Scientific Research Publishing

Open Journal of Social Sciences, 2025, 13(10), 1-17
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Detecting Bias in AI: A Multi-Label RoBERTa Classification Model to Detect Bias in LLM-Generated Diversity Reports

Mathew Sunil Abraham, Nicole Lee, Rezza Moieni

Diversity Atlas, Melbourne, Australia
Email: msabraham98@gmail.com, nicole.lee@diversityatlas.io, rezza.moieni@culturalinfusion.com

How to cite this paper: Abraham, M. S., Lee, N., & Moieni, R. (2025). Detecting Bias in AI: A Multi-Label RoBERTa Classification Model to Detect Bias in LLM-Generated Diversity Reports. *Open Journal of Social Sciences, 13,* 1-17.
https://doi.org/10.4236/jss.2025.1310001

Received: June 30, 2025
Accepted: September 26, 2025
Published: September 29, 2025

Abstract

With increasing use of generative AI in creating textual summaries and dashboard reports, there is significant concern regarding diverse forms of bias such as gender, geographical, religious, cultural, and language biases. This research investigates the presence of bias in AI-generated diversity reports and presents a sentence-level bias detection model to quantify and classify different types of bias. The study focuses on five key bias categories: gender, religion, age, disability, and sexuality. We train and evaluate a multi-label bias classifier with a RoBERTa-based deep learning model, integrating manual confidence-weighted annotation practices to ensure reliable labelling. Synthetic diversity reports were generated using the Gemini 1.5-flash language model to simulate real-world corporate content. We use this model to analyse around 1000 reports (10,000+ sentences) for bias and assess the nature and distribution of different bias types. The model demonstrated high accuracy and recall rates, effectively detecting both overt and subtle biases across categories. Analysis of over 10,000 sentences revealed measurable bias in the generated reports, with disability, gender, and religion biases being the most frequently detected. These findings highlight that even when using inclusive prompts, large language models can produce biased content. The model's strong performance and ability to detect nuanced and intersectional biases make it a practical tool for organisations aiming to audit AI-generated communications. We expect this to contribute to the growing field of ethical AI by supporting more transparent, fair, and inclusive corporate reporting.

Keywords

Artificial Intelligence (AI), Ethnical AI, Bias Detection, Large Language Models (LLMs), Diversity Data Analysis

1. Introduction

The emergence of generative Artificial Intelligence (AI) tools has transformed organisational communication, particularly in creating diversity and inclusion reports. Large Language Models (LLMs) like GPT-4, GPT-4o, and Gemini are increasingly utilised to automate content for internal and external distribution. However, this reliance on AI introduces challenges related to fairness and representation, as these models may unintentionally perpetuate societal biases present in their training data (Bender et al., 2021). Such biases can appear across dimensions like gender, religion, age, disability, and sexuality, affecting how various groups are portrayed in organisational narratives. To demonstrate, consider this hypothetical sentence generated for a diversity report: "Our senior leadership team, headed by capable male executives, collaborates with supportive female team members to help our dynamic, youthful workforce navigate modern workplace challenges while addressing the needs of employees with disabilities." This one sentence reveals several subtle biases: it links leadership with men and support roles with women (gender bias), suggests that only younger employees possess energy (age bias), and presents disability accommodations as "special" instead of being regarded as standard practice (disability bias). These types of seemingly neutral statements can subtly reinforce harmful stereotypes while appearing professional and inclusive on the surface.

This is particularly problematic for diversity reports, which aim to foster inclusivity and fairness. If AI-generated content contains biases, it can contradict the intended goals of these reports. Addressing and reducing these biases is essential to uphold credibility, advance equity, and ensure alignment with organisational values.

Previous studies conducted within the organisation have also addressed these concerns. Raichur et al. (2023) demonstrated the persistent gender biases in government communication despite policy-driven inclusivity efforts, while Mao et al. (2023) developed a large-scale language model specifically to detect and mitigate gender and age biases in Australian job advertisements. These foundational studies inform and support the direction of the current project.

The main aim of this study is to detect and quantify bias in AI-generated reports. The objectives were to:

- Build a multi-label classifier capable of detecting five types of bias: gender, religion, age, disability and sexuality.
- Generate synthetic diversity reports using multiple LLMs to simulate real-world corporate content.
- Create a manually annotated dataset using majority vote and confidence weighting strategies to ensure high-quality training data.
- Train and evaluate a RoBERTa-based deep learning model for multi-label bias classification.
- Use this model to analyse around 1000 reports (10,000+ sentences) for bias and

assess the nature and distribution of different bias types.

In the context of this study, *bias* refers to systematic and unfair favouritism or marginalisation of particular groups within AI-generated content. The following categories of bias were specifically assessed:

- **Gender Bias:** Assumptions, stereotypes, or language that reinforce traditional gender roles or exclude gender-diverse identities.
- **Religion Bias:** Language that favours majority religions, overlooks minority faiths, or frames religious diversity as problematic.
- **Age Bias:** Stereotypes that associate age with declining abilities or assign capabilities based on generational labels.
- **Disability Bias:** Language that frames disability accommodations as special or unusual, reinforcing exclusion rather than normalisation.
- **Sexuality Bias:** Assumptions or language that prioritise heterosexual identities while overlooking or excluding non-heterosexual individuals.

