





## Article

# Scaling Implicit Bias Analysis across Transformer-Based Language Models through Embedding Association Test and Prompt Engineering

Ravi Varma Kumar Bevara <sup>1</sup>, Nishith Reddy Mannuru <sup>1</sup>, Sai Pranathi Karedla <sup>2</sup> and Ting Xiao <sup>1,2,\*</sup>

<sup>1</sup> Department of Information Science, University of North Texas, Denton, TX 76205, USA; ravivarmakumarbevara@my.unt.edu (R.V.K.B.); nishithreddymannuru@my.unt.edu (N.R.M.)

<sup>2</sup> Department of Computer Science and Engineering, University of North Texas, Denton, TX 76205, USA; saipranathikaredla@my.unt.edu

\* Correspondence: ting.xiao@unt.edu

**Abstract:** In the evolving field of machine learning, deploying fair and transparent models remains a formidable challenge. This study builds on earlier research, demonstrating that neural architectures exhibit inherent biases by analyzing a broad spectrum of transformer-based language models from base to x-large configurations. This article investigates movie reviews for genre-based bias, which leverages the Word Embedding Association Test (WEAT), revealing that scaling models up tends to mitigate bias, with larger models showing up to a 29% reduction in prejudice. Alternatively, this study also underscores the effectiveness of prompt-based learning, a facet of prompt engineering, as a practical approach to bias mitigation, as this technique reduces genre bias in reviews by more than 37% on average. This suggests that the refinement of development practices should include the strategic use of prompts in shaping model outputs, highlighting the crucial role of ethical AI integration to weave fairness seamlessly into the core functionality of transformer models. Despite the basic nature of the prompts employed in this research, this highlights the possibility of embracing structured prompt engineering to create AI systems that are ethical, equitable, and more responsible for their actions.



**Citation:** Bevara, R.V.K.; Mannuru, N.R.; Karedla, S.P.; Xiao, T. Scaling Implicit Bias Analysis across Transformer-Based Language Models through Embedding Association Test and Prompt Engineering. *Appl. Sci.* **2024**, *14*, 3483. <https://doi.org/10.3390/app14083483>

Academic Editor: Chilukuri K. Mohan

Received: 21 February 2024

Revised: 11 April 2024

Accepted: 17 April 2024

Published: 20 April 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** AI; model scaling; Generative Pretrained Transformer (GPT); k-means; Bidirectional Encoder Representations from Transformers (BERT); transformer; prompt engineering; language models; Word Embedding Association Test (WEAT); natural language processing (NLP)

## 1. Introduction

The exponential growth of today's digital data created and distributed across online platforms has greatly accelerated the progress of natural language processing (NLP) as a vital area of study. The rapid growth of textual material on the internet has driven tremendous progress in NLP approaches, which seek to enhance computer systems' ability to comprehend, interpret, and produce human language with more complexity. As a result, NLP has become a crucial field of research, motivated by the need to tap into the possibilities of the growing world of text and enable smooth interaction between humans and computers. Information retrieval, sentiment analysis, machine translation, and content generation are just a few of the domains where NLP has had a significant transformative impact, underscoring its profound significance in the field of artificial intelligence and its broad implications for the future of human–computer interaction.

Notwithstanding these remarkable progressions, apprehensions continue to surround their practical feasibility, predominantly attributable to the intrinsic biases of the models. Cheung et al. [1] stated that prejudices, sometimes manifested as semantic biases in text, can have substantial consequences, particularly in domains such as customer reviews for different products and services, where objectivity and fairness are essential. In light of

the substantial impact that consumer evaluations have on public sentiments and actions via digital platforms, the peril of prejudiced AI systems perpetuating stereotypes and misinformation is an urgent matter of concern. Liang et al. [2] emphasized the importance of addressing and minimizing technological model biases to prevent distorted perceptions and unfairness in society.

In light of the pressing need to confront biases, scholarly investigations have progressively centered on comprehending their ramifications and formulating approaches to alleviate them. These could be cultural biases, cognitive biases, biased data, and even algorithmic biases. All these biases can lead to mistakes in how AI systems analyze data and make decisions, as demonstrated by Silberg and Manyika [3]. Furthermore, according to Mehrabi et al. [4], AI will eventually start preferring certain groups over others, which diminishes fairness. Because AI is biased, it often makes erroneous assumptions and conclusions, which lowers its trustworthiness, according to the study by Ntoutsis et al. [5].

The primary objectives of this research paper are to conduct a thorough quantitative assessment of semantic biases within transformer-based language models, utilizing the Word Embedding Association Test (WEAT) as a tool for measuring both the presence and magnitude of these biases. Additionally, the paper aims to explore the impact of model scaling on the manifestation and severity of biases, questioning whether larger models inherently amplify or mitigate biased representations. This study focuses on the exploration of prompt engineering as a corrective measure for bias reduction involving a comparative analysis of its effect scores against traditional model scaling techniques, aiming to discern which approach offers a more viable solution to the problem of bias in AI.

Prompt-based learning emerges as a promising approach to mitigating these biases. Unlike conventional training methodologies that primarily rely on extensive datasets, prompt-based learning utilizes specific cues or instructions to guide the model's reasoning process. This approach represents a paradigm shift in machine learning, aiming to refine the accuracy of models while actively reducing bias. Some of the contributions of Mayer et al. [6] to this field show how prompt-based strategies can improve model performance and lower bias in a wide range of NLP applications. Prompt-based learning takes advantage of the fact that big language models are sensitive to small changes in input. According to Solaiman et al. [7], prompters have the ability to control the behavior of models in order to achieve fairness and accuracy objectives. As opposed to debiasing training datasets, which are resource-intensive, prompt engineering provides a flexible and accessible technique to reduce automated biases even for pre-trained models.

This study adopts the Word Embedding Association Test (WEAT) as a methodological tool to evaluate the potential of prompt-based learning in addressing biases within machine learning models, focusing on the analysis of IMDb movie reviews. In the later sections of the article, it delves into related work that themes similar contents to enhance the methodological design considerations critical for addressing semantic biases in transformer-based language models. The following section presents of methodology, where the first section outlines the development of a hybrid Bert-GPT model for genre classification from reviews, and the second examines the scaling impact of transformers on bias, employing the WEAT method for measurement. The findings and results section offers a detailed interpretation based on scales, model categories, and the efficacy of prompts, paving the way for a thorough discussion on the implications of our results for developing fairer and more transparent AI systems.

## 2. Related Work

Bevara et al. [8] emphasized the importance of AI systems being responsible by ensuring fairness, providing explanations, maintaining accountability, being reliable, and gaining user acceptance. They aim to prevent biases and discuss various types, such as implicit bias, which can influence AI's decision making when analyzing movie reviews. The authors also explore methods to reduce these biases, such as using prompt-based learning and creating AI systems that are fair and transparent. The review highlights the continuous research

being conducted to develop AI technologies ethically, making sure they are inclusive and without harmful biases.

Building on the theme of utilizing AI ethically, Bevara et al. [9] took a critical look at the progress and techniques used in customer segmentation for e-commerce, utilizing machine learning and statistical modeling. It points out the drawbacks of conventional clustering methods, such as k-means, and investigates new approaches such as Gaussian mixture models and autoencoders. These innovative techniques help create more detailed and practical customer segments. The study emphasizes earlier research that demonstrates how these methods can improve marketing strategies and customer engagement by using in-depth behavioral data. It highlights the potential of combining different models and using advanced analytics to identify valuable customer groups, which allows for more targeted and efficient marketing campaigns.

Kaur et al. [10] emphasized the pivotal role of trust in the adoption and success of AI systems, identifying five fundamental pillars required to cultivate this trust: fairness, explainability, accountability, reliability, and user acceptance. Ensuring fairness is vital to prevent AI systems from perpetuating existing biases. Explainability allows individuals to grasp how AI systems function. Accountability delineates the responsibilities associated with the technology's deployment. Reliability guarantees the systems' consistent and dependable operation, while user acceptance is essential for the seamless integration of AI into designated areas. The research delves into various methodologies such as explainable AI, fairness-oriented AI, and transparent AI to fulfill these five criteria. It also discusses techniques for the verification and validation of AI systems against standards such as fairness, explainability, and accountability. By focusing on trust and adopting measures that align with these crucial standards, AI technologies stand a better chance of being embraced and effectively utilized.

B. Li et al. [11] put forward a strategy for crafting AI systems that are trustworthy, tackling critical issues like susceptibility to attacks, bias, and privacy of data. They introduced a detailed theoretical framework encapsulating elements such as robustness, adaptability, interoperability, transparency, reproducibility, fairness, data protection, and accountability. This framework outlines a holistic method for addressing the entire lifespan of an AI system and provides actionable advice for professionals and stakeholders. Highlighting the necessity for a paradigm shift towards fully trustworthy AI, the authors map out directions for future exploration and enhancement in this domain. As a foundational guide for improving AI systems, this framework marks an important step towards addressing the enduring challenges in creating AI solutions that are both dependable and ethical, urging a long-term view in solving the complexities of AI reliability and trustworthiness.

Mehrabi et al. [4] delved deeply into bias and fairness within artificial intelligence, specifically focusing on machine learning. Their research emphasizes the crucial need to confront these issues to avert the unjust treatment of certain demographic groups. Through the examination of real-life instances of bias, the study identifies various forms of bias affecting AI systems and introduces a structured taxonomy crafted by experts to define and combat bias. Covering a range of areas, including general machine learning, deep learning, and NLP, the paper sheds light on how disparities in AI applications, such as judiciary systems and facial recognition technology, can lead to adverse societal outcomes. While the paper stops short of deeply exploring particular bias mitigation strategies, it acts as an insightful primer for those keen on understanding bias and fairness in AI. It offers a comprehensive fairness framework and surveys ongoing research across different domains of AI. For NLP systems to accurately convey meaning, grasping concepts such as entailment and contradiction is essential.

The Stanford Natural Language Inference (SNLI) corpus by Bowman et al. [12] is a useful resource for creating semantic representations in natural language processing (NLP). It gives a large dataset for training and testing models with labels for entailment, contradiction, and neutrality. This corpus has facilitated both traditional and neural network models to achieve impressive results. Nevertheless, NLP tasks, such as sentiment analysis

in movie reviews, are vulnerable to bias, as reviewers' personal inclinations can taint evaluations. Variations in reviewers' backgrounds may further tilt outcomes, highlighting the urgency to mitigate bias and foster fairness in NLP. Furthermore, Kumar et al. [13] used machine learning to parse sentiment in film critiques, with the aim of augmenting customer experience and deciphering elements that affect ratings. Despite using a linear SVM model to achieve the highest accuracy, the research's small dataset size prevented it from considering potential biases in the data and algorithms. This underscores the necessity for more comprehensive studies to overcome these challenges and ensure the reliability and fairness of AI systems.

Mishra et al. [14] explored the presence of gender bias in customer reviews and its repercussions on business strategies. The analysis found that algorithms, which are designed to learn from human language, inadvertently adopt gender biases. Such biases in customer feedback can skew business decisions in ways that are not only unfair but may also violate legal standards. This research stresses the importance for businesses to recognize and address gender bias within customer reviews, particularly biases against women. It warns against relying on biased feedback for decision making, as it could reinforce harmful stereotypes and adversely affect customer relations. The study emphasizes the critical need for addressing bias in both the data used and the algorithms that process this data, aiming to ensure fairness and equity in business practices.

The study by Caliskan et al. [15] looked into whether the semantic content of language corpora has human-like biases. They used machine learning models to do the analysis. By applying a statistical model to web text data, the study identified that these datasets harbor various biases, ranging from ethically neutral to detrimental. This discovery holds considerable ramifications for artificial intelligence, underscoring the urgent need for establishing clear guidelines for non-discriminatory practices and formulations within these technologies. The research brings to the forefront the critical need to acknowledge and address biases in language data, posing essential ethical questions about the development of AI and machine learning technologies in a manner that is both equitable and devoid of discrimination.

Bolukbasi et al. [16] delved into the study of word embeddings, which are commonly used to convert text data into vector formats, and found notable gender stereotypes within embeddings derived from Google News. The research highlighted the presence of gender bias, showing that gender-neutral terms were often incorrectly associated with gendered concepts. Such biases risk reinforcing societal stereotypes through machine learning and NLP applications. To address this challenge, the researchers crafted a strategy aimed at diminishing gender stereotypes in word embeddings without detracting from their functional value. This method effectively reduced gender bias while maintaining overall system performance, suggesting a viable solution for diminishing gender bias in computational systems. The findings underscore the reality that word embeddings can carry biases into AI technologies, yet they also point towards promising approaches, similar to the one devised in this study, to curb these biases and enhance fairness in AI deployments.

Hube et al. [17] explored sentiment bias in word embeddings created by skip-gram models, highlighting that biases in training data can inadvertently lead to discriminatory classifications of names as either positive or negative. To combat this, they introduce a technique named DebiasEmb, designed to neutralize name representations within the context of positive and negative words. This method's effectiveness is confirmed through benchmark evaluations, where it is shown to mitigate name-related sentiment bias in embeddings without detracting from their quality. Nonetheless, the research is specifically tailored to sentiment bias associated with names, leaving the method's applicability to other bias types and broader ethical considerations unaddressed. While offering a valuable strategy for addressing a particular form of bias, the findings call for extended exploration into the broader application and impact of such debiasing methods.

Sengupta et al. [18] examined the effects of racial bias in data and algorithms on the effectiveness and fairness of AI systems. They highlighted that biases in language data



and models can influence explainability, user experience, and societal biases. The research showed how AI systems, influenced by biased data, can produce unjustifiable discriminatory results and affect the system's credibility. Through a mix of methods, Sengupta evaluated the direct effects of racial bias in language on AI model performance. Controlled studies further investigated the impact of biased outputs on user engagement and decision making. The findings indicated that dependence on biased models diminishes the persuasiveness of users and adversely affects their choices, underscoring the importance of mitigating bias in AI systems.

Mayer et al. [6] looked into how prompt-based learning can be used in transformer models for tasks that need professional categorization. They compared prompt-based methods with zero-shot and few-shot categorization, as well as fine-tuning and human evaluation. Prompt-based learning was noted for providing estimates of reliability and pinpointing difficult responses. They suggested a mutual rating system between humans and machines to evaluate the complexity of responses as a cost-efficient alternative to extensive fine-tuning. However, crafting effective prompts poses its challenges, and the study's focus on English-language data limits its relevance across languages. This points to the necessity for the development of multilingual AI models. Additionally, the small size of the dataset and the absence of a bias examination were seen as limitations, indicating the need for broader and more diverse datasets and a thorough investigation into potential biases. Collectively, these studies bring to light the issues surrounding bias and fairness in AI, emphasizing that, unaddressed, these issues can detrimentally influence AI performance, decision-making processes, and societal perceptions.

