 2021).

The comparison can be facilitated through statistical analysis, examining measures such as correlation coefficients and mean squared error between LLM and human scores. This quantitative approach provides an objective assessment of the LLM's accuracy and highlights the areas where it excels or falls short.

b) Consistency and Improvement

Ensuring consistency in LLM responses across different MSS and reporting years is crucial for reliable benchmarking. Inconsistent responses can undermine the credibility of the LLM's assessments and hinder its utility in longitudinal studies. To address this, feedback loops are employed to iteratively improve the LLM's scoring accuracy and reliability.

Feedback loops involve a continuous cycle of evaluation, feedback, and refinement. After initial benchmarking, human evaluators provide detailed feedback on the LLM's performance, identifying specific instances where the LLM's assessments diverged from human judgments. This feedback is used to refine the LLM's algorithms and retrain the model, improving its ability to accurately score MSS (Wamba et al., 2017).

In addition to feedback from human evaluators, automated validation checks can be implemented to ensure consistency. For example, consistency checks can be performed to verify that the LLM's responses to similar prompts remain stable over time and across different documents. Any identified inconsistencies can be addressed through further model training and parameter adjustments.

c) Longitudinal Analysis

Longitudinal analysis involves monitoring changes in corporate disclosures over time to detect any improvements or regressions. By using LLMs to analyse MSS across multiple reporting periods, organisations can gain insights into the effectiveness of their anti-slavery policies and practices (Ivanov et al., 2019).

LLMs can generate comprehensive summaries and risk assessments based on longitudinal data, highlighting trends and patterns in corporate disclosures. For instance, an LLM can track the evolution of a company's policies on modern slavery, noting any enhancements in detail and transparency or identifying areas where the company's statements have become more ambiguous or less comprehensive.

This longitudinal approach not only provides a dynamic view of corporate performance but also helps in identifying best practices and areas for improvement. Companies can benchmark their progress against industry standards and peers, fostering a culture of continuous improvement.

d) Case Study Examples

The writing of this chapter aligned with the summer dissertation period for MSc students at the University of Manchester. To introduce them to the application of these advanced concepts, a 4-day bootcamp was organised. Following this, a group of five students applied these concepts in their dissertations, each conducting a case study involving 3-5 companies across various sectors, including Clothing & Textile, Supermarkets, Hotels, Beverage & Tobacco, and Food. Initially, the LLM's scores for the companies' MSS were benchmarked against human evaluations. The comparison revealed that while the LLM accurately captured detailed policy descriptions, it also uncovered subtle instances of strategic ambiguity. Feedback from human evaluators was then used to refine the LLM prompts, resulting in more concise output and enhanced detection of vague language in subsequent assessments. Additionally, the students employed longitudinal LLM analysis to identify trends in reporting standards.

By integrating comparative analysis and benchmarking, organisations can harness the power of LLMs to significantly enhance the assessment of MSS. This approach not only guarantees consistent and reliable evaluations but also drives continuous improvement in corporate transparency and accountability. The longitudinal analysis provides crucial insights into the effectiveness of anti-slavery initiatives over time, reinforcing the broader mission to eliminate modern slavery from global supply chains. Moreover, the students involved in these pilot studies gain valuable and transferable AI skills that will serve them well in their future careers. As ESG considerations become increasingly central to investment decisions and climate change regulations, LLM-driven analytical tools will play a critical role in advancing the industry towards more sustainable business practices.

9. Discussion and Conclusion

The UK Modern Slavery Act 2015 represents a significant step forward in addressing modern slavery and human trafficking within business operations and supply chains. However, its effectiveness has been a subject of debate. LeBaron and Rühmkorf (2019) provide an insightful analysis into the political dynamics that shaped the Act, highlighting how corporate lobbying and domestic politics led to a focus on voluntary corporate disclosures over mandatory regulations. This emphasis on reporting, without stringent penalties for non-compliance, has raised concerns about the actual impact of the Act on eradicating modern slavery.

One major criticism is that the Act primarily targets larger companies, potentially leaving smaller businesses and more hidden parts of supply chains unregulated and thus more vulnerable to exploitation. This loophole means that while larger corporations may improve their practices, smaller entities might continue to operate under the radar, perpetuating modern slavery conditions. Additionally, businesses are afforded significant leeway in interpreting and applying the requirements of the Act, which can

dilute its overall effectiveness. This flexibility can result in superficial compliance, where companies produce detailed reports that do not necessarily translate into meaningful action against modern slavery.