This project holds significant importance for several reasons. First, it introduces a structured framework for detecting bias at the sentence level within long-form AI-generated text. Second, it integrates human-centred annotation practices with modern deep learning techniques, making the resulting model both robust and grounded in real-world perceptions. Third, it provides actionable insights for organisations seeking to audit and enhance the inclusivity of their automated communication systems. Finally, by focusing on corporate diversity reporting, it addresses a gap in current NLP research, which has primarily concentrated on bias in social media and short-form text.

2. Literature Review

Artificial Intelligence (AI), particularly Large Language Models (LLMs) such as GPT, has revolutionised how content is generated, summarised, and presented across industries. However, growing reliance on these models has also raised ethical concerns regarding the biases they inherit and propagate. These biases can influence representation, reinforce stereotypes, and affect societal equity (Shuford, 2024; Marinucci et al., 2023). Recent studies further reinforce this concern by demonstrating how gender and age biases are deeply embedded in both public communications and recruitment materials, even when organisations intend to promote inclusivity (Raichur et al., 2023; Mao et al., 2023). As institutions increasingly use AI to automate content, especially public-facing reports, identifying and mitigating biases in AI-generated text becomes essential to ensure fairness, inclusivity, and trustworthiness.

This literature review explores academic and applied research related to bias detection in AI-generated language and discusses the classification of multi-dimensional biases (e.g., gender, religion or worldview, age, disability, sexuality). It draws on studies from Machine Learning (ML), Natural Language Processing (NLP), media theory, and ethical AI governance to build a foundation for bias quantification in institutional reporting.

2.1. Defining and Understanding Bias in Text

Bias in language models is typically defined as a systematic and unfair skew in output toward or against certain groups, perspectives, or regions. Mehrabi et al. (2021) describes bias as multi-dimensional and context-dependent, often arising from biased training datasets or societal power structures. Gender bias, for instance, is observed when leadership roles are associated with masculine terms (Zhao et al., 2018), while religious bias may reinforce dominant worldviews or stereotypes (Abrar et al., 2025). Geographic bias is often observed when Global North regions (e.g., the United States, the United Kingdom, Australia) are overrepresented in training data or outputs (Li et al., 2024). These biases are not mutually exclusive and can intersect, compounding marginalisation. Zhao et al. (2018) demonstrated how word embeddings trained on common text sources reflect gender stereotypes, such as associating "man" with "programmer" and "woman" with "homemaker". Abrar et al. (2025) explored religious narratives within NLP, identifying skewed representation of Abrahamic versus Eastern religions. Li et al. (2024) found that GPT-based summaries favoured data and references from Global North countries, limiting global contextual awareness. These studies underline the necessity for deliberate design and evaluation to ensure equitable outputs across cultural and demographic lines.

The theoretical basis for understanding bias in AI-generated organisational reports can be traced to principles of strategic corporate communication. Cornelissen (2023) highlights that organisational communication extends beyond information delivery, focusing instead on shaping perceptions, building meaning, and guiding interpretation through framing techniques. In the case of AI-produced content, such as diversity reports or HR summaries, the manner of presentation can greatly affect stakeholder comprehension, trust, and involvement. Sentence-level framing, tone, and emphasis might inadvertently perpetuate existing social hierarchies or overlook marginalised viewpoints, emphasising the need for a thorough assessment of representational fairness in automated reporting.

2.2. Role of AI in DEI Initiatives

As Diversity, Equity, and Inclusion (DEI) becomes a strategic priority across the tech sector, organisations are increasingly relying on AI tools to generate structured reports that highlight progress and gaps in workplace equity. AI-generated reports enable companies to streamline internal documentation, reduce human error, and scale communication efforts without expanding operational costs. These tools are especially useful for creating regular performance updates, recruitment analytics, inclusive policy summaries, and compliance documentation.

According to Kondra et al. (2025), AI systems are being integrated across HR and DEI functions for tasks such as monitoring demographic representation, evaluating inclusive hiring practices, and automating the creation of diversity dashboards. This automation helps reduce bias in human decision-making but may also embed new forms of algorithmic bias, making transparency and interpreta-

bility in AI outputs essential.

For example, large tech firms like Microsoft, IBM, and Accenture have reported using generative AI to assist in producing inclusion reports, employee experience summaries, and ethical compliance updates. Thus, this project's focus on AI-generated DEI reports reflects a real-world trend: the growing dependence on language models for inclusive corporate communication and the corresponding need to ensure those outputs are equitable and unbiased.