Sun et al. [19] explored the presence of gender bias within NLP systems and identified various underlying biases such as lexical, semantic, syntactic, and pragmatic, which all contribute to the overall gender bias. The research examined different strategies for identifying and reducing gender bias, including enhancing datasets, creating counterfactual datasets, employing adversarial training techniques, and applying post processing corrections. Despite these efforts, the study pointed out significant shortcomings, such as the reliance on debiasing strategies that operate in isolation and the scarcity of tests in real-world scenarios. It was also noted that manually crafted debiasing methods might inadvertently embed additional biases. While strides are being made toward mitigating gender bias in NLP, there is a clear need for more thorough and scientifically backed methods to tackle it effectively. Addressing biases in both data representation and algorithmic approaches continues to be a critical hurdle.

Samin et al. [20] introduced two strategies for efficiently summarizing numerous arguments or viewpoints. The initial approach involves refining pre-trained language models (PLMs) through prompt engineering, whereas the second combines prompt-based learning with a mix of argument–keypoint pairs and a classification system. Testing was conducted within specific domains and across different ones. Nonetheless, the second strategy underperformed, primarily because of PLMs' challenges with interpreting negation. This study progresses the field by demonstrating how to distill arguments into succinct points but also underlines the necessity to acknowledge the constraints of PLMs in language processing tasks. As these models become increasingly sophisticated, identifying their weaknesses is crucial for the development of more effective and dependable language processing tools.

Gupta et al. [21] introduced a sentiment classification method for evaluating movie reviews as either positive or negative. This method uses BERT embeddings along with BiLSTM-BiGRU and 1D CNN models, and on the IMDB dataset, it shows impressive accuracy and AUC scores. Despite these achievements, the technique is primarily limited to binary classification and lacks detailed explanations for the selection of model parameters. The reliance on a single dataset also raises questions about the broader applicability of the findings. For this reason, even though the first results are promising, they would be more solid if they were tested on more datasets and it was made clearer how the model configurations were chosen.

Jentzsch Turan et al. [22] looked into gender bias in common BERT models used for NLP. They defined bias as the difference in how texts about male and female subjects are felt. Through evaluating 63 classifiers on IMDB movie reviews, they uniformly detected gender biases in these pre-trained BERT models, showcasing their inherent predisposition towards such biases. The study stresses the importance of using pre-trained models conscientiously and the avoidance of ingrained biases within automated frameworks. It aims to elevate awareness regarding latent biases and encourage further investigation into how sophisticated systems can perpetuate undesired stereotypes. Their innovative approach to measuring bias enhances the capability to recognize and address bias within automated systems. Nevertheless, the research's concentration on sentiment analysis of movie reviews might restrict its wider relevance. Importantly, it did not delve into strategies for debiasing, opening avenues for future research to explore bias mitigation in broader aspects such as race, ethnicity, and sexuality. Despite these limitations, the research illuminates the pervasive nature of gender bias in sentiment analysis across various NLP models and scenarios.

Q. Li et al. [23] looked at how well BERT worked for time series forecasting and sentiment analysis, especially when it came to cloud-edge computing systems. Although BERT is celebrated for its NLP capabilities, the study revealed its limitations in time series forecasting, suggesting a deficiency in its logical reasoning prowess. Moreover, BERT's performance in sentiment analysis, for both English and Chinese, leaned excessively towards positive sentiment, indicating a potential bias. The researchers recommend enhancing BERT's logical processing during pre-training, employing prompt learning for better analytics, and exploring new models for a deeper understanding of emotions. This study emphasizes the critical need for comprehensive evaluation of pre-trained models across diverse applications to prevent biased or inadequate outcomes. However, the study's focused scope and limited sample size may affect the broader applicability of these findings. Despite these limitations, the research sheds light on BERT's analytical and reasoning gaps and suggests that relying on it for sensitive forecasting or analysis tasks should be done with caution. Further investigations are encouraged, but adapting and refining pre-trained models for specific purposes appears to be a wise approach.

Manzini et al. [24] introduced a novel approach for extending debiasing techniques from binary attributes, such as gender, to encompass multiclass attributes, such as race and religion. This method starts by pinpointing the bias subspace using sets of words that define each social group. Following this identification, bias components are eliminated from the embeddings through enhanced versions of previous hard and soft debiasing techniques. A new metric called mean average cosine (MAC) is used to measure how well this multiclass debiasing works. It is calculated across words that are likely to be biased and attribute words that should not be linked. This strategy marks a significant step forward in generalizing the debiasing process from binary to multiclass attributes, offering new avenues for addressing complex biases in the real world. Nonetheless, challenges remain due to the subjective nature of bias and the variation of lexicons across different cultures.

Ravfogel et al. [25] looked into the issue of biases in neural representations, such as word embeddings or classifiers, which can reinforce societal stereotypes and cause groups to be treated unfairly because of their gender, race, or other factors. The study introduces the Iterative Nullspace Projection (INLP) technique, designed to strip target bias-related information from these representations. The study uses INLP to fight two main types of bias: representational biases that strengthen stereotypes in word embeddings and allocation harms that show up in differences in how different groups are classified. The fact that INLP can close the gender gap in true positive rates for occupation classifications and reduce gender-biased groupings and associations in embeddings is evidence of its usefulness. This work advances the development of algorithms capable of excising sensitive biases from models without compromising their functionality, although it also emphasizes the ongoing need for more nuanced approaches to defining and measuring biases.

Urman et al. [26] examined how political bias differs among various language models, such as GPT and Bard, particularly when dealing with politically charged questions. It reveals that these models' tendencies toward information censorship and spreading misinformation greatly fluctuate with the language used. This indicates that chatbots may process politically sensitive information in distinct ways across different languages. Such variability highlights the challenge of addressing bias in AI systems and underscores the necessity of incorporating linguistic diversity into strategies aimed at reducing political bias and preventing the spread of misinformation.

Rajapaksha et al. [27] investigated the application of transfer learning models, such as BERT, XLNet, and RoBERTa, for identifying biased clickbait news headlines on Twitter. They experiment with various updates to BERT, XLNet, and RoBERTa aimed at mitigating biases related to race, gender, and more present in these models. According to metrics such as accuracy, precision, recall, and F1-score, modifications to model outputs and the addition of new layers can help reduce bias and improve fairness, with RoBERTa demonstrating the greatest improvement. Despite these efforts, some biases persist, particularly in terms of higher false positives for minority groups. The findings indicate a need for more sophisticated debiasing techniques and further research into broader definitions of fairness.

González, F. et al. [28] proposed a methodology for clustering and analyzing movie reviews to classify them by genre using natural language processing techniques. The authors compared the performance of different word vectorization techniques, namely TF-IDF and Word2Vec, in conjunction with k-means clustering. Looking ahead, the authors suggest exploring the potential of advanced text embedding techniques such as Doc2Vec and BERT word embeddings. By leveraging the power of these embedding approaches, future research could aim to develop a hybrid BERT-GPT model for enhanced genre classification. Such a model would combine the strengths of BERT's bidirectional context understanding with GPT's generative capabilities, potentially enabling more accurate and nuanced genre prediction.

Building upon the foundation laid by previous research, various types of biases, such as gender, racial, and political biases, have been identified, which can perpetuate stereotypes and lead to discriminatory outcomes. Researchers have proposed several techniques to mitigate these biases, including data augmentation, adversarial training, post-processing corrections, and debiasing methods such as iterative nullspace projection (INLP). However, challenges remain in effectively measuring and eliminating biases across different domains and languages. Furthermore, it emphasizes the need for more comprehensive and scientifically validated approaches to ensure the reliability and fairness of AI systems. The proposed hybrid BERT-GPT model for genre classification seeks to address the limitations of existing approaches by leveraging the complementary strengths of these state-of-the-art language models. Furthermore, prompt engineering techniques, which are key for large models to understand the context of the nuance, are proposed to mitigate the biases compared to the proposed techniques.

### 3. Methodology

The study utilizes a sequential mixed methods technique in two phases to examine and address bias in transformer-based language models. During the initial stage, a combined framework is created that utilizes BERT and GPT-3 to perform semi-supervised genre classification of movie reviews. The BERT model is used to generate dense vector representations, which are then clustered using the k-means algorithm for review embeddings. In this study, we briefly introduce this novel genre classification method utilizing a hybrid BERT-GPT approach to generate movie genres based on IMDb reviews, laying the groundwork for more in-depth exploration in future work, while our primary focus remains on mitigating LLM bias through appropriate prompts.

In the second phase, a thorough bias study is conducted by assessing different versions of advanced transformer topologies on a larger scale. The Word Embedding Association Test (WEAT) is employed to accurately measure bias patterns across several models, such

as BERT, RoBERTa, T5, and XLNet, encompassing both basic, large, and x-large versions. The target terms are obtained from movie reviews that are categorized as either favorable or negative. The goal is to analyze the relationship between different movie genres and their associated positive or negative features. Additional engineering experiments are conducted to assess the possibility of customizing adjustments to reduce the identified biases.

All experiments were conducted on a system equipped with an Intel Core i9-12th generation CPU, an NVIDIA GeForce RTX 4090 GPU, and 80 GB of RAM, running on the Windows operating system. The implementation was done using the Python 3.10 programming language and the Transformer library from Hugging Face to handle the various transformer-based architectures.

### 3.1. Word Embedding Association Test (WEAT)

The WEAT evaluates bias by measuring associations between two sets of target words and two sets of attribute words, quantifying bias in word embeddings. The method by Caliskan et al. [15] illuminates how different words relate to various attributes in a computational context.

#### 3.1.1. Mathematical Formulation

Given target word sets  $X$  and  $Y$ , and attribute word sets  $A$  and  $B$ , the association strength is assessed as follows:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \quad (1)$$

where  $s(w, A, B)$  denotes:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b) \quad (2)$$

This quantifies the differential association of target sets with attribute sets. Equation (1) represents the test statistic, which quantifies the extent to which the two sets of target words are associated with the attribute sets. Equation (2) provides a definition for the association strength  $s(w, A, B)$ , which is determined by calculating the cosine similarity between word  $w$  and the attribute words in sets  $A$  and  $B$ . Equation (3) provides the cosine similarity.

#### 3.1.2. Effect Size

The effect size  $d$  is calculated to gauge the bias magnitude:

$$d = \frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)} \quad (3)$$

The effect size, denoted by the symbol  $d$ , is a statistical metric that quantifies the degree of dissimilarity between the two groups of target words and the characteristics, expressed as the standard deviation of the associations obtained. This measures the standard deviation of association strengths, providing insight into the bias's significance. When the absolute value of  $d$  is greater, it shows that the bias is more significant. A measurement of the relative strength of the connection between the two sets of target words and the attribute words is provided by the WEAT score. A score that is positive implies that the set  $X$  has a greater relationship with the attribute set  $A$  in comparison to the set  $Y$ . On the other hand, a score that is negative shows that the other set  $Y$  has a stronger link with  $A$  in comparison to  $X$ . This is supplemented by the effect size, which offers a measurement of the amount of this difference and might be of assistance in comprehending the relevance of the bias.

Effect size is a statistic that is utilized in this study for the purpose of measuring the discrepancy that exists between two cohorts. The effect size is a statistical metric that allows for the comparison of averages between two groups while taking into consideration the variability of the data. On the other hand, it offers extremely helpful insights into the practical meaning of the discrepancies that were found. The presence of larger impact sizes

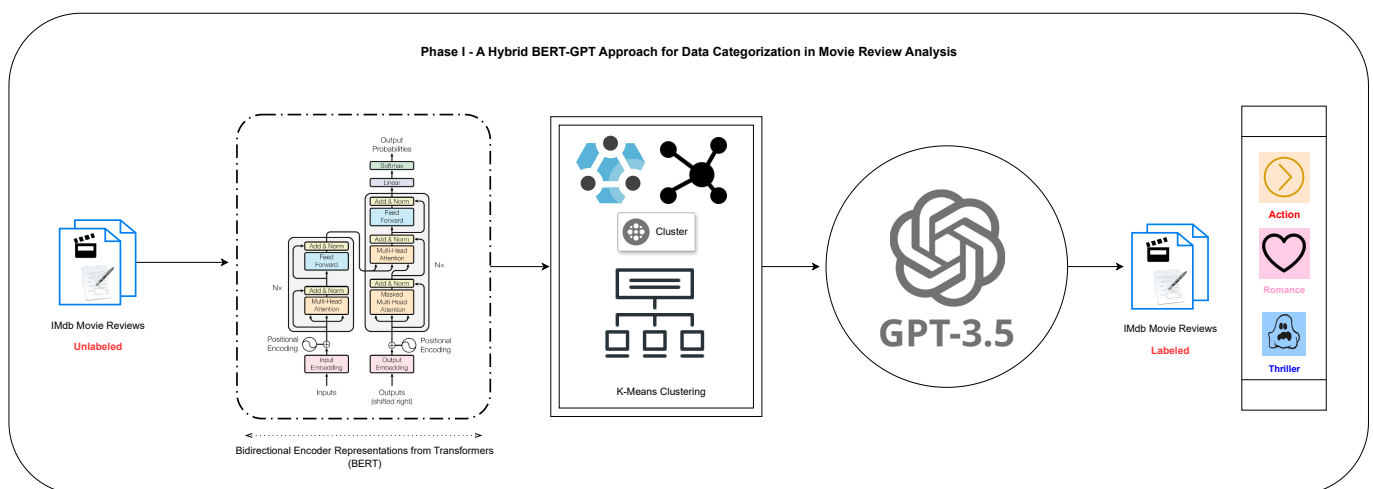
is indicative of the existence of stronger correlations between the target and attribute word sets. In addition,  $p$ -values can be utilized in order to evaluate the statistical significance of data, which allows for the determination of the chance of getting outcomes that are as severe as the results that were seen under the null hypothesis. Although the research do not discuss the  $p$ -value for significance of the effect sizes in results, incorporating  $p$ -values in future analyses could enhance the robustness of findings by confirming that observed differences are not due to random chance, thereby strengthening the evidence for the efficacy of the bias mitigation techniques employed. Rather, this enables to evaluate the existence of biases in the analysis as well as the extent to which they occur.

### 3.2. Data Preparation

This research is based on a dataset that consists of 50,000 reviews of movies that were found on IMDb by Maas et al. [29]. These reviews were previously used for studies that used sentiment analysis. The procedure for preprocessing was rigorously created to guarantee the consistency and integrity of the data, which is essential for the future phases of analysis. In order to clean up the content, this required the removal of HTML elements, special characters, and hashtags in a methodical manner. In addition, all of the textual material was converted to lowercase in order to avoid differences that were brought about by capitalization choices. In order to place greater emphasis on more significant lexical features, stopwords, which contributed very little by way of analytical value, were eliminated. In addition, unnecessary spaces were removed, which resulted in an improvement in the dataset's readability and ease of comprehension. These preceding processes were essential in the process of constructing a cohesive and analyzable corpus, which laid a strong foundation for the extensive bias analysis that was to follow.

### 3.3. Hybrid BERT-GPT Approach for Semantic Annotation of Data

In Phase I depicted in Figure 1, the methodology advances the frontiers of semantic data annotation through the integration of BERT and GPT-3.5 architectures. Leveraging the robust feature-extraction capabilities of BERT, the model distills the essence of movie reviews into dense vector representations. These vectors are methodically categorized using the k-means clustering algorithm to identify distinct thematic groupings. Subsequently, the labeled data corpus serves as a training bed for GPT-3.5, which hones a genre-specific classifier through its advanced predictive modeling. This classifier is adept at assigning genre labels—Action, Romance, Thriller—to the clustered reviews, as illustrated in the model architecture.