Crane (2013) further elaborates on the organisational conditions that facilitate human exploitation. These include complex and opaque supply chains, high pressure for cost reductions, weak enforcement of labour standards, and a profit-driven corporate culture that often prioritises financial performance over ethical considerations. Such an environment is ripe for exploitative practices to thrive unnoticed and unchallenged. Addressing these issues requires a comprehensive approach that enhances supply chain transparency, strengthens labour standards, promotes accountability, and fosters ethical leadership. Crane proposes a framework involving supply chain mapping, rigorous audits, robust reporting mechanisms, public disclosures, and empowering workers through representation and support.

The implementation of LLMs in analysing MSS offers a promising solution to some of these challenges identified by LeBaron and Rühmkorf (2019) and Crane (2013). LLMs can process and analyse extensive textual data quickly and consistently, identifying patterns and trends that may be missed in manual reviews. This automation can enhance the efficiency and scalability of MSS analysis, ensuring a thorough evaluation of a larger number of statements. By providing detailed insights into the content of MSS, LLMs can help highlight areas of compliance and non-compliance, generating summaries and recommendations that can inform policy decisions and drive improvements in corporate practices.

Specifically, LLMs can address some of the issues identified by LeBaron and Rühmkorf. For instance, the flexibility in interpreting and applying the Act's requirements could be mitigated by LLMs' ability to standardise the evaluation of MSS. By applying consistent criteria across all statements, LLMs reduce the risk of superficial compliance and ensure a more uniform assessment of corporate efforts against modern slavery. This could indirectly pressure smaller companies to enhance their transparency and compliance, even if they are not explicitly targeted by the Act.

Regarding the conditions identified by Crane (2013), LLMs can significantly contribute to enhancing supply chain transparency and accountability. By automating the analysis of supply chain disclosures, LLMs can identify gaps and inconsistencies that may indicate exploitative practices. They can also highlight best practices and effective policies, promoting a culture of transparency and accountability. However, while LLMs can enhance the detection and reporting of such practices, they are not a panacea. Effective implementation still requires strong regulatory frameworks and active enforcement to ensure that the insights generated by LLMs lead to tangible improvements in corporate behaviour.

However, the use of LLMs is not without its challenges. Ensuring the accuracy and reliability of LLM-generated insights requires continuous refinement and validation. Comparative analysis and benchmarking against human evaluations are essential to

identify areas where LLMs perform well and where they need improvement. For instance, while LLMs might excel in identifying well-documented policies, they may struggle with detecting vague language or strategic ambiguities often used to obscure non-compliance. Feedback loops involving human reviewers are crucial for refining LLM algorithms and enhancing their ability to detect and categorise such ambiguities.

Addressing strategic ambiguities in MSS is particularly important for ensuring genuine corporate accountability. As Stevenson and Cole (2018) and Meehan and Pinnington (2021) suggest, companies often use vague language, defensive reassurances, and other tactics to obscure accountability and present an illusion of compliance. Using LLMs to identify and categorise these ambiguities can significantly enhance the detection of misleading statements, ensuring that companies are held accountable for their commitments. This involves creating synthetic examples of common strategic ambiguities and using targeted prompts to guide LLMs in identifying vague or misleading language. Continuous refinement of these prompts, based on feedback from human annotators, can improve the LLMs' accuracy and reliability in detecting strategic ambiguities.

In conclusion, while the UK Modern Slavery Act 2015 has laid a foundation for addressing modern slavery in business operations, its effectiveness is limited by the focus on voluntary disclosures and the lack of stringent enforcement mechanisms. The integration of LLMs in the analysis of MSS presents a powerful tool for enhancing compliance assessments, provided that their use is complemented by continuous refinement and validation processes. By addressing the challenges of strategic ambiguity and leveraging the advanced capabilities of LLMs, researchers and policymakers can drive significant improvements in corporate transparency and accountability, contributing to the global effort to combat modern slavery.