2.3. Bias Detection in AI

A variety of techniques have been employed to detect bias within AI models and their outputs. These approaches utilise diverse methodologies, from relying on lexicon-based methods to employing sophisticated embedding-based models for comprehensive detection. Lexicon-based methods use pre-compiled lists of terms to identify bias but lack contextual sensitivity (Fast et al., 2016), whereas embedding-based models like BERT and RoBERTa are more effective in detecting subtle bias through semantic analysis (Jentzsch & Turan, 2022). This aligns with previous work where RoBERTa outperformed other models in bias detection tasks within Australian job advertisements, specifically for gender and age bias detection (Mao et al., 2023). Supervised models trained on datasets like StereoSet (Nadeem et al., 2020) enable sentence-level detection and classification of multiple bias types. Fairness evaluation metrics like Equality of Odds and Disparate Impact further assess model performance in bias-sensitive tasks (Dixon et al., 2018). Sentence-level analysis is crucial, as it provides granular insight into how specific lines within a larger report may carry bias, enabling better downstream filtering and correction.

Building on these detection techniques, scoring and classification models then provide a structured way to quantify identified biases. Supervised ML classifiers such as logistic regression, support vector machine, or fine-tuned transformers can assign binary labels (biased/unbiased) or bias type codes (e.g., 1G = gender bias). Furthermore, probabilistic scoring allows assigning a confidence-based score between 0 and 1 to indicate the likelihood of bias, which can support nuanced mitigation strategies (Quan & Pu, 2023; Hossin & Sulaiman, 2015).

2.4. Broader Impacts and Policy Implications

Bias in AI-generated reports can influence decision-making, amplify inequality, and reduce user trust. This is particularly concerning in education, hiring, and journalism. Governments and institutions are thus calling for transparency in AI development. For instance, Amazon's recruitment AI system was scrapped after it was found to downgrade resumes containing the word "women's", reflecting gender bias in the training data. Another well-known example includes Twitter's image-cropping algorithm, which was found to disproportionately favour lighter-skinned faces over darker ones in preview images, raising concerns about racial bias in visual AI systems. R'boul (2021) cautions that intercultural messaging,

even when promoting inclusion, may reinforce existing hierarchies. Ethical AI frameworks recommend continuous auditing, inclusive dataset creation, and human-in-the-loop systems to reduce unintended consequences. Bias detection models like the one developed in this study contribute directly to mitigation efforts by identifying harmful patterns early, enabling organisations to retrain models, refine prompts, and prevent the reinforcement of systemic inequalities.

2.5. Gaps and Future Directions

Despite significant progress, current models struggle with intersectional bias, where multiple identity-based biases compound within AI outputs affecting individuals who belong to multiple marginalised groups simultaneously (Mehrabi et al., 2021). Additionally, most bias benchmarks are based on Western language corpora (Li et al., 2024). Our project addresses these gaps by developing a multi-label bias detection model trained on a manually annotated, sentence-level dataset that covers five distinct bias categories: gender, religion, age, disability, and sexuality. By generating 990 AI-generated diversity reports using Gemini LLM, we introduced diversity in model behaviour and content sources. Furthermore, our dataset includes annotations from multiple human evaluators using a majority vote system to ensure reliability and reduce subjectivity. This approach directly confronts the issue of intersectional and cultural bias by applying consistent, multi-annotator labelling across diverse topics and prompts. In doing so, the project moves toward building a scalable, adaptable framework for detecting representational bias in real-world organisational content. Future work should include multilingual datasets, domain-specific audits, and real-time evaluation tools for AI-generated content in live systems. Integrating human annotations with automated predictions could further enhance accuracy and trustworthiness.

3. Methodology

This section outlines the methodology employed in this study, encompassing dataset preparation, manual annotation, and model development. We explain the generation and refinement of the synthetic diversity reports, the systematic process of human annotation for bias identification, and the subsequent training of our model. Figure 1 offers a comprehensive overview of these stages.

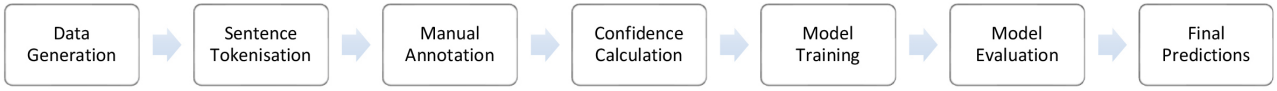

Figure 1. Methodology workflow outlining the key stages from dataset generation to final bias prediction.