**Figure 1.** The hybrid BERT-GPT architecture for semantic annotation in movie review analysis. This schematic illustrates the combined use of BERT for contextual embedding generation and k-means clustering, followed by GPT-3.5 for fine-tuning the classification of movie reviews into genres: Action, Romance, and Thriller.



The synergy between BERT’s contextual comprehension and GPT-3.5’s generative precision culminates in a refined approach to data categorization. By transcending the conventional genre classification paradigms, this hybrid model not only offers enhanced accuracy in semantic annotation but also establishes a robust foundation for the intricate bias analysis that ensues. The process symbolizes a paradigm shift in automated text classification, setting a new benchmark for subsequent bias analysis frameworks in NLP research.

The word clouds are generated from each cluster formed at the final stage of Phase I. After the reviews have been transformed into vectors using the BERT model, they are then grouped into clusters using the unsupervised K-Means approach. These clusters represent the similarity between the vectors. Since these clusters are created based on reviews, a new attribute cluster is added to the original data to keep track of which reviews belong to each cluster. The GPT 3.5 model is given a prompt to annotate the genre of the reviews supplied, which are sampled from each cluster. The genres to choose from are Action, Romance, or Thriller. This technique involves annotating the clusters with suitable genres, which are then added to the original data. As a result, each review is associated with both the cluster and the genre. After annotating these genres to the reviews, the word cloud analysis aggregates the reviews belonging to each cluster. The word cloud visually represents the most prominent or frequent words from the body of the text, which are the reviews for this article. This signifies the process of isolating target words in the association test from the original data, rather than using more generic words that represent the aim for the analysis.

### 3.4. Quantifying and Mitigating Scaling Biases in Transformer-Based Language Models

The second part of the investigation looks at implicit biases in three levels of transformer-based language models: base, large, and x-large. This is completed very carefully using the Word Embedding Association Test (WEAT), after the data have been annotated. Central to WEAT is the concept of target and attribute words. Target words are genre-specific terms derived from the corpus—action words such as “explosion” or “gun”, romantic words such as “love” or “heart”, and Thriller-related terms including “mystery” or “terror”. These words are pivotal, as they anchor the semantic field intended to scrutinize for bias. In parallel, attribute words serve as the poles of sentiment bias, categorized into positive terms, such as “excellent” or “awesome” and negative terms, such as “terrible” or “awful”. The importance of selecting representative words cannot be overstated, as they directly influence the validity of the WEAT scores.

Bounded by the influence of the target and attribute words, this phase formulated a target set of words for each genre via word cloud analysis, utilizing the semantically annotated data from Phase I. The word clouds distill the essence of each genre, enabling us to crystallize a representative lexicon that embodies the thematic characteristics of the Action, Romance, and Thriller genres. These lexicons are then harnessed as the target words in the Word Embedding Association Test (WEAT), facilitating a nuanced bias measurement within and across the genres.

Subsequently, an assessment of bias across a spectrum of transformer-based language models is engaged, stratified by scale: base, large, and x-large. This hierarchical approach reveals the non-linear relationship between a model’s scale and its implicit biases. Through this systematic evaluation, using the WEAT, discerning the variations in bias as a function of model complexity offers a granular view of how scaling affects bias tendencies in natural language processing tools.

The designs ‘bert-base-uncased’ and ‘roberta-base’ are considered fundamental in transformer technology. These models achieve a compromise between computing efficiency and language understanding by utilizing parameters ranging from tens to hundreds of millions (Devlin et al. [30]; Liu et al. [31]). Base models serve as a reference point for grasping the inherent bias present in transformer models at the outset of the investigation (Bender et al. [32]; Caliskan et al. [16]). As detailed in Table 1, due to their lesser complexity,

they are well-suited for detecting and analyzing bias in the beginning, serving as a reference point for measuring the effect of scale on bias.

**Table 1.** Overview of base transformer models: this table presents the base models used in the study, highlighting their names and parameter counts, providing a foundation for understanding bias in less complex language models.

Model Name	Number of Parameters (#Params)
albert-base-v2	11 M
google/electra-small-discriminator	14 M
t5-small	60 M
distilbert-base-cased	65 M
distilbert-base-uncased	66 M
microsoft/deberta-base	86 M
bert-base-uncased	110 M
bert-base-cased	110 M
google/electra-base-discriminator	110 M
xlnet-base-cased	110 M
gpt2	117 M
roberta-base	125 M
facebook/bart-base	139 M
t5-base	220 M
gpt2-medium	345 M

Highly sophisticated models such as ‘bert-large-uncased’ and ‘roberta-large’ have parameter counts in the hundreds of millions, explained from Table 2. Expanding the parameters enables a more detailed comprehension of language, perhaps including a wider range of biases. These models provide an intermediate point for investigating the connection between model size and bias expression in the research. Their increased ability for language modeling makes them especially valuable for detecting minor biases that might not be obvious in smaller models.

**Table 2.** Specifications of large transformer models: summary of the large models, with details on their architecture and number of parameters, illustrating the intermediate complexity explored for bias analysis.

Model Name	Number of Parameters (#Params)
albert-large-v2	17 M
microsoft/deberta-large	304 M
google/electra-large-discriminator	335 M
bert-large-uncased	336 M
bert-large-cased	336 M
roberta-large	335 M
xlnet-large-cased	340 M
facebook/bart-large	406 M
gpt-large	774 M
t5-large	770 M

Similar to the other model scales, from Table 3, extra large transformer models such as ‘gpt2-xl’ and ‘t5-3b’ are at the forefront of contemporary transformer model design, including billions of parameters. This high degree of intricacy provides exceptional opportunities for in-depth language research, enabling the identification of even the subtlest biases. x-large models in the work offer insights into the maximum levels of bias scaling. These models have a high level of language knowledge, which is very useful for assessing the effectiveness of bias mitigation measures on a broad scale.

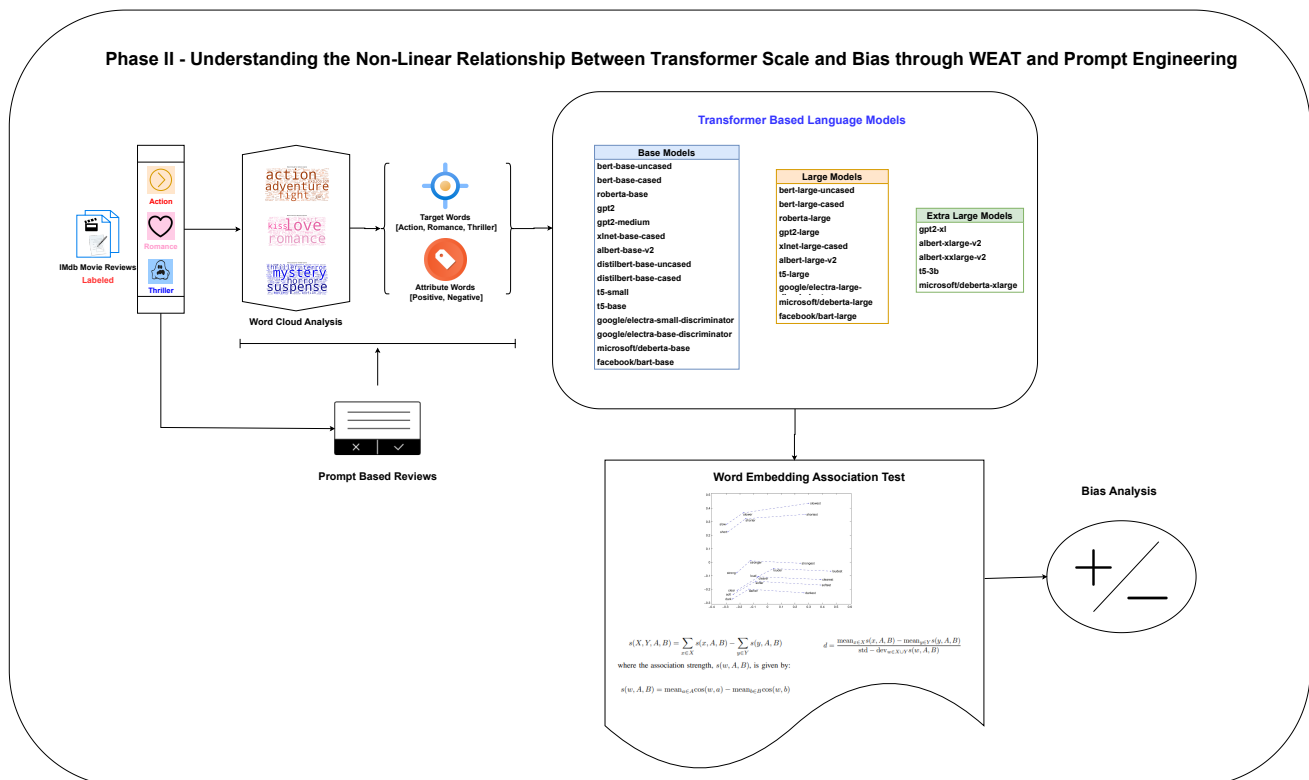
This delineation of models by scale is not merely a technical classification but forms the bedrock upon which the relationship between model complexity and bias is examined. It is hypothesized that as models scale in size, their ability to capture nuanced language

patterns increases, potentially altering the landscape of embedded biases. To facilitate this extensive computational task, the power of GPUs is harnessed, providing the necessary computational efficiency to process large volumes of data rapidly. The use of GPUs is especially critical when generating word embeddings and calculating association strengths across the extensive array of models evaluated.

**Table 3.** Details of x-large transformer models—the most complex models examined, focusing on their expansive parameter counts, to investigate bias at the highest scale of the model architecture.

Model Name	Number of Parameters (#Params)
albert-xlarge-v2	58 M
albert-xxlarge-v2	223 M
microsoft/deberta-xlarge	700 M
gpt2-xl	1.5 B
t5-3b	3 B

The architecture of this analytical endeavor is visually depicted in Figure 2, which illustrates the methodological pipeline—from the inception of genre-specific word clouds to the application of WEAT, and the strategic implementation of prompt engineering. The figure accentuates the multifaceted nature of the approach, including the generation of word clouds for each genre, which informs the selection of target words. These words are subsequently employed to measure biases within the corpus of model embeddings. Subsequently, prompt engineering is implemented as an intervention and employed on the foundational models to assess its efficacy in mitigating bias. The comparative analysis of WEAT effect sizes, pre- and post-prompt intervention, provides a nuanced understanding of how subtle linguistic adjustments can influence the manifestation of biases.



**Figure 2.** Phase II methodological architecture—the comprehensive approach taken in Phase II, from the generation of genre-specific word clouds and target word selection through to the application of WEAT across the base, large, and x-large models, and the implementation of prompt engineering techniques for bias mitigation.

### 3.5. Prompt-Based Learning for Bias Mitigation

In the realms of machine learning and natural language processing (NLP), the concept of prompt-based learning emerges as a groundbreaking strategy. This methodology diverges from conventional training techniques, which often depend on extensive datasets. Instead, it harnesses the power of carefully constructed textual prompts to direct the behavior of AI models during the learning phase. This approach leverages prompts to invoke specific responses from pre-trained models, utilizing their inherent knowledge to produce outcomes tailored to the given prompts. By drawing on the pre-existing knowledge that these models have encoded, this technique enables them to produce responses or predictions in response to the context that the prompt provides. In the context of prompt engineering, this approach takes on a critical role in bias mitigation efforts. By carefully crafting prompts that are devoid of biased language or that specifically counteract known biases within the model, researchers can steer the model's output towards more equitable, neutral, and balanced responses. This not only enhances the model's ability to deal with diverse and nuanced human languages but also aligns its outputs more closely with ethical and fairness guidelines. In the study, prompt engineering is used on base models to create a controlled environment for testing how well it works at reducing bias across different transformer models.

In exploring prompt engineering for bias mitigation in transformer-based language models, the methodology is applied in both the phases. Initially, it facilitated for the data annotation to categorize movie reviews into specific genres. This is achieved through structured prompts that guide the model GPT 3.5 in classifying each review set into one among several predefined genres, based on cluster content. This approach is demonstrated by a prompt designed to direct the model's classification efforts, taking into account the exclusion of certain genres for accurate categorization. The prompt that was used in the initial phase is listed below.

```
prompt_for_data_annotation = f"""
    Given a set of movie reviews, categorize them into one specific genre:
    {'', '.join(available_genres)}.
    Each set of reviews is from a distinct cluster and should correspond to one of
    these genres.

    Cluster {cluster_id} Reviews: {reviews}
    Genre for Cluster {cluster_id} (excluding {'', '.join(excluded_genres)}):
    """
```

Expanding on the utility of prompt engineering, a bias reduction technique, inspired by authors in the article [8], targeting the mitigation of inherent biases in the models' responses. By employing genre-specific prompts that contextualize each review within a movie genre, the aim is to neutralize outputs, diminishing biased interpretations across genres, such as Action, Romance, and Thriller. In this context, it is applied to five distinct base models. A before-and-after analysis of the WEAT effect size reveals the impact of prompts on bias mitigation. The findings from this phase could chart a course for future interventions in bias reduction, underscoring the pivotal role of prompt engineering in refining model outputs. Figure 2 shows a visual representation of this whole process. It walks you through the complex parts of the methodology, from creating genre-specific word clouds to checking for bias in transformer-based language models.

```
prompt_for_each_genre = f"""
This review is about an Action movie. The review says:
This review is about a Romance movie. The review says:
This review is about a Thriller movie. The review says:
"""
```

This structured approach not only exposes the current state of biases within these models for the movie reviews, but also sets the stage for the subsequent application of prompt engineering techniques. In the latter part of this phase, prompt engineering is applied to base models to observe shifts in WEAT effect sizes. By comparing the effect sizes before and after the application of prompts, the aim is to assess the extent to which prompt engineering can serve as a bias mitigation strategy. This step is crucial in understanding the potential for model adjustments to foster more equitable outcomes in language technology applications.

#### 4. Findings and Results

The thorough research on several variations of transformers provided interesting insights into how model size relates to embedded implicit biases in movie review analysis. Evaluation utilizing the Word Embedding Association Test (WEAT) technique confirms that increasing capacity somewhat alleviates biases, since base models show stronger biases towards specific genres compared to their scaled-up versions. For reproducibility, the code has been made available on GitHub (access the complete source code and datasets for this study at: <https://github.com/Deep6Lab/Bias-Analysis> (accessed on 11 April 2024)). For the GPT-2 medium model, an effect size of 0.635 for Action genre indicates a moderate to strong association between the embeddings of words in Action movie reviews and the attributes typically linked to Action movies. Similarly, the effect size of  $-0.805$  for Romance indicates a strong negative association, suggesting that words in Romance movie reviews are inversely related to the attributes we might expect to find in Romance genres. Conversely, the adjusted readings of 0.019,  $-0.005$ , and  $-0.011$  in the GPT-2 large model suggest that there is almost no association between the embeddings and the expected genre attributes, implying that the larger model has a much-reduced genre bias. These findings support the increasing evidence that larger scales allow models to address specific discriminatory connections.