References:

- Bales, K. (2007). Ending Slavery: How We Free Today's Slaves. University of California Press.
- Bales, K., & Soodalter, R. (2009). *The Slave Next Door: Human Trafficking and Slavery in America Today*. University of California Press.
- Bommarito, M. J., & Katz, D. M. (2020). GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About. *MIT Technology Review*.
- Brown, T. et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Business & Human Rights Resource Centre (BHRRC). (2023). Modern Slavery Registry. Available at: Modern Slavery Registry - <https://www.modernslaveryregistry.org/>
- Crane, A. (2013). Modern Slavery as a Management Practice: Exploring the Conditions and Capabilities for Human Exploitation. *Academy of Management Review*, 38(1), 49-69.
- Crane, A., LeBaron, G., Phung, K., Behbahani, L., & Allain, J. (2017). Governance Gaps in Eradicating Forced Labor: From Global to Domestic Supply Chains. *Regulation & Governance*, 13(1), 86-106.
- Devlin, J., et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805.
- Gold, S., Trautrimas, A., & Trodd, Z. (2015). Modern Slavery Challenges to Supply Chain Management. *Supply Chain Management: An International Journal*, 20(5), 485-494.
- Howard, J., and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328-339.
- Gao, J., et al. (2023). "Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation." arXiv preprint arXiv:2305.07375.
- Kara, Siddharth. Modern slavery: A global perspective. Columbia University Press, 2017.
- Ivanov, D., Dolgui, A., & Sokolov, B. (2019). "The impact of digital technology and industry 4.0 on the ripple effect and supply chain risk analytics." *International Journal of Production Research*, 57(3), 829-846.
- Jackson, B., Lo, C., Murray, J., Villasenor, J., & Tambe, M. (2020). Efficient Deployment of Privacy-Preserving AI for Public Good. *Communications of the ACM*, 63(9), 53-56.
- LeBaron, G., & Lister, J. (2021). "Ethical audits and the supply chains of global corporations." *Review of International Political Economy*, 28(3), 719-745.

- LeBaron, G., and Rühmkorf, A. (2017). Steering CSR Through Home State Regulation: A Comparison of the Impact of the UK Bribery Act and Modern Slavery Act on Global Supply Chain Governance. *Global Policy*, 8(S3), 15-28
- LeBaron, G., and Rühmkorf, A. (2019). "The domestic politics of corporate accountability legislation: Struggles over the 2015 UK Modern Slavery Act." *Socio-Economic Review*.
- Marcus, Gary and Ernest Davis (2020). GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About. *MIT Technology Review*. Last accessed: 21/08/2024 <https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>
- Meehan, J., & Pinnington, B. D. (2021). "Modern slavery in supply chains: insights through strategic ambiguity." *International Journal of Operations & Production Management*, 41(2), 77-101.
- New, S. J. (2015). Modern Slavery and the Supply Chain: The Limits of Corporate Social Responsibility? *Supply Chain Management: An International Journal*, 20(6), 697–707. DOI: 10.1108/SCM-06-2015-0201
- Pinnington, B., & Meehan, J. (2023). "Learning to see modern slavery in supply chains through paradoxical sensemaking." *Journal of Supply Chain Management*, 59(4), 22-41.
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. (2019) "Language models are unsupervised multitask learners." OpenAI blog 1, no. 8: 9.
- Rasche, A. (2010). "The Limits of Corporate Responsibility Standards." *Business Ethics: A European Review*, 19(3), 280–291. DOI: 10.1111/j.1467-8608.2010.01592.x.
- Skrivankova, Klara. (2010) "Between decent work and forced labour: examining the continuum of exploitation." York: Joseph Rowntree Foundation, #16.
- UK Home Office. (2015). Transparency in supply chains etc. A practical guide. accessed 21/08/2024, https://assets.publishing.service.gov.uk/media/61b7401d8fa8f5037778c389/Tra_nsparency_in_Supply_Chains_A_Practical_Guide_2017_final.pdf.
- Vaswani, A., et al. (2017). "Attention Is All You Need." *Advances in Neural Information Processing Systems*, 30, 5998-6008.
- Wamba, S. F., et al. (2017). "Big data analytics and firm performance: Effects of dynamic capabilities." *Journal of Business Research*, 70, 356-365.
- Wang, X., et al. (2023). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." arXiv preprint arXiv:2201.11903.

Zhang, H., et al. (2023). "Active Learning for Large-Scale Text Classification: A Comparative Study." Proceedings of the AAAI Conference on Artificial Intelligence, 37(1), 123-130.