3.1. Dataset Preparation

A total of 990 synthetic diversity reports were generated using prompts applied to three prominent large language models: GPT-4, GPT-4o, and Gemini 1.5-flash. Each report was based on one of nine standardised diversity-related prompts, covering topics such as multilingual workplace support, inclusive religious practices,

accommodations for long-term conditions, generational diversity, data-driven diversity and inclusion strategies, planning, social media content, and internal communications to HR or executive leadership. Examples of these prompts include:

- Multilingual_Workplace_Support: "*Please provide a 'Multilingual Workplace Support' plan that can be implemented in the short term. The goal is to establish a supportive and effective multilingual environment by addressing communication challenges, preventing conflicts, and ensuring seamless interactions with customers. Please pay particular attention to communication issues, provide advice, and suggest solutions to avoid challenges with employees. Focus more on language data and the percentage of multilingual employees. Additionally, propose a solution to prevent potential conflicts among multilingual employees. Be specific and use some of the data provided.*"
- Long_Term_Condition_Accommodations: "*Please provide a 'Long-Term Condition Accommodations' plan that can be implemented in the short term. Pay attention to the number of participants with long-term conditions, and the percentage of participants with long-term conditions across different generations. Also, consider the most common identified long-term condition and provide relevant suggestions. Propose steps to address potential mental health issues. Suggest ways to prevent employees with long-term conditions from leaving the workforce and worsening their health conditions. Recommend strategies to increase the productivity of these employees. Please be specific and use the data provided.*"

These prompts were distributed across 110 major technology companies to ensure diverse industry representation.

Following an evaluation of the outputs, reports generated by Gemini 1.5-flash were selected for further analysis. This decision was primarily driven by Gemini's consistent ability to complete sentences and maintain logical coherence, an area where other models occasionally produced incomplete or abruptly ended sentences. Particularly, GPT-4 and GPT-4o often produced incomplete sentences, exhibited inconsistent formatting that made segmentation more difficult, and generated repetitive content that could distort bias detection. These models struggled to keep context clear in longer documents, leading to contradictory statements and extraneous annotation. The choice of model may have impacted the bias patterns present in the training data, as various large language models display different bias inclinations based on their respective training sets. In addition, Gemini's outputs exhibited a more uniform structure and clearer language, contributing to more reliable annotation and enhanced performance in downstream model training.

The final dataset consisted of over 10,000 sentences. Reports were segmented into individual sentences using Python-based natural language processing techniques, and relevant metadata such as report ID and prompt type were retained. To minimise cognitive bias during the annotation process, company names were anonymised prior to review.

3.2. Manual Annotation and Label Creation

A total of 300 randomly selected reports, comprising approximately 4700 sentences, were independently annotated by four reviewers. Each sentence was assessed for the presence (labelled as 1) or absence (labelled as 0) of bias across five key dimensions: gender bias, religious bias, age bias, disability bias, and sexuality bias. Reviewers carefully analysed each sentence and categorised it into the appropriate bias classifications according to their evaluation of biased wording, stereotypes, or unjust representation.

To evaluate the reliability of the annotations, inter-annotator agreement was determined for all bias categories prior to implementing the weighted majority voting method. This offered a quantitative assessment of the consistency among the four reviewers in recognising bias within the sentences they manually annotated.

To enhance the reliability of the annotation process, a weighted majority voting system was employed.

- When three or four annotators agreed on the presence of bias, the sentence received a label of 1 with high confidence (confidence value = 1).
- If two annotators agreed, the label was assigned with medium confidence (0.67).
- If only one annotator identified bias, the sentence was labelled with low confidence (0.33).
- In cases where no consensus on bias was reached, the sentence was classified as unbiased (label = 0).

This approach facilitated the generation of both binary bias indicators and confidence-weighted labels, enabling more nuanced model training and evaluation.

3.3. Model Development

The RoBERTa model was chosen for its superior performance in sentence-level classification and proven success in bias detection, attributed to its enhanced pretraining efficiency, robust contextual understanding, and stable training compared to BERT. Its ability to handle large datasets, process long sequences, and support flexible fine-tuning for multi-label tasks made it ideal for identifying nuanced biases across diverse text domains. Additionally, RoBERTa's balance of high performance, computational efficiency, and strong community support in frameworks like Hugging Face ensures reliable and practical implementation.

Training data was split into 80% for training and 20% for validation. Training was performed over 5 epochs using the AdamW optimiser with a learning rate of 2e−5 and batch size of 8. A linear learning rate scheduler was employed, and early stopping was considered to prevent overfitting. Sentences and labels were converted into PyTorch-compatible datasets with appropriate attention masks and input embeddings. Sentences were tokenised using the RoBERTa tokeniser, applying padding and truncation to a maximum length of 256 tokens. This ensured consistent input shapes and preserved key contextual information for robust sen-

tence-level understanding.

The final trained model file size was approximately 487 MB, indicating that the base model was efficiently fine-tuned without excessive parameter overhead. This balance between model capacity and storage size made it practical for sentence-level bias detection tasks.

The initial analysis revealed a low prevalence of bias across most categories, with certain dimensions such as sexuality bias appearing in only approximately 1% of the annotated sentences. To address the resulting class imbalance and enhance model learning, several strategies were employed:

- **Minority class augmentation**: It was applied by oversampling positive examples to ensure better representation of underrepresented bias categories.
- **Focal loss function**: It was utilised to assign greater penalty to the misclassification of these minority classes, thereby improving the model's sensitivity to subtle bias indicators.
- **Confidence weights:** The weights derived from the annotation process were incorporated during training, allowing the model to prioritise learning from examples labelled with higher certainty.