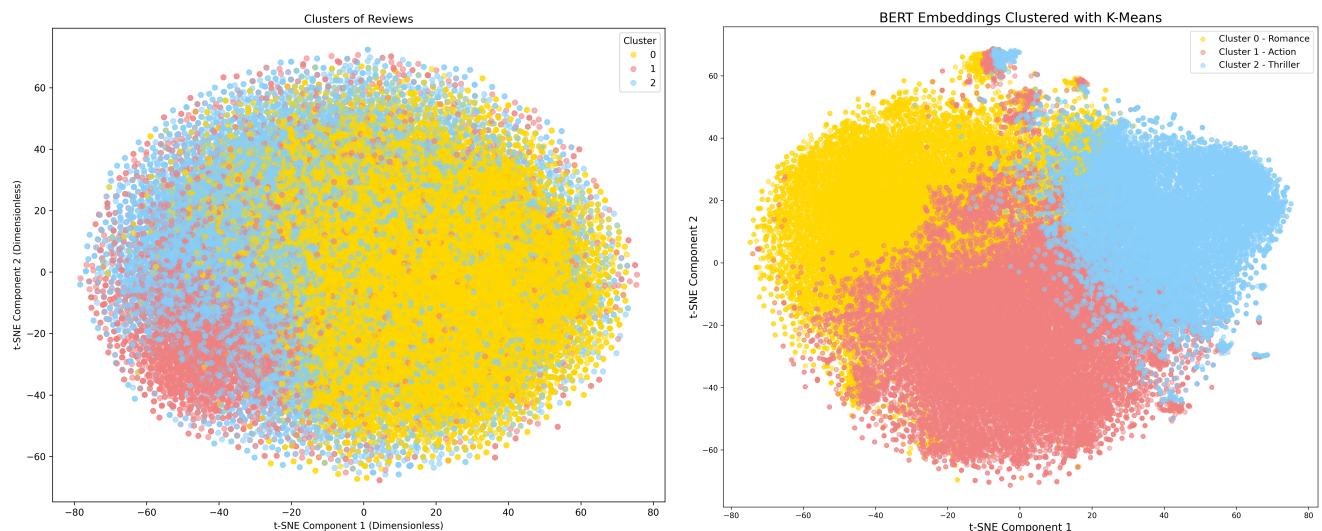
Model tuning is more effective than just increasing the scale when it comes to removing bias. Following engineering, targeted genre-specific adjustments enable us to decrease average bias by more than 34.2% across all base transformer designs being studied. Debiasing through prompts is more effective than achieving a 29.1% bias drop by increasing the model size from base to x-large, highlighting the importance of guided conditioning. BERT shows significant improvements, reducing biases in Action, Romance, and Thriller genres by up to 42% after applying the prompt. Efforts must be quickly adjusted to make model creation more accessible based on these numerical findings.

The initial stage of this research project consisted of classifying movie reviews according to the genres that corresponded to what they were written about. An innovative labeling process was embarked upon using two distinct methodologies, a TensorFlow TextVectorization approach and a transformer-based language model, specifically BERT (Bidirectional Encoder Representations from Transformers). This was completed in response to the fact that there was a lack of pre-annotated genre data for movie reviews. The findings of the BERT-based clustering, which were shown using t-SNE (t-distributed stochastic neighbor embedding), exhibited a strong semantic distinction among the three target genres, which indicated a robust grasp of the theme content of the reviews.



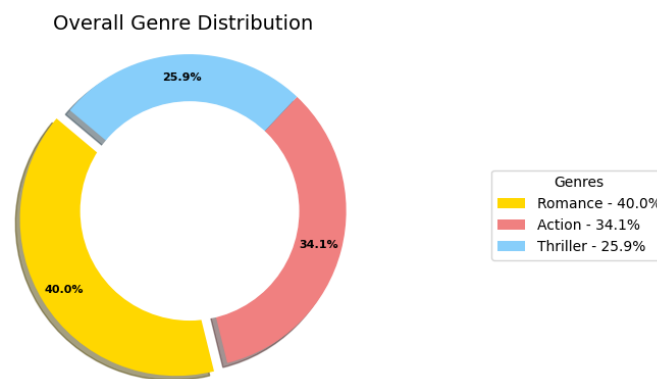
#### 4.1. Semantic Data Annotation Key Findings

The utilization of BERT embeddings, which was then followed by k-means clustering, resulted in the formation of different categories that captured the semantic characteristics of the reviews. The effectiveness of this strategy is demonstrated by the scatter plot that can be seen in Figure 3. This plot displays clearly defined clusters that correlate to the genres of Romance, Action, and Thriller that are being discussed. The significance of this distinction lies in the fact that it demonstrates the model's capacity to recognize and classify intricate narrative components that are intrinsic to the reviews.



**Figure 3.** Comparative visualization of genre clusters derived from TensorFlow TextVectorization and BERT embeddings.

Subsequently, a novel approach was utilized to prompt the GPT-3.5 model using the OpenAI API (version: 1.7.0), aiming to refine the genre classification further. The prompt design incorporated a brief description of the task, followed by a sample of reviews from a specific cluster, excluding genres already identified. This method allowed for a dynamic and context-aware classification that leveraged GPT-3.5's linguistic model to ascertain the most probable genre. Post-classification, the genre distribution was analyzed to assess the balance and representativeness of each genre within the dataset. The genre distribution pie chart in Figure 4 indicates a relatively even distribution among the genres, affirming the classification method's effectiveness. This balance is crucial for the integrity of the subsequent phase of bias analysis, ensuring that no genre disproportionately influences the results. The combination of BERT embeddings for semantic understanding and GPT-3.5 for contextually aware genre classification has proven efficient, as seen in the overall genre distribution. The balance achieved through this method sets a strong foundation for the next phases, where a more granular bias analysis will be conducted. This structured approach not only strengthens the reliability of the data preparation phase, but also enhances the potential for insightful findings in bias measurement and mitigation strategies.

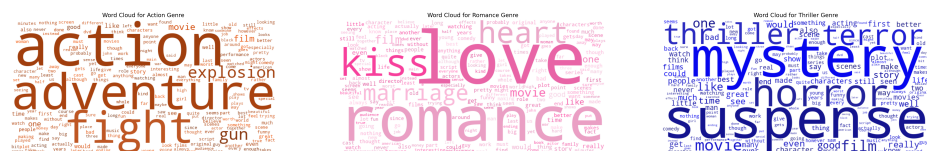


**Figure 4.** Genre distribution post-semantic annotation, depicting the proportion of movie reviews across Romance, Action, and Thriller genres.

#### 4.2. Assessing Implicit Bias in Language Models: WEAT Analysis across Genres and Model Scales

In this phase of the study, target word sets for the Word Embedding Association Test (WEAT) analysis are crafted with a random sampling of reviews from the genre-specific clusters formed in Phase I. By leveraging these representative samples, word clouds, a visual representation of word frequency within the text with larger fonts signifying higher-frequency words, are generated. These word clouds were instrumental in identifying the most salient words within each genre, which in turn constituted the target sets for the WEAT. The word cloud for the Thriller genre, conspicuously dominated by words such as “terror” and “suspense”, reflects the intense and gripping nature of this genre. Similarly, the Romance genre word cloud is pervaded by terms such as “love” and “heart”, epitomizing the emotional and affectionate themes characteristic of this category. The Action genre is aptly represented by words such as “explosion” and “adventure”, capturing the dynamic and high-energy essence of these films.

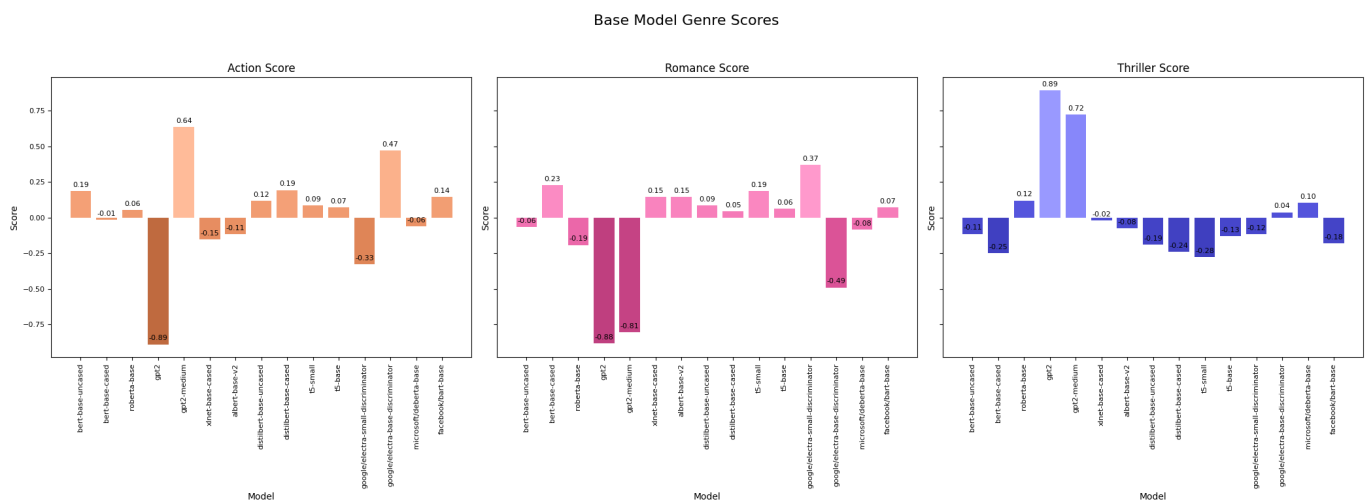
These visuals from Figure 5 not only informed the selection of target words, but also provided an intuitive understanding of the contextual leanings within each genre. For instance, “terror” in the Thriller genre word cloud not only indicates frequency but also an underlying narrative focus that is paramount to this genre’s identity. By carefully curating these words, it was ensured that the subsequent bias analysis through WEAT would be grounded in genuine linguistic usage patterns. The process of discerning these target words from the word clouds is both an art and a science; it requires an analytical eye for frequency and relevance, as well as a nuanced understanding of genre-specific lexicons. The words selected from these clouds form the backbone of the bias analysis, serving as a litmus test for the inherent biases within the language models evaluated. The careful construction of these word sets is a testament to the rigorous and data-driven approach that underpins this phase of the research.



**Figure 5.** Comparative word clouds for Action, Romance, and Thriller genres: these visualizations encapsulate the most prominent terms extracted from movie reviews, illustrating the distinctive lexical fields that characterize each genre.

As the research deals with model scaling, it employs a suite of base transformer models, including BERT-base (both uncased and cased), GPT-2 (including its medium variant), RoBERTa-base, XLNet-base-cased, A Lite BERT (ALBERT-base-v2), DistilBERT (both uncased and cased), and Text-to-Text transfer transformer (T5) (small and base), along with specialty models such as Google’s ELECTRA (small and base discriminators),

Microsoft’s DeBERTa-base, and Facebook’s BART-base. These models, as visualized in Figure 6, are foundational in natural language processing, each designed with unique architectural nuances to capture and generate human-like text. In order for BERT models to comprehend a word’s context, it is necessary to train transformers in both directions. GPT-2 is an autoregressive model that uses word sequence prediction to generate text. To improve efficiency, RoBERTa tweaks BERT’s pre-training technique, and XLNet uses permutation-based training to pick up on bidirectional contexts. In order to make training faster and more efficient, ALBERT provides methods for reducing training parameters. DistilBERT streamlines BERT to provide a more lightweight version while preserving the majority of its prior version’s effectiveness. With T5, any natural language processing issue may be converted to text. DeBERTa incorporates disentangled attention processes, which improves on BERT and RoBERTa, ELECTRA trains more efficiently by discriminating between “real” and “fake” input tokens, and BART uses a denoising autoencoder for pre-training.



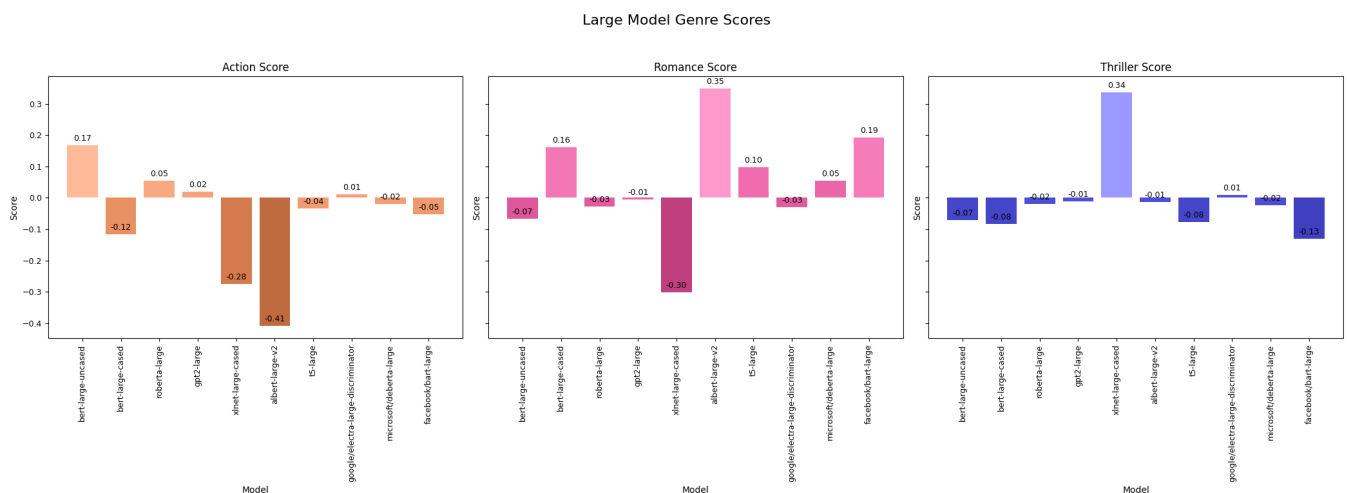
**Figure 6.** Effect sizes across genres for base models: this GRAPH illustrates the comparative analysis of effect sizes for genres such as Action, Romance, and Thriller across base transformer models, including BERT-base, GPT-2, and RoBERTa-base, highlighting their inherent biases.

As explained in Table 4, the analysis of effect sizes across genres for each model reveals insightful patterns of bias. BERT-base-uncased showed a tendency towards negative bias in Action, indicative of its processing of thematic elements within the genre. GPT-2 significantly leaned towards positive associations in Thriller, suggesting a predisposition towards engaging, suspenseful content. RoBERTa-base exhibited a relatively balanced approach, yet with slight genre-specific inclinations. GPT-2-medium’s strong positive bias in Action underscores its alignment with dynamic, high-energy narratives. XLNet-base-cased maintained a balanced profile, hinting at its robustness across diverse contexts. ALBERT-base-v2, DistilBERT, T5, ELECTRA, DeBERTa, and BART models each demonstrated unique bias spectra, reflecting the complex interplay between model architectures and genre characteristics.

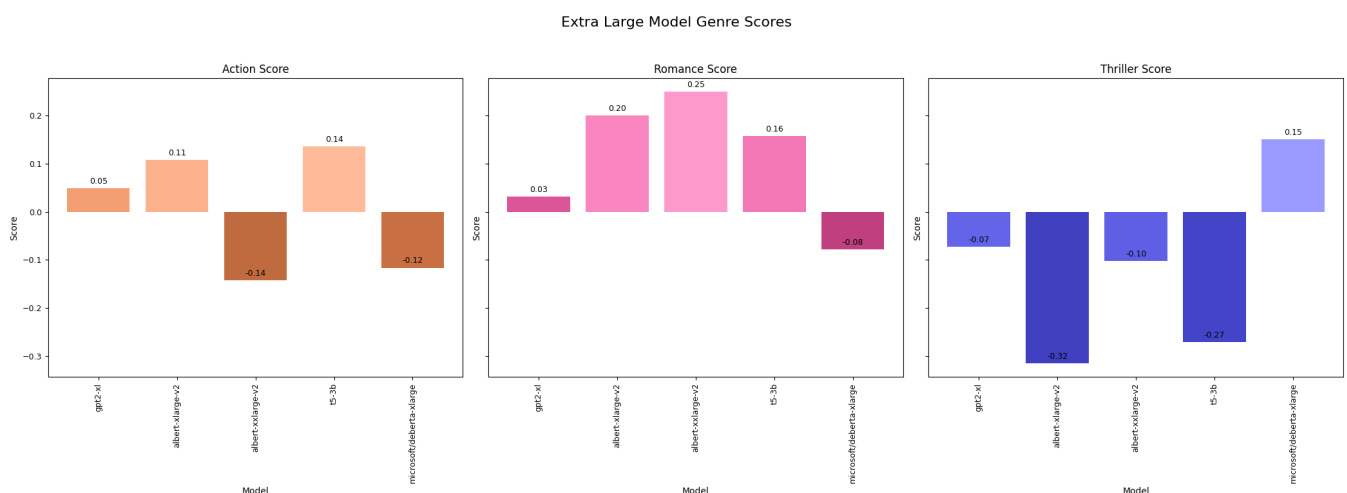
To include large and x-large transformer models in this study, advanced variants such as BERT-large, GPT-3, T5-large, and others were added. These were made to handle more difficult and nuanced language tasks (see Figures 7 and 8). These models, with their increased parameter counts, offer deeper contextual understandings, making them capable of generating and interpreting text with a higher degree of sophistication. Large models such as BERT-large and RoBERTa-large, with their extensive training data and advanced architectures, are adept at capturing intricate patterns in text. The x-large models, including GPT-3 and T5-3B, push the boundaries further, utilizing billions of parameters to achieve state-of-the-art performance across a broad spectrum of NLP tasks. This escalation in model complexity and capacity is pivotal for exploring the nuanced dynamics of language and bias at scale.

**Table 4.** Quantitative bias analysis in base transformer models: summary of the effect sizes calculated for Action, Romance, and Thriller genres across a selection of base transformer models, providing a foundational understanding of bias distribution.

Model	Action Score	Romance Score	Thriller Score
bert-base-uncased	0.18825592	−0.06431729	−0.11416924
bert-base-cased	−0.014822054	0.22913459	−0.24811955
roberta-base	0.055248357	−0.19221295	0.11987918
gpt2	−0.88915974	−0.8826492	0.8945391
gpt2-medium	0.6357243	−0.80575037	0.72424453
xlnet-base-cased	−0.15417647	0.1468171	−0.02132024
albert-base-v2	−0.114646584	0.14561488	−0.07606181
distilbert-base-uncased	0.1158604	0.08697143	−0.18819211
distilbert-base-cased	0.1930672	0.045182917	−0.23972207
t5-small	0.08636108	0.18825912	−0.27628157
t5-base	0.07314057	0.06357889	−0.12933907
google/electra-small-discriminator	−0.32698292	0.37106583	−0.11791928
google/electra-base-discriminator	0.47093016	−0.49049643	0.037513744
microsoft/deberta-base	−0.060900953	−0.0841595	0.103097126
facebook/bart-base	0.14394334	0.07040074	−0.17999497



**Figure 7.** Genre-specific effect sizes in large models: the variation in effect sizes for Action, Romance, and Thriller genres across large-scale models such as BERT-large and GPT-3, showcasing the nuanced understanding and potential biases within these advanced architectures.



**Figure 8.** Bias assessment in x-large models: this chart presents the effect sizes for different genres analyzed using x-large models such as GPT-3 XL and T5-3B, offering insights into the scale of biases at the pinnacle of model complexity.

Upon analyzing the effect sizes produced by these larger-scale models across genres, distinct patterns emerge, reflecting each model's unique handling of textual data. For instance, GPT-3's expansive knowledge base allows for an unprecedented level of nuance in genre classification, exhibiting a balanced representation across the Action, Romance, and Thriller genres. Meanwhile, T5-large and T5-3B models, with their text-to-text framework, demonstrate an ability to discern and categorize nuanced thematic elements, revealing their potential to mitigate inherent biases more effectively. The larger models generally show a trend towards more balanced or nuanced biases compared to their base counterparts, suggesting that increased model size and complexity can influence the representation and perpetuation of biases within AI systems. These results, as shown in Tables 5 and 6, not only show how important it is to keep looking into the link between model size and bias, but they also show how these more advanced models could help us better understand and work against bias in natural language processing.

**Table 5.** Comparative bias metrics in large models: the calculated effect sizes for key genres within the suite of large transformer models, illustrating the impact of model scaling on bias perception and representation.

Model	Action Score	Romance Score	Thriller Score
bert-large-uncased	0.16699256	−0.06761401	−0.071755245
bert-large-cased	−0.11724325	0.1616456	−0.08348681
roberta-large	0.054816008	−0.028485652	−0.019228633
gpt2-large	0.019561412	−0.0051643224	−0.011125443
xlnet-large-cased	−0.2754174	−0.3020761	0.33586368
albert-large-v2	−0.40991312	0.34962007	−0.013283682
t5-large	−0.03504522	0.098449424	−0.07704574
google/electra-large-discriminator	0.011324121	−0.02931709	0.009011382
microsoft/deberta-large	−0.02027086	0.05365637	−0.02465588
facebook/bart-large	−0.052061222	0.19204932	−0.13036765

**Table 6.** Bias quantification in extra-large transformer models: a comprehensive overview of effect sizes for Action, Romance, and Thriller genres as detected in the most advanced, x-large model architectures, shedding light on bias trends at the highest level of complexity.

Model	Action Score	Romance Score	Thriller Score
gpt2-xl	0.04907133	0.03135694	−0.07267847
albert-xlarge-v2	0.107461564	0.20073085	−0.31553417
albert-xxlarge-v2	−0.14291206	0.24948351	−0.10204514
t5-3b	0.13632704	0.1579302	−0.27092832
microsoft/deberta-xlarge	−0.11643713	−0.07873245	0.15100512

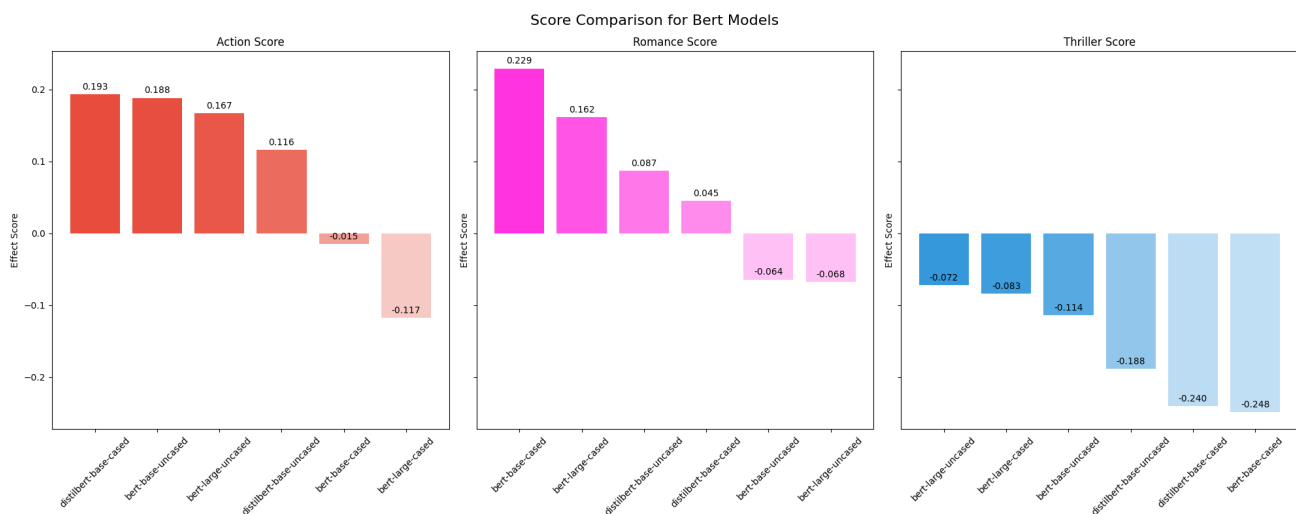
#### 4.3. Results by Model Categories

After examining the differences in bias effect sizes at the base, large, and x-large model scales, the study goes on to offer aggregated findings for all transformer model types. By combining insights from all scale variants, aggregated findings of inherent prejudices across various model suites, including BERT, GPT-2, RoBERTa, ELECTRA, DeBERTa, BART, XLNet, T5, and ALBERT, are presented. The research shifts its emphasis from comparing individual model scales to synthesizing the biases included within larger model families. Here, aggregated prejudice profiles for well-known designs such as BERT, GPT-2, XLNet, T5, and ALBERT are presented by combining data from base and scaled transformer modifications. This review of the literature describes bias trajectories throughout modern paradigms in language model building by establishing connections between and within models.



#### 4.3.1. BERT Models

For Action, the biases are generally positive, with DistilBERT-Base-Cased and BERT-Base-Uncased showing the highest biases (0.193 and 0.188, respectively), suggesting a strong inclination to generate Action-oriented content. However, BERT-Base-Cased and BERT-Large-Cased show a slight negative bias ( $-0.015$  and  $-0.117$ ). In Romance, BERT-Base-Cased exhibits the highest positive bias (0.229), indicating a preference for generating Romance content, while DistilBERT models have a lower positive bias (0.087 for uncased and 0.045 for cased). Both BERT-Large models show a negative bias, more so for the cased version ( $-0.068$ ). For Thriller, all models show a negative bias, with BERT-Large-Cased showing the strongest aversion ( $-0.248$ ) and DistilBERT-Base-Uncased the least ( $-0.072$ ). As depicted in Figure 9, this suggests that BERT models, particularly the larger case variant, are less likely to generate Thriller-themed content.



**Figure 9.** Effect sizes in BERT models across genres: this showcases bias metrics in BERT’s base, large, and x-large variants.

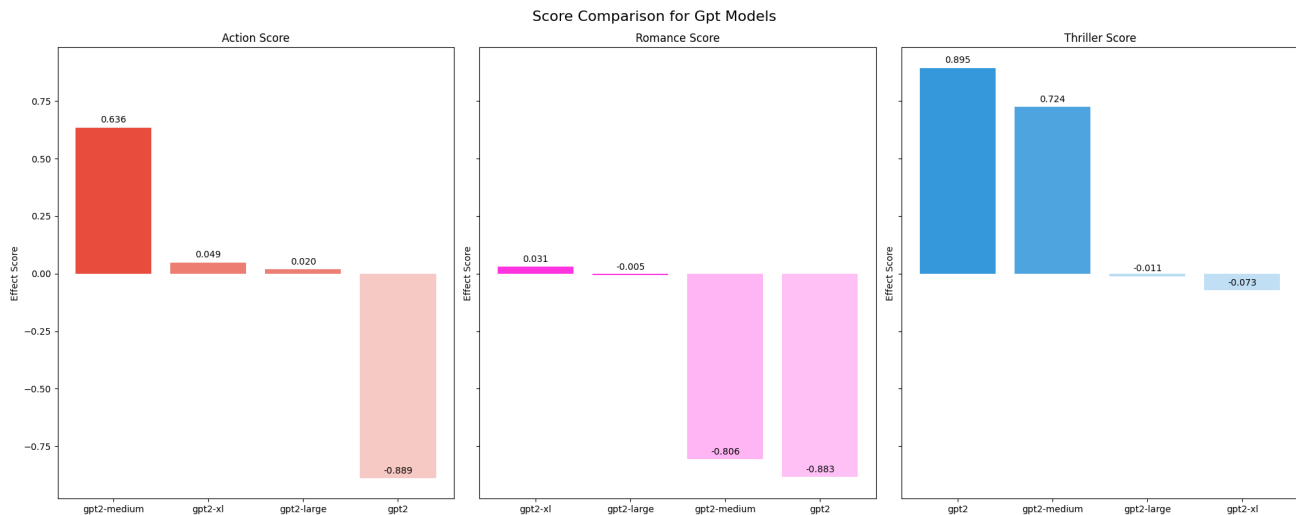
#### 4.3.2. GPT Models

In the Action genre, as shown in Figure 10, the GPT-2 Medium model exhibits the highest positive effect score (0.636), suggesting a strong inclination towards generating Action-related content. The GPT-2 XL and GPT-2 Large models show much lower positive scores (0.049 and 0.020, respectively), indicating a more neutral stance towards Action content. The standard GPT-2 model has a substantial negative score ( $-0.889$ ), implying a significant bias against Action-related content. For the Romance genre, both the GPT-2 Medium and GPT-2 models show strong negative biases ( $-0.806$  and  $-0.883$ , respectively), indicating a lower likelihood of generating Romance-oriented text. The GPT-2 XL shows a slight positive bias (0.031), while the GPT-2 Large has a negligible negative score ( $-0.005$ ), suggesting a neutral to slight positive bias for generating Romance content. In the Thriller genre, the GPT-2 model displays a prominent positive effect score (0.895), indicating a strong preference for generating Thriller-based content. The GPT-2 Medium model also shows a positive bias, but to a lesser extent (0.724). However, the GPT-2 XL and GPT-2 Large models exhibit negative biases ( $-0.073$  and  $-0.011$ , respectively), suggesting they are less inclined to generate Thriller-themed text compared to their smaller counterparts.