A multi-label classification setup was used, as each sentence could simultaneously exhibit more than one type of bias. This approach allowed the model to independently assess all five bias categories, enhancing detection accuracy for complex and intersectional biases. The bias detection models produced probability scores for each bias label rather than binary outputs. Instead of applying a fixed threshold such as 0.5 across all labels, the study computed optimal thresholds for each category individually by maximising the F1-score, with a particular emphasis on improving recall. This approach enabled the model to better identify subtle forms of bias, reduce the likelihood of false negatives, and enhance the overall accuracy of sentence classification when applied in real-world scenarios.

4. Results and Discussions

This section presents the findings from the application of our RoBERTa-based model, showcasing the distribution of detected bias, model performance metrics, and the practical implications of these findings for organisations.

4.1. Bias Detection Findings

The RoBERTa-based multi-label classification model was applied to a final dataset comprising 10,862 sentences sourced from synthetic diversity reports generated using the Gemini 1.5-flash large language model. The model independently evaluated each sentence for the presence of five distinct types of bias: gender, religion, age, disability, and sexuality.

The predicted distribution of bias revealed that gender bias was detected in 681 sentences (6.27%), religion bias in 675 sentences (6.21%), age bias in 549 sentences (5.05%), disability bias in 829 sentences (7.63%), and sexuality bias in 111 sentences (1.02%). Figure 2 represents these total bias detections across the dataset.

These findings indicate that, despite the use of inclusive prompts during the generation process, the LLM-generated reports contained measurable bias across multiple dimensions.

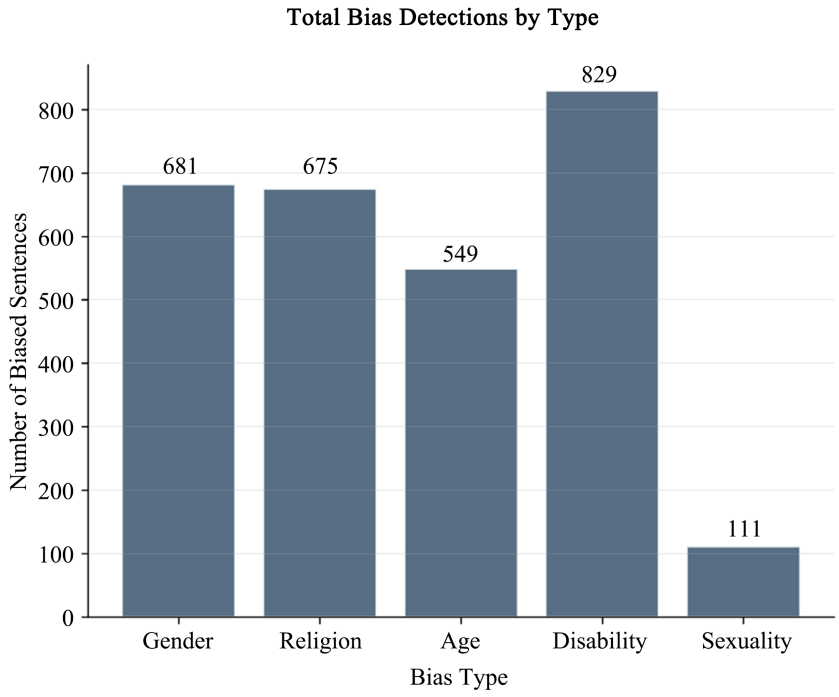

Figure 2. Number of biased sentences detected by bias type.

The analysis revealed that disability bias was the most frequently detected bias, suggesting that even with diversity-focused prompts, the language model consistently generated content that subtly marginalised individuals with disabilities. For instance, the sentence "*This plan addresses the needs of employees with long-term conditions, aiming to improve workplace wellbeing and productivity.*" positions employees with disabilities as requiring special attention rather than being part of the standard organisational fabric, reinforcing the notion of separation rather than inclusion.

Gender and religion biases were also notably prevalent, each appearing in over 6% of the total sentences, indicating persistent representational imbalances in these categories. Gender bias frequently manifested in the form of binary gender assumptions and demographic generalisations. For example, the sentence "*Assume your workforce is 60% female, 40% male*" portrays a binary framing that excludes gender-diverse individuals and oversimplifies gender representation.

Religion bias often surfaced in organisational narratives that framed religious diversity as a potential management issue rather than an inclusive opportunity. The sentence "*This plan aims to address potential issues arising from diverse religious and worldview beliefs within your organisation*" subtly positions diversity as a challenge, which can unintentionally marginalise individuals with diverse religious backgrounds.

Sexuality bias, though less frequently detected, was sometimes reflected in language that indirectly connected sexual orientation with health-related concerns. For instance, the sentence "*The percentage of individuals identifying as non-heterosexual and experiencing long-term health conditions varies across generations, impacting individual needs and preferences.*" subtly implies an association between non-heterosexual identities and health conditions, which risks perpetuating harmful stereotypes.