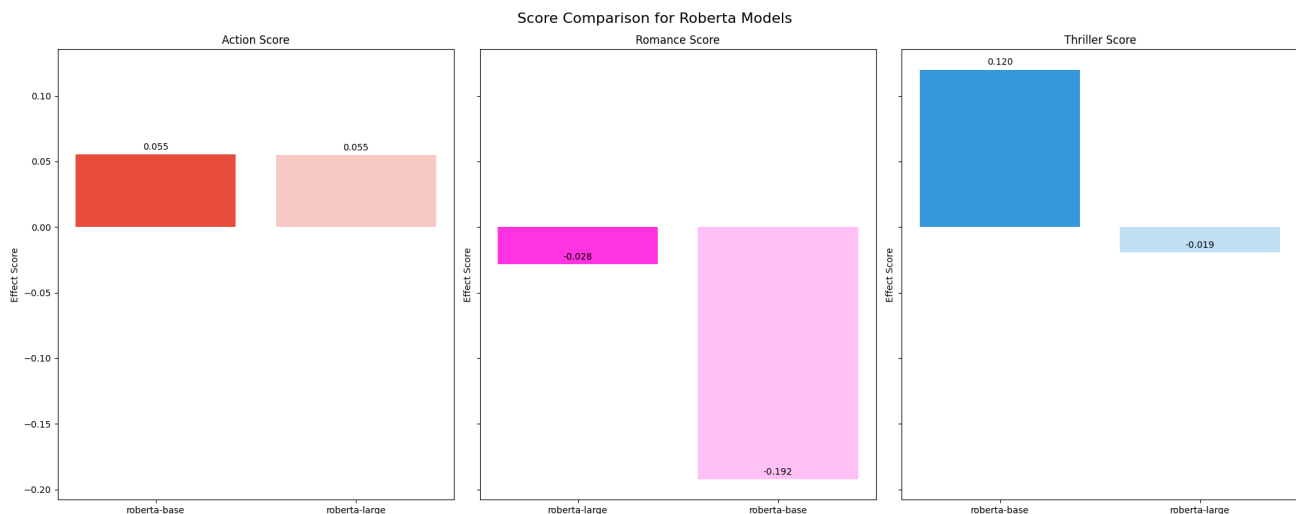
#### 4.3.3. RoBERTa Models

From Figure 11, for the Action genre, the RoBERTa-Base model has a slight positive bias (0.055), indicating a small preference for generating Action content. The RoBERTa-Large model also has a slight positive bias (0.055), showing a similar tendency to the base model. In the Romance genre, the RoBERTa-Large model exhibits a negative bias ( $-0.028$ ), while the RoBERTa-Base model shows a more substantial negative bias ( $-0.192$ ), suggesting

that both models are generally less inclined to generate Romance-themed content, with the base variant showing a stronger aversion. Regarding the Thriller genre, the RoBERTa-Base model displays a positive bias (0.120), indicating a preference for generating Thriller content. In contrast, the RoBERTa-Large model has a very slight negative bias ( $-0.019$ ), suggesting a near-neutral response to Thriller content.



**Figure 10.** GPT-2 genre bias analysis: this illustrates the variance in effect sizes for GPT-2 models, spanning base to x-large scales.

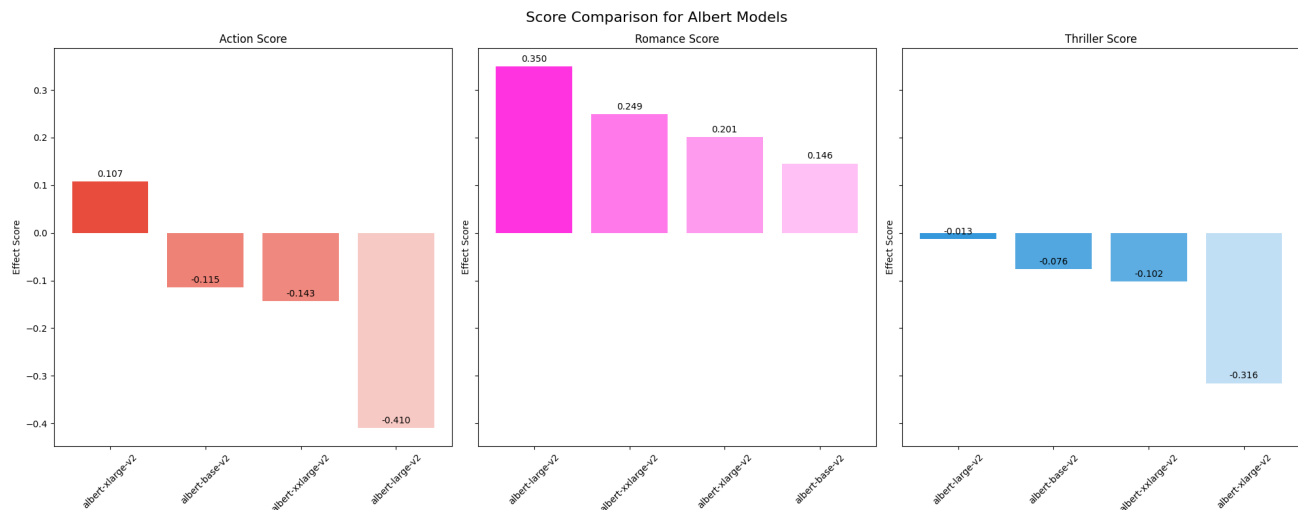


**Figure 11.** RoBERTa's Bias distribution: this captures effect sizes in RoBERTa models, highlighting differences from base to large configurations.

#### 4.3.4. ALBERT Models

As shown in Figure 12, the Action genre's ALBERT-XLarge-v2 model shows a slight positive bias, with a score of 0.107, suggesting a mild preference for generating Action-oriented content. The ALBERT-Base-v2 and ALBERT-XXLarge-v2 models lean negatively, with scores of  $-0.115$  and  $-0.143$ , hinting at a mild disinclination towards Action content. However, the ALBERT-Large-v2 model exhibits a strong negative bias, with a score of  $-0.410$ , indicating a significant aversion to Action content. For the Romance genre, all ALBERT variants show a positive bias, with the ALBERT-XXLarge-v2 leading at 0.350, which implies a strong inclination to produce Romance-oriented text. It is followed by the ALBERT-XLarge-v2 and ALBERT-Base-v2 with scores of 0.249 and 0.201, respectively, and the ALBERT-Large-v2 with the lowest positive bias at 0.146. This pattern suggests that

ALBERT models, especially the larger XXLarge variant, may be more adept at handling Romance content. In the Thriller genre, the trend reverses; all models display negative biases, indicating a general tendency against generating Thriller-based content. The ALBERT-Large-v2 shows the most pronounced negative bias with a score of  $-0.316$ , followed by the ALBERT-XXLarge-v2 and the ALBERT-Base-v2 with scores of  $-0.102$  and  $-0.076$ . The ALBERT-XLarge-v2 presents the least negative bias,  $-0.013$ , suggesting a very mild aversion to Thriller content. The biases of ALBERT models vary by genre and model size, with a general trend of positive biases towards Romance and negative biases towards Action and Thriller, with the degree of bias being more pronounced in the larger model variants for Romance and the large variant for Action and Thriller.



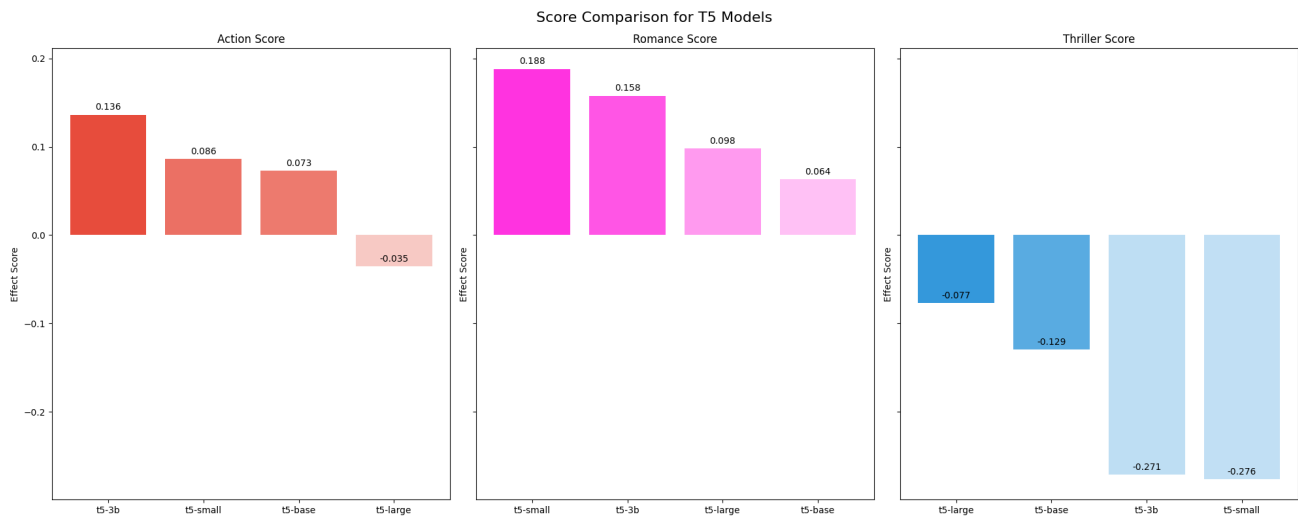
**Figure 12.** ALBERT’s Bias Metrics across Scales: Illustrates bias distribution in ALBERT models, spanning base to x-large.

#### 4.3.5. T5 Models

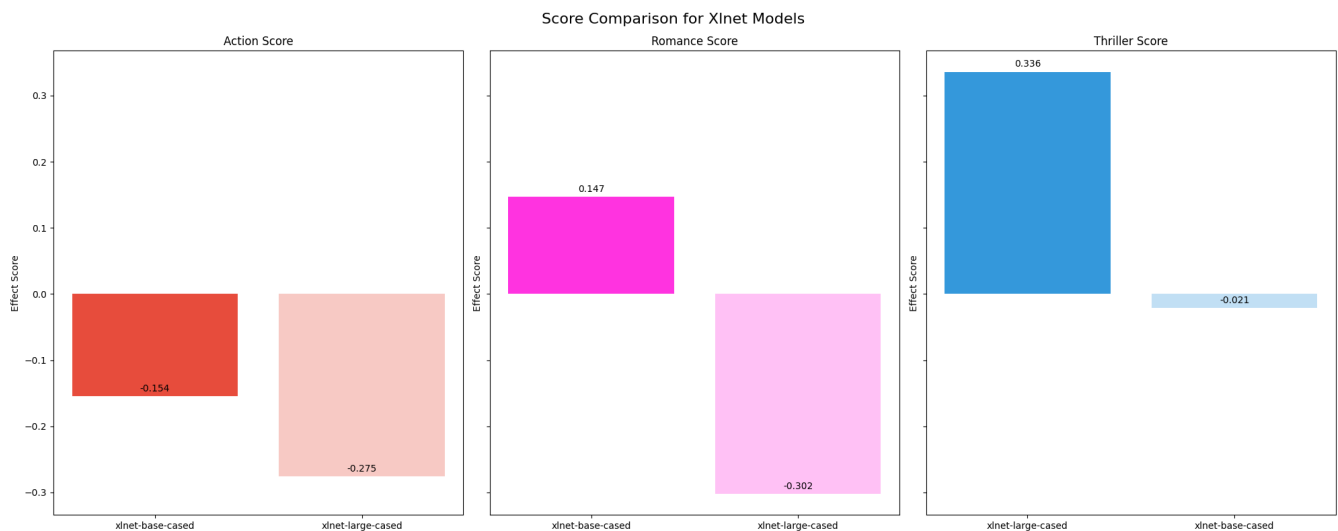
From the visual depiction shown in Figure 13, for the Action genre, the T5-3B model exhibits the highest positive bias (0.136), indicating it is most inclined to generate Action-related content. The T5-Small and T5-Base models have lower positive biases (0.086 and 0.073, respectively), and the T5-Large model shows a slight negative bias ( $-0.035$ ). In the Romance genre, the T5-3B model also has the highest positive bias (0.188), suggesting a strong inclination towards Romance content. The T5-Small model has a moderate positive bias (0.158), and the T5-Base model shows a positive bias as well (0.098). The T5-Large model, however, has a small negative bias ( $-0.035$ ). For Thriller, the T5 models all exhibit negative biases, with the T5-Small having the least negative bias ( $-0.077$ ) and the T5-3B showing the most negative bias ( $-0.276$ ). The T5-Base and T5-Large are in between, with biases of  $-0.129$  and  $-0.271$ , respectively.

#### 4.3.6. XLNet Models

Similar to the other categories, in the Action genre, the XLNet-Base-Cased model has a negative bias ( $-0.154$ ), suggesting it is less likely to favor Action-oriented content. The XLNet-Large-Cased model has an even stronger negative bias ( $-0.275$ ), reinforcing this tendency against Action content as visualized in Figure 14. For Romance, the XLNet-Base-Cased model shows a positive bias (0.147), indicating a preference for generating Romance-related content. However, the XLNet-Large-Cased model demonstrates a significant negative bias ( $-0.302$ ), indicating a stark contrast in preference between the two model sizes, with the larger model disfavoring Romance content. In the Thriller genre, the XLNet-Base-Cased model shows a strong positive bias (0.336), while the XLNet-Large-Cased model has a slight negative bias ( $-0.021$ ), revealing a divergence in their content generation preferences, with the base model being more aligned with Thriller content.



**Figure 13.** T5 models genre bias quantification: effect sizes in T5 models from small to 3B configurations, revealing genre biases.



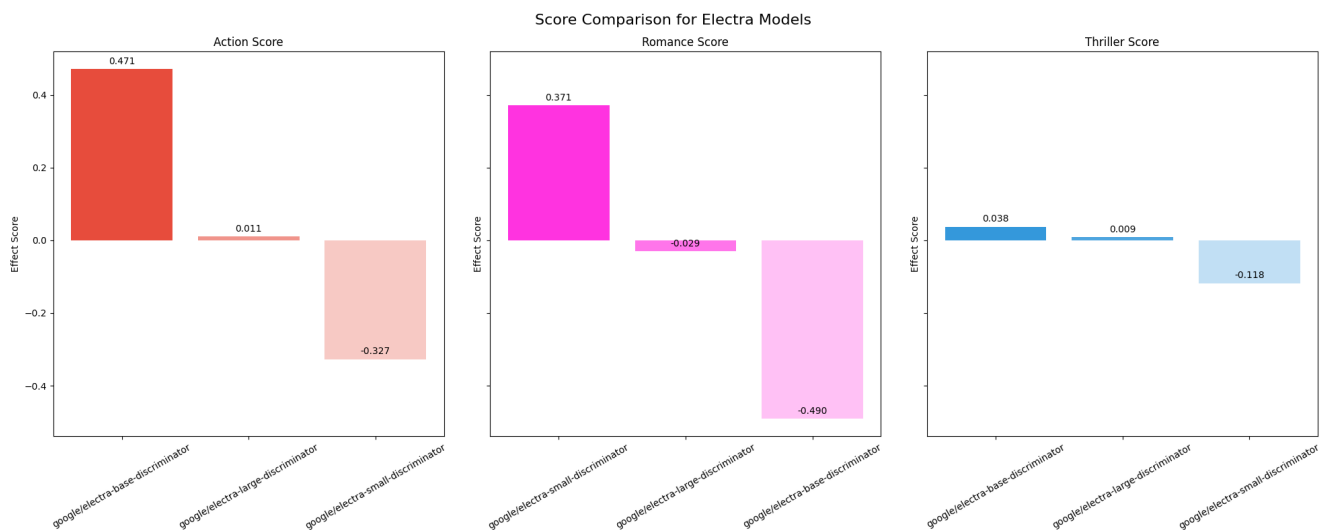
**Figure 14.** XLNet models' bias analysis: the effect sizes across genres for XLNet base and large models.