Furthermore, Figure 3 below illustrates ten companies with the highest total bias detections across the analysed synthetic diversity reports. Okta recorded the highest number of biased sentences (38), followed closely by ASML and GitHub (Microsoft) with 36 and 35 detections respectively. This distribution highlights that even among leading technology companies, measurable bias persists in AI-generated content.

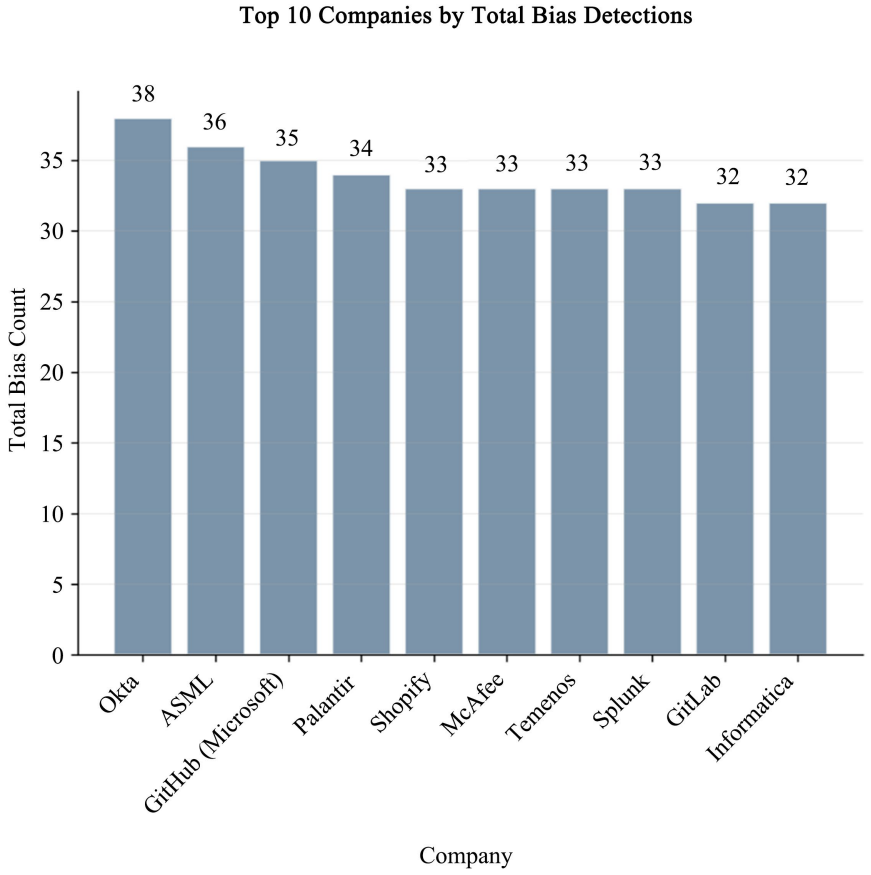

Figure 3. Ten companies with the highest number of biased sentences detected.

4.2. Model Performance Metrics

The model's predictive performance was evaluated on a validation set comprising 20% of the manually annotated dataset. The following metrics—accuracy, precision, recall, and F1-score—were recorded and are presented in Table 1 below. The overall model accuracy across all labels was 0.99.

Table 1. Values of accuracy, precision, recall, and F1-Score by bias type.

Bias Type	Accuracy	Precision	Recall	F1-Score
Gender Bias	0.99	0.81	0.97	0.89
Religion Bias	0.98	0.90	0.82	0.86
Age Bias	0.99	0.93	0.89	0.91
Disability Bias	0.97	0.69	0.97	0.80
Sexuality Bias	1.00	0.86	1.00	0.92

The model exhibited excellent sensitivity, particularly through consistently high recall scores across all bias types. It performed especially well in detecting gender, disability, and sexuality biases:

- **Gender bias:** The detection achieved a robust F1-score of 0.89 and near-perfect recall, indicating the model's strong capability to correctly identify nearly all biased sentences flagged by human annotators.
- **Religion and age bias:** The detection showed balanced performance, maintaining a solid trade-off between precision and recall.
- **Disability bias:** An important trade-off was observed. While the model achieved an outstanding recall of 97%, its precision dropped to 69%. This indicates that although the model was highly sensitive to biased sentences, it also generated some false positives by incorrectly classifying neutral sentences as biased. In bias detection tasks, however, such a trade-off is often acceptable and desirable, as high recall is typically prioritised to minimise the risk of missing biased content. This is especially important in the context of organisational diversity reporting, where unnoticed bias can reinforce damaging stereotypes and undermine institutional credibility, whereas false positives can be corrected through human evaluation (Dixon et al., 2018).
- **Sexuality bias**: The detection achieved perfect recall and very high precision, suggesting it was highly effective in this category. Nonetheless, this result should be interpreted with caution due to the small number of positive examples for sexuality bias in the dataset (36), which may have introduced the risk of overfitting. While the model's performance in this category is promising, further evaluation on more diverse and balanced datasets is recommended to confirm its generalisability.