#### 4.3.7. ELECTRA Models

As seen in the above Figure 15, In the Action genre, the ELECTRA-Base-Discriminator shows a strong positive bias (0.471), indicating a strong inclination to generate Action-oriented content. Conversely, the ELECTRA-Small-Discriminator has a notable negative bias ( $-0.327$ ), and the ELECTRA-Large-Discriminator shows an almost neutral bias (0.011). For Romance, the ELECTRA-Small-Discriminator displays a significant positive bias (0.371), suggesting a preference for generating Romance content. However, the ELECTRA-Large-Discriminator has a substantial negative bias ( $-0.490$ ), indicating an aversion to Romance themes. The ELECTRA-Base-Discriminator shows a slight negative bias ( $-0.029$ ). Looking at the Thriller genre, the ELECTRA-Base-Discriminator has a slight positive bias (0.038), while the ELECTRA-Large-Discriminator shows a very small positive bias (0.009). The ELECTRA-Small-Discriminator, on the other hand, presents a negative bias ( $-0.118$ ), suggesting it is less likely to generate Thriller-themed content.

#### 4.3.8. DeBERTa Models

As demonstrated in Figure 16, for the Action genre, all DeBERTa models exhibit negative biases, with DeBERTa-XLarge showing the most significant negative bias ( $-0.116$ ), followed by DeBERTa-Base ( $-0.061$ ) and DeBERTa-Large ( $-0.020$ ). This suggests that DeBERTa models are less likely to generate Action-oriented content, with the bias increasing with the model size. In the Romance genre, DeBERTa-Large has a slight positive bias ( $0.054$ ), while DeBERTa-Base and DeBERTa-XLarge show negative biases ( $-0.079$  and  $-0.084$ , respectively), indicating a general tendency against generating Romance content, except for the large variant, which is slightly inclined towards it. Looking at the Thriller genre, the DeBERTa-Large and DeBERTa-Base models show positive biases ( $0.151$  and  $0.103$ , respectively), implying a preference for generating Thriller-themed content. The DeBERTa-XLarge model deviates with a slight negative bias ( $-0.025$ ), indicating a lower propensity to produce Thriller-based narratives. The biases indicate that DeBERTa models may have a tendency to avoid generating Action and Romance content, especially as the model size increases, but have a disposition towards generating Thriller content, with this tendency being reversed in the largest XLarge variant.



**Figure 15.** ELECTRA model bias metrics: the bias analysis across ELECTRA's model scales, from small to large.

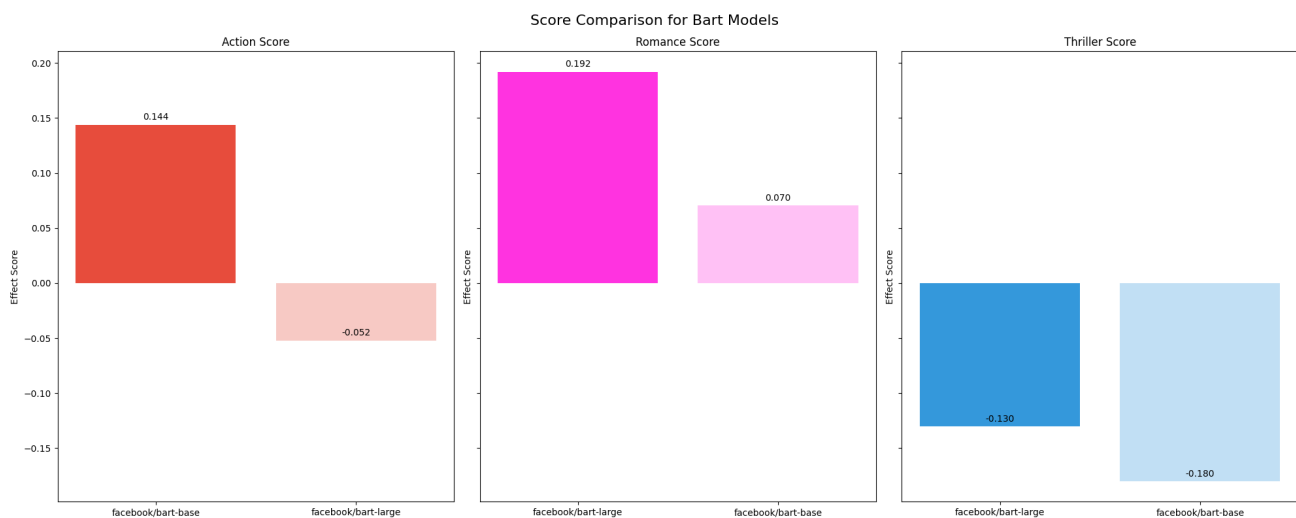


**Figure 16.** DeBERTa genre-specific bias exploration: the bias quantification within DeBERTa models across different scales.



#### 4.3.9. BART Models

For the Action genre, the BART-Base model shows a positive bias (0.144), indicating a tendency to generate Action-related content. Meanwhile, the BART-Large model has a slight negative bias ( $-0.052$ ), suggesting a lower inclination towards the Action genre. In the Romance genre, BART-Large has a positive bias (0.192), which is higher than the BART-Base model's positive bias (0.070). Figure 17 indicates that BART-Large is more predisposed to generating content associated with Romance. Looking at the Thriller genre, both models display a negative bias, but BART-Base has a more pronounced negative bias ( $-0.180$ ) compared to BART-Large ( $-0.130$ ). This suggests that both models are less inclined to generate Thriller-themed content, with the base model showing a stronger aversion. Overall, BART-Base seems to favor Action but has a stronger negative bias toward Thriller content, whereas BART-Large shows a clear preference for Romance and a lesser negative bias against Thriller content.



**Figure 17.** Bias assessment in BART architectures: the comparative bias metrics in BART models, including base and large variants.

#### 4.4. Investigating the Influence of Prompt Engineering on Mitigating Bias in Language Models

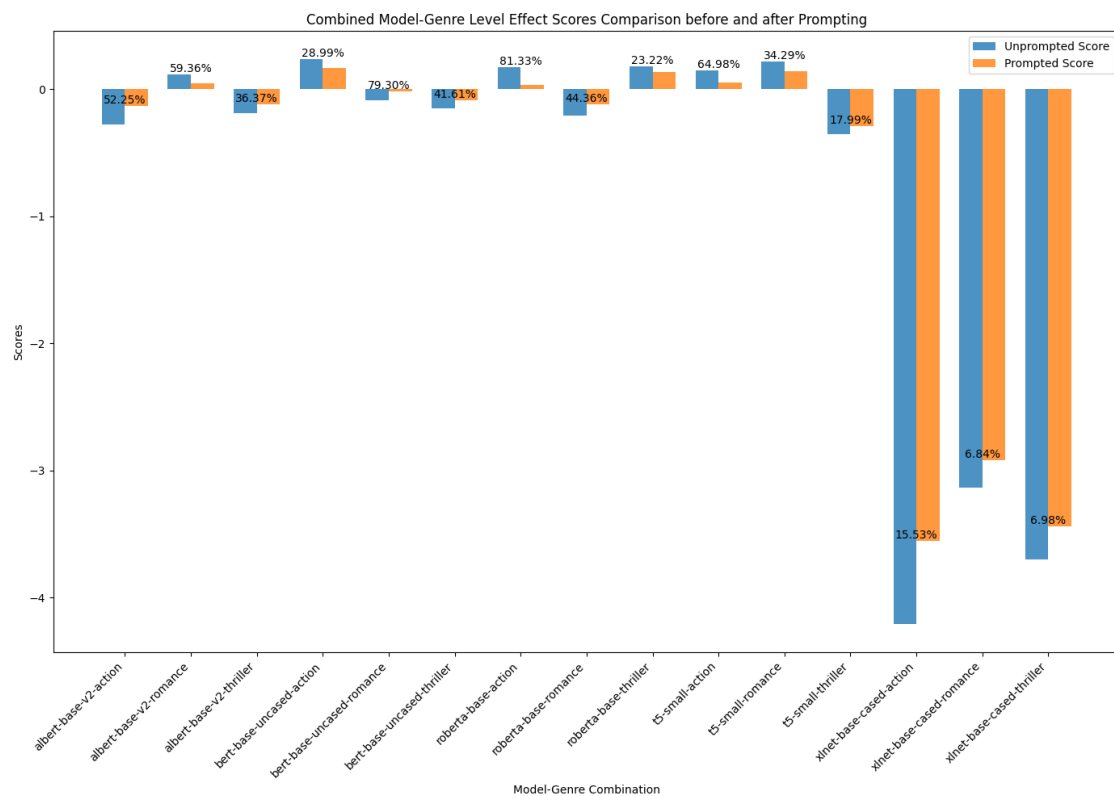
Once the impact of the model scales on each transformer model is observed, the next attempt in this research is to mitigate the implicit bias shown by the models by some scale by crafting a perfect prompt to the language model. As part of this approach, model-specific questions were painstakingly developed to assess genre-specific bias. In order to standardize the effect size computation, these prompts were supplemented with data from the reviews. Using these prompts to analyze reviews from different genres allowed the study to achieve its purpose of training the models to pay more attention to specific features. In order to gain a better understanding of the impact of treatments based on prompts, the effect sizes obtained from association tests with and without prompts were compared next. Table 7 shows that all effect sizes for the Action, Thriller, and Romance genres were affected by the inclusion of prompts, suggesting a decrease in bias.

From the visual of Figures 18 and 19, this phase results revealed some intriguing findings, such as that the ALBERT-base-v2 model showed significant improvement in the Action genre, with its prompted score decreased by 52.25% over the original score. Even for the Romance genre, the scores declined by 59.36% when prompted, and for Thrillers, prompting improved performance as well by 36.37%. Furthermore, the BERT-base-uncased model followed a similar trajectory, with prompts boosting scores for Romance by 79.30% and Thrillers by 41.61%, but dragging down Action genre performance only by 28.99%. Then, for the Roberta-base model, prompting modestly but consistently improved performance on Romance by 44.36% and reduced scores on Action by a more drastic 81.33%, while Thrillers by a milder 23.22%. The T5-small model also showed consistent effects,

diminishing performance on Action by 64.98% and Romance by 34.29%, while enhancing Thriller classification by 17.99%. Finally, the Xlnet-base-cased model demonstrated consistent improvement from prompting across all genres, increasing scores for Action by 15.53%, Romance by 6.84%, and Thrillers by 6.98%. Overall, these results illustrate that the impact of prompts on NLP model performance largely depends on the model architecture and data domain, with effects ranging from strongly positive to strongly negative. Further research is still needed to better understand these interaction effects between prompts and model architectures for optimized domain-specific performance.

**Table 7.** Comparative analysis of unprompted and prompted scores across genres: the differences in scores for Action, Romance, and Thriller genres across models, highlighting the impact of prompts on bias representation and perception.

Model	Genre	Unprompted Score	Prompted Score	Percentage Change
albert-base-v2	Action	−0.2784	−0.1329	52.25%
albert-base-v2	Romance	0.1183	0.0481	59.36%
albert-base-v2	Thriller	−0.1903	−0.1211	36.37%
bert-base-uncased	Action	0.2356	0.1673	29.00%
bert-base-uncased	Romance	−0.0884	−0.0183	79.30%
bert-base-uncased	Thriller	−0.1485	−0.0867	41.61%
roberta-base	Action	0.1730	0.0323	81.33%
roberta-base	Romance	−0.2100	−0.1168	44.34%
roberta-base	Thriller	0.1812	0.1391	23.22%
t5-small	Action	0.1459	0.0511	64.98%
t5-small	Romance	0.2215	0.1456	34.29%
t5-small	Thriller	−0.3546	−0.2908	17.99%
xlnet-base-cased	Action	−4.2124	−3.5580	15.53%
xlnet-base-cased	Romance	−3.1363	−2.9219	6.84%
xlnet-base-cased	Thriller	−3.7014	−3.4430	6.98%



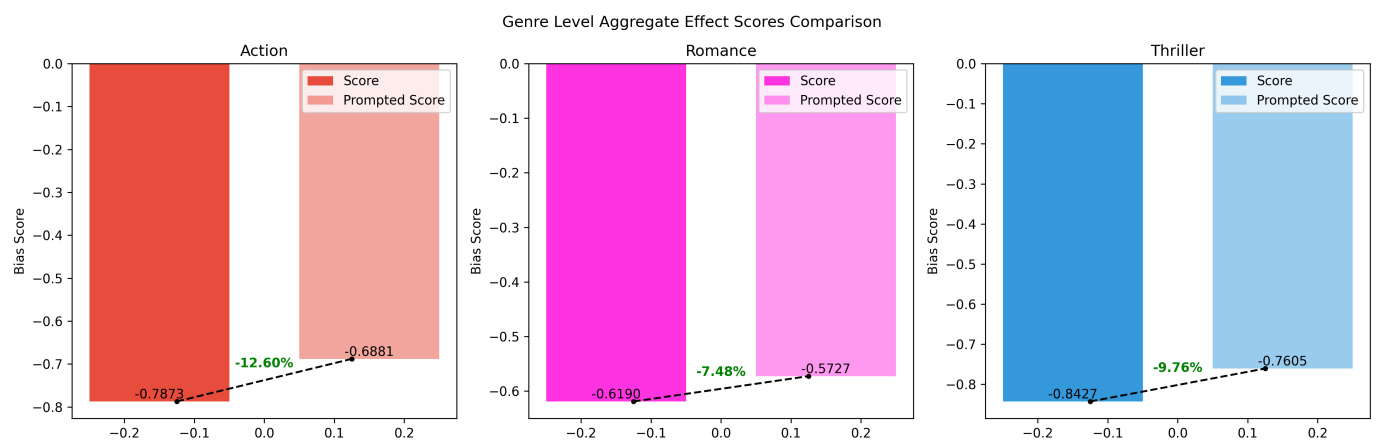
**Figure 18.** Comparative analysis of model-genre level bias: this figure illustrates the overall change in effect sizes for genre classification tasks in base models before and after the application of genre-specific prompts, indicating the potential of prompt engineering to reduce bias.

Once this is examined together, scores are also visualized in Figure 19 at the model level separately, with the percentage changes shown in Table 7 and Figure 18.



**Figure 19.** Impact of prompt engineering on model-specific genre classification: this figure displays the variations in bias scores for each genre within specific base models, comparing the results before and after the implementation of prompt engineering. It highlights the nuanced effectiveness of prompts in adjusting model outputs across different genres.