4.3. Model Strengths and Considerations

A key strength of this model lies in its use of a confidence-weighted annotation process. By applying varying weights based on the level of annotator agreement classified as low, medium, or high confidence, the model was trained to prioritise highly reliable examples while still learning from more ambiguous cases. This approach proved particularly valuable, as it improved the model's ability to detect subtle bias indicators that may not have been consistently identified by all annotators. It also reduced the risk of overfitting to only the clearest examples, ensuring

that the model developed sensitivity to a wide range of bias expressions, including less overt cases. The inclusion of confidence-weighted learning allowed the model to make more nuanced decisions during training, contributing to its excellent generalisation on the validation set. By learning from both highly agreed-upon and less certain examples, the model became particularly robust in handling the complex nature of bias detection.

Throughout this study, the bias detection model demonstrated several significant strengths. One of its most notable advantages was its high sensitivity, with consistently strong recall scores across all bias categories. This ensured that the model successfully detected most biased sentences, making it particularly reliable for screening large volumes of AI-generated content where manual review would be impractical. The model also exhibited strong multi-label flexibility, correctly identifying intersectional bias where a single sentence could display multiple types of bias simultaneously. This capability is crucial for real-world applications, where complex layers of bias often coexist.

Furthermore, the model effectively managed class imbalance within the dataset. The use of focal loss and strategies for minority class augmentation, especially for sexuality bias, allowed the model to learn better from underrepresented categories and improve its sensitivity to less frequently detected biases. Another key strength was the optimisation of decision thresholds for each bias type. Instead of applying a generic cut-off, the model was fine-tuned to select thresholds that maximised the F1-scores for each specific category, thereby enhancing its overall accuracy and practical performance.

4.4. Practical Implications for Organisations

The predicted results highlight that even when using carefully designed, diversity-focused prompts, large language models continue to produce measurable bias across multiple dimensions. The presence of such bias in AI-generated content raises significant concerns for organisations, especially when the most frequently detected biases like disability, gender, and religion closely align with areas that are central to corporate diversity, equity, and inclusion (DEI) reporting. Failing to detect and address these biases can seriously undermine an organisation's DEI commitments and may lead to the dissemination of content that is inconsistent with the organisation's core values. Such inconsistencies can damage the credibility of AI-generated reports and erode trust among both internal stakeholders and the broader public. If biased content is publicly released, it also poses reputational risks that can be difficult to recover from.

These findings emphasise the importance of integrating automated bias detection tools, such as the model developed in this study, into organisational workflows. Proactively auditing AI-generated content can help organisations identify and mitigate representational biases early in the content creation process, ensuring that their communications remain fair, inclusive, and aligned with their stated diversity goals. A feasible approach would include creating a review workflow that

incorporates human oversight, where sentences identified by the bias detection system are sent directly to DEI specialists for assessment and modification. Organisations could establish bias threshold levels to focus their review on the most concerning content while still ensuring production efficiency.

5. Limitations and Future Works

While the model performed strongly in many areas, it is important to acknowledge its limitations. The very high accuracy and recall, particularly in the sexuality bias category, may indicate a degree of overfitting to the augmented data used during training. This could potentially limit the model's ability to maintain the same level of performance when exposed to entirely new or more complex datasets. In addition, the model's validation was conducted using a holdout set from the same synthetic dataset on which it was trained. Although the internal validation results were promising, further testing on real-world, externally sourced reports is essential to confirm the model's generalisability and practical robustness. Another limitation stems from the exclusive use of synthetic reports generated by Gemini 1.5-flash. Broader testing on reports generated by other large language models, as well as actual organisational documents, would provide a more comprehensive assessment of the model's reliability and adaptability across different content sources. Finally, despite efforts to enhance annotation quality through majority voting and confidence weighting, the inherent subjectivity involved in manual bias annotation remains a factor that may have influenced the final labels. Human interpretation of bias is inherently variable, and these subjective differences could have affected both the training data and the model's subsequent learning outcomes.

6. Conclusion

This study developed a multi-label classification model based on RoBERTa to detect five key types of bias: gender, religion, age, disability, and sexuality, within AI-generated diversity reports. The model was applied to over 10,000 sentences sourced from synthetic reports generated using the Gemini 1.5-flash language model. The results revealed measurable bias across all categories, with disability, gender, and religion biases being the most frequently detected. These findings highlight that even when using carefully constructed, diversity-oriented prompts, large language models continue to produce content that reflects subtle but persistent representational biases.

A key strength of this research lies in its confidence-weighted annotation process, which allowed the model to learn from both highly reliable and more ambiguous examples. By incorporating varying levels of annotator agreement, the model developed a nuanced sensitivity to both overt and subtle expressions of bias. Its strong predictive performance, particularly with consistently high recall across all bias types, makes it highly suitable for practical organisational use. The model can reliably screen large volumes of AI-generated content to proactively identify and flag potentially biased sentences before publication.

The study also has important implications for organisations. It demonstrates that automated systems designed to promote diversity and inclusion are not inherently immune to bias and require careful auditing. Integrating automated bias detection tools, like the one developed in this project, is essential to safeguard organisational credibility, support inclusive communication, and mitigate reputational risks.