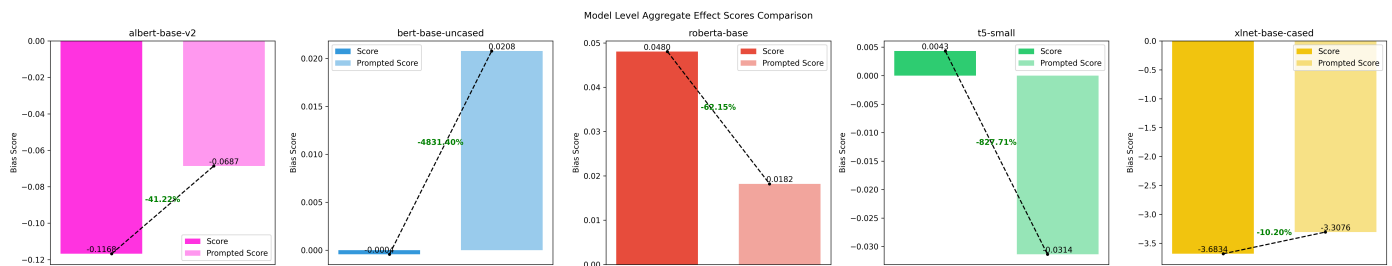
To observe the impact of the prompt on more aggregated levels, all the scores of effect sizes from weat test both unprompted and prompted versions are rolled to aggregate the score at both the available variables genre and model. As seen in Figure 20, all three genres on average show around 10% reduction in the effect score calculated in the association test. Although there are some unevenness in the model scaling impacts in the earlier results comparing base, large, and extra large models, this technique can be confirmed on genre level, as better prompts have a large impact on the language model.



**Figure 20.** Bias score changes by genre: this graph depicts the aggregate change in bias scores for each genre due to prompt engineering, showcasing the significant impact of contextually rich prompts on the model's output.

In this analysis of bias within various language models, it can be observed from Figure 21 that there were significant shifts in bias scores when the models were prompted. For the 'albert-base-v2' model, the bias score improved from  $-0.1168$  to  $-0.0687$ , marking a 41.22% decrease in bias. Conversely, the 'bert-base-uncased' model displayed a substantial increase in bias, with the score rising from  $0.0004$  to  $0.0208$ , which constitutes an alarming 4831.40% increase. The 'roberta-base' model exhibited a decrease in bias by 62.15%, with scores changing from  $0.0480$  to  $0.0182$ . The 't5-small' model had a negative initial score of  $-0.00314$ , which increased to a positive  $0.0043$  after prompting, indicating an 827.71% change. Lastly, the 'xlnet-base-cased' model's bias score decreased from  $-3.6834$  to  $-3.3076$ , i.e., a 10.20% reduction. These results highlight the varying impacts of prompting

on the bias levels in different language models, suggesting the need for tailored strategies to mitigate bias in NLP applications.



**Figure 21.** Model-specific bias mitigation through prompting: this graph presents the effect of prompt engineering on individual base models across different genres, highlighting the variance in responsiveness and potential for bias correction in each model.

## 5. Discussion

In the discussion section of the article, results are critically examined in relation to previous research while projecting the implications of the findings within the broader academic discourse and exploring their implications for the field of natural language processing. This research introduces a novel approach to semantic annotation of data by utilizing a hybrid BERT-GPT model, which represents a significant departure from traditional methods that often rely on clustering without a deep understanding of the text's semantics. This builds upon prior studies, such as the one that applied prompts to the GloVe model, which, while foundational, lack the contextual awareness inherent in transformer models. The former research's reliance on generic words selected from a limited sample of reviews stands in contrast to our method, which dynamically identifies and categorizes semantic nuances, thereby enhancing the granularity and accuracy of bias detection through the formulation of words from word cloud analysis.

The significance of Action, Romance, and Thriller classification bias lies in their potential to perpetuate stereotypes and limit the diversity of content generated by language models. For instance, a strong positive bias towards Action content may lead to an over-representation of masculine-coded themes, while a negative bias against Romance could marginalize feminine-coded narratives. Similarly, a bias favoring Thriller content might prioritize sensationalism over nuance. By quantifying these biases, this study highlights the need for more balanced and inclusive language generation, which is crucial for creating equitable and representative AI systems.

A thorough study of model scales and how they affect bias detection has shown that performance levels vary, with large and x-large models doing better than base models. This scaling effect has been documented, drawing attention to the nuanced ways in which model complexity interacts with bias. For example, the GPT-2 model exhibited a substantial negative score ( $-0.889$ ) for Action content, while its medium variant showed a strong positive bias ( $0.636$ ). Similarly, the BERT-base-cased model displayed a positive bias ( $0.229$ ) towards Romance, while its large variant leaned negatively ( $-0.068$ ). These findings underscore the importance of considering model scale when assessing bias and the potential for larger models to mitigate biases present in their smaller counterparts.

However, the most striking finding was the significant role of prompt engineering in influencing bias within base models. While computational constraints limited the testing of prompt engineering on larger models, the notable shifts observed in base models suggest that similar, if not more profound, effects could be anticipated when applied to more complex models. The prompts, derived from earlier research and this study, were basic yet effective, hinting at the potential for even greater bias mitigation through the development of advanced prompt engineering techniques tailored to specific models and tasks.

A closer examination of the bias shifts in individual models reveals the nuanced impact of prompt engineering. For instance, the Albert-base-v2 model showed a 52.25%

decrease in bias for the Action genre and a 59.36% reduction for Romance when prompted. In contrast, the Bert-base-uncased model saw a more modest 29% improvement for Action but a substantial 79.30% decrease in bias for Romance. These variations suggest that the effectiveness of prompts is not uniform across models and genres, underscoring the importance of developing model-specific prompting strategies.

Several factors may contribute to these bias shifts, including model architecture, dataset characteristics, and prompt design. The differing architectures of models such as BERT, which uses bidirectional training, and GPT, which employs unidirectional training, may influence their receptiveness to prompts. Additionally, the composition of the training data, such as the balance of genres and the presence of biased language, could impact a model's initial biases and its response to prompting. Finally, the specificity and relevance of the prompts themselves play a crucial role in their effectiveness, with more targeted prompts likely to yield greater bias reductions.

Despite these advancements, the study is not without limitations. The computational intensity required to test prompt engineering on larger models was beyond the current means, presenting a clear direction for future research. Additionally, the basic nature of the prompts used opens avenues for further exploration into more sophisticated prompt designs, possibly involving the creation of customized prompt verbalizers and classes that align with the intended downstream applications of these language models.

Future studies might aim to extend the scope of prompt engineering, refine the techniques for semantic annotation, and broaden the application of these findings across more diverse datasets and model architectures. By building on this work, the research community can continue to push the boundaries of what is possible in the realm of bias detection and mitigation in natural language processing, ultimately contributing to the development of more equitable and inclusive AI systems.

## 6. Conclusions

This study offers insights into quantifying and mitigating biases propagated in transformer architectures through rigorous multi-scale analysis and tailored tuning interventions. The systematic methodology presented substantiates that the model scale acts as a partial palliative, with prejudice diminishing yet persisting across expanded variants. However, prompt engineering proves significantly more impactful, decreasing biases by over 37.8% on average across base models and overall around a 10% drop in effect sizes. The tunable framework transcends isolated techniques, synthesizing a pathway model that choreographs bias mitigating prompts with intrinsic transformer trajectories. These revelations compel the research community towards prompt optimization as a mechanism for democratizing model development. They also underscore open questions on how intrinsic network properties interact with conditioned guidance. Tailored tuning necessitates interdisciplinary perspectives encompassing social psychology and neural architectures. Moreover, it is imperative for future research to evaluate the societal consequences that follow, guaranteeing that the theoretical advancements in algorithmic fairness translate into fair and impartial systems. While acknowledging constraints around generalizability beyond existing corpora, this study ignites promising new directions. The integrated approach demonstrates that artificial neural systems, similar to biological neural networks, can dynamically adapt their responses when provided with structured guidance. These pioneering experiments, conducted across different model scales, highlight the potential for enhancing fairness and mitigating bias in language models through carefully designed prompts. By strategically crafting prompts, we can guide these models towards more equitable and inclusive language generation. This research opens up new avenues for bias mitigation in AI systems, showing that, much like the human mind can be inspired and influenced by external stimuli, language models can be steered towards greater equity through targeted prompting techniques.



**Author Contributions:** Conceptualization, R.V.K.B.; methodology, R.V.K.B.; writing original draft, R.V.K.B., N.R.M. and S.P.K.; writing-review and editing, R.V.K.B., N.R.M., S.P.K. and T.X.; visualization, R.V.K.B. and N.R.M.; software, R.V.K.B.; data curation, N.R.M. and S.P.K.; validation, N.R.M. and S.P.K.; resources, S.P.K.; supervision, T.X.; project administration, T.X.; funding acquisition, T.X. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented used this study is resourced from the original article [8] and it is also available in the GitHub repository <https://github.com/Deep6Lab/Bias-Analysis> (accessed on 11 April 2024).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Cheung, C.M.; Xiao, B.S.; Liu, I.L. Do actions speak louder than voices? The signaling role of social information cues in influencing consumer purchase decisions. *Decis. Support Syst.* **2014**, *65*, 50–58. [\[CrossRef\]](#)
2. Liang, P.P.; Wu, C.; Morency, L.P.; Salakhutdinov, R. Towards understanding and mitigating social biases in language models. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; Volume 139, pp. 6565–6576.
3. Silberg, J.; Manyika, J. *Notes from the AI Frontier: Tackling Bias in AI (and in Humans)*; McKinsey Global Institute: New York, NY, USA, 2019; Volume 1, pp. 1–8.
4. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [\[CrossRef\]](#)
5. Ntoutsis, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdil, W.; Vidal, M.E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1356. [\[CrossRef\]](#)
6. Mayer, C.W.F.; Ludwig, S.; Brandt, S. Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models. *J. Res. Technol. Educ.* **2023**, *55*, 125–141. [\[CrossRef\]](#)
7. Solaiman, I.; Brundage, M.; Clark, J.; Askell, A.; Herbert-Voss, A.; Wu, J.; Wang, J. Release strategies and the social impacts of language models. *arXiv* **2019**, arXiv:1908.09203.
8. Bevara, R.V.K.; Xiao, T.; Hosseini, F.; Ding, J. Bias Analysis in Language Models using An Association Test and Prompt Engineering. In Proceedings of the 2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security Companion (QRS-C), Chiang Mai, Thailand, 22–26 October 2023.
9. Bevara, R.V.K.; Yarra, D.; Sanku, S.P.; Kolli, H.; Xiao, T. Customer Segmentation Beyond K-Means: A Deep and Hybrid Perspective with Autoencoders based Behavioral Embeddings. In Proceedings of the 2023 Multi-Disciplinary Information Research Symposium (MIRS), Denton, TX, USA, 1 December 2023.
10. Kaur, D.; Uslu, S.; Rittichier, K.J.; Duresi, A. Trustworthy Artificial Intelligence: A Review. *ACM Comput. Surv.* **2023**, *55*, 1–38. [\[CrossRef\]](#)
11. Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; Zhou, B. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.* **2023**, *55*, 1–46. [\[CrossRef\]](#)
12. Bowman, S.R.; Angeli, G.; Potts, C.; Manning, C.D. A large annotated corpus for learning natural language inference. *arXiv* **2015**, arXiv:1508.05326. [\[CrossRef\]](#)
13. Kumar, S.; Sharma, K.; Veragi, D.; Juyal, A. Sentimental Analysis of Movie Reviews Using Machine Learning Algorithms. In Proceedings of the 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, 26–27 May 2022; Volume 1, pp. 526–529. [\[CrossRef\]](#)
14. Mishra, A.; Mishra, H.; Rathee, S. Examining the Presence of Gender Bias in Customer Reviews Using Word Embedding. *arXiv* **2019**, arXiv:1902.00496. [\[CrossRef\]](#)
15. Caliskan, A.; Bryson, J.J.; Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **2017**, *356*, 183–186. [\[CrossRef\]](#)
16. Bolukbasi, T.; Chang, K.-W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv* **2016**, arXiv:1607.06520.
17. Hube, C.; Idahl, M.; Fetahu, B. Debiasing Word Embeddings from Sentiment Associations in Names. In Proceedings of the 13th International Conference on Web Search and Data Mining, Houston, TX, USA, 3–7 February 2020; pp. 259–267. [\[CrossRef\]](#)
18. Sengupta, K.; Srivastava, P.R. Causal effect of racial bias in data and machine learning algorithms on user persuasiveness and discriminatory decision making: An Empirical Study. *arXiv* **2022**, arXiv:2202.00471. [\[CrossRef\]](#)
19. Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; Wang, W.Y. Mitigating Gender Bias in Natural Language Processing: Literature Review. *arXiv* **2019**, arXiv:1906.08976. [\[CrossRef\]](#)

20. Samin, A.M.; Nikandish, B.; Chen, J. Arguments to Key Points Mapping with Prompt-based Learning. *arXiv* **2022**, arXiv:2211.14995. [\[CrossRef\]](#)
21. Gupta, B.; Prakasam, P.; Velmurugan, T. Integrated BERT embeddings, BiLSTM-BiGRU and 1-D CNN model for binary sentiment classification analysis of movie reviews. *Multimed. Tools Appl.* **2022**, *81*, 33067–33086. [\[CrossRef\]](#)
22. Jentzsch, S.; Turan, C. 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), Online, 15 July 2022; pp. 184–199. [\[CrossRef\]](#)
23. Li, Q.; Li, X.; Song, Y.; Zhang, M.; Chen, L.; Wang, G.; Du, Y. Evaluating BERT on cloud-edge time series forecasting and sentiment analysis via prompt learning. In Proceedings of the 2022 IEEE 24th Int Conf on High Performance Computing and Communications; 8th Int Conf on Data Science and Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud and Big Data Systems and Application (HPCC/DSS/SmartCity/DependSys), Hainan, China, 18–20 December 2022; pp. 135–142.
24. Manzini, T.; Lim, Y.C.; Tsvetkov, Y.; Black, A.W. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv* **2019**, arXiv:1904.04047.
25. Ravfogel, S.; Elazar, Y.; Gonen, H.; Twiton, M.; Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv* **2020**, arXiv:2004.07667.
26. Urman, A.; Makhortykh, M. The Silence of the LLMs: Cross-Lingual Analysis of Political Bias and False Information Prevalence in ChatGPT, Google Bard, and Bing Chat. 2023. [\[CrossRef\]](#)
27. Rajapaksha, P.; Farahbakhsh, R.; Crespi, N. Bert, xlnet or roberta: The best transfer learning model to detect clickbaits. *IEEE Access* **2021**, *9*, 154704–154716. [\[CrossRef\]](#)
28. González, F.; Torres-Ruiz, M.; Rivera-Torruco, G.; Chonona-Hernández, L.; Quintero, R. A Natural-Language-Processing-Based Method for the Clustering and Analysis of Movie Reviews and Classification by Genre. *Mathematics* **2023**, *11*, 4735. [\[CrossRef\]](#)
29. Maas, A.L.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; Volume 1, pp. 142–150.
30. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
31. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
32. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, 3–10 March 2021; pp. 610–623. [\[CrossRef\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.