However, several limitations should be acknowledged. The reliance on synthetic reports, the use of augmented data for minority classes, and the inherent subjectivity in manual annotation may have influenced the model's generalisability. Future research should focus on testing the model with real-world organisational documents and reports generated by different language models, as well as expanding datasets to include more diverse cultural and linguistic contexts. Overall, this research presents a scalable and adaptable framework for bias detection and contributes meaningfully to the growing field of ethical AI.

Acknowledgements

This project's success relied on the guidance, collaboration, and support of numerous individuals and organisations. Heartfelt thanks go to Cultural Infusion and Diversity Atlas for offering the platform and opportunity to investigate the convergence of AI and diversity through this research.

Appreciation is expressed to the wider research and development community at Cultural Infusion and Diversity Atlas and Cultural Infusion, whose discussions, diverse perspectives, and commitment to inclusivity profoundly shaped the project's vision and objectives, although they may not agree with all the interpretations or conclusions of the results. The authors would like to thank Peter Mousaferiadis, Mary Legrand, MyLinh Le, Aida Hakimi, Quincy Hall, Catherine Mccredie and Osvaldo Branquinho for their support in this research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Abrar, A., Oeshy, N. T., Kabir, M., & Ananiadou, S. (2025). *Religious Bias Landscape in Language and Text-to-Image Models: Analysis, Detection, and Debiasing Strategies.* arXiv: 2501.08441.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). ACM. https://doi.org/10.1145/3442188.3445922

Cornelissen, J. P. (2023). *Corporate Communication: A Guide to Theory and Practice.* SAGE Publications Ltd.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67-73). ACM. https://doi.org/10.1145/3278721.3278729

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4647-4657). ACM. https://doi.org/10.1145/2858036.2858535

Hossin, M., & Sulaiman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process, 5,* 1-11. https://doi.org/10.5121/ijdkp.2015.5201

Jentzsch, S., & Turan, C. (2022). Gender Bias in BERT—Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)* (pp. 184-199). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.gebnlp-1.20

Kondra, S., Medapati, S., Koripalli, M., Chandra Nandula, S. R. S., & Zink, J. (2025). AI and Diversity, Equity, and Inclusion (DEI): Examining the Potential for AI to Mitigate Bias and Promote Inclusive Communication. *Journal of Artificial intelligence and Machine Learning, 3,* 1-10. https://doi.org/10.55124/jaim.v3i1.249

Li, B., Haider, S., & Callison-Burch, C. (2024). *This Land Is {Your, My} Land: Evaluating Geopolitical Biases in Language Models.* arXiv: 2305.14610.

Mao, R., Tan, L., & Moieni, R. (2023). Developing a Large-Scale Language Model to Unveil and Alleviate Gender and Age Biases in Australian Job Ads. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 4176-4185). IEEE. https://doi.org/10.1109/bigdata59044.2023.10386083

Marinucci, L., Mazzuca, C., & Gangemi, A. (2023). Exposing Implicit Biases and Stereotypes in Human and Artificial Intelligence: State of the Art and Challenges with a Focus on Gender. *AI & SOCIETY, 38,* 747-761. https://doi.org/10.1007/s00146-022-01474-3

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys, 54,* 1-35. https://doi.org/10.1145/3457607

Nadeem, M., Bethke, A., & Reddy, S. (2020). *StereoSet: Measuring Stereotypical Bias in Pretrained Language Models.* arXiv: 2004.09456.

Quan, Z., & Pu, L. (2023). An Improved Accurate Classification Method for Online Education Resources Based on Support Vector Machine (SVM): Algorithm and Experiment. *Education and Information Technologies, 28,* 8097-8111. https://doi.org/10.1007/s10639-022-11514-6

R'boul, H. (2021). North/South Imbalances in Intercultural Communication Education. *Language and Intercultural Communication, 21,* 144-157. https://doi.org/10.1080/14708477.2020.1866593

Raichur, A., Lee, N., & Moieni, R. (2023). A Natural Language Processing Approach to Promote Gender Equality: Analysing the Progress of Gender-Inclusive Language on the Victorian Government Website. *Open Journal of Social Sciences, 11,* 513-529. https://doi.org/10.4236/jss.2023.119033

Shuford, J. (2024). Examining Ethical Aspects of AI: Addressing Bias and Equity in the Discipline. *Journal of Artificial Intelligence General Science (JAIGS), 3,* 262-280. https://doi.org/10.60087/jaigs.v3i1.119

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). *Gender Bias in Co-Reference Resolution: Evaluation and Debiasing Methods.* arXiv: 1804.06876.

Appendix

The complete set of code for this project is securely stored in a private GitHub repository. Access has been granted to the project collaborators. The repository can be found at the following link:

https://github.com/CulturalInfusion/Detecting-Bias-in-AI-A-Case-Study-of-LLM-Generated-Diversity-Reports-for-Tech-Companies-.git