LUISS

Degree in Marketing Analytics & Metrics

Business and Marketing Analytics

The Double Edge of Generative AI: Identifying and Addressing Gender Bias through LLMs

Prof. Adrea De Mauro

SUPERVISOR

Prof. Luigi Monsurro

CO-SUPERVISOR

Valentina Ferreri - 775731

CANDIDATE

Accedemic Year 2024/2025

Index

1. Introduction

Advances in Generative Artificial Intelligence (GenAI) have generated new and far-reaching applications across an expanding array of domains, including, but not limited to, the production of natural language and images, as well as automated decision-making systems. As AI systems, including GenAI, continue to develop rapidly, providing unexpected opportunities for innovation and efficiency, emergent and potentially urgent ethical and social issues arise. One notable ethical issue is the ability of AI systems to reproduce gender stereotypes encoded in the text and images they generate, which ultimately has the potential to reiterate existing norms, exacerbate existing inequalities, and undermine contemporary efforts to cultivate diversity and inclusion in technology and society.

Achieving fairness in AI is a multidisciplinary challenge that must include computer scientists, ethicists, legal scholars, and social scientists (Ferrara et al. 2023). Collaborative efforts from these fields can help design not only technically robust systems but also systems that are socially responsible, transparent, and fair. However, the road to reveal unbiased AI is still being developed. While there have been notable advances in areas like dataset audits, prompt engineering, and fairness metrics, in general, our current approaches are historic and, again, incomplete. This indicates a need for renewed focus on research, particularly as it relates to the detection and mitigation of bias in generative systems.

Representation of gender within STEM fields is also indicative of trends in technology as a whole, women remain underrepresented in both participation and visibility. However, recent bibliometric analyses are starting to show that women are increasingly represented as lead authors in highly cited research. Therefore, we should view this as a slow but notable move towards equity in academic contribution and as researchers, we must ensure that the tools and technologies developed are consistent with the values of inclusion.

This thesis supports both of these issues. The first aim of this thesis is to examine how bias manifests and exists in our society, while the second aim is to design and evaluate a functional framework to detect and mitigate bias in text. With the use of prompt-engineered LLMs and a user-friendly interface, this research is intended to

contribute to the responsible development of AI by developing more than critiques, but functional tools for promoting equitable communication in digital spaces.

This work of research makes both a theoretical and practical contribution. In addition to reviewing the literature and identifying ethical concerns regarding gender bias in AI systems, this work proposes and validates a methodology based on prompt-engineered Large Language Models to detect such bias. Furthermore, it introduces a proof-of-concept mobile application to support individuals in real-time bias detection and inclusive language reformulation. This dual contribution underscores the paradox of GenAI: while it can replicate societal bias, it can also be a powerful ally in identifying and mitigating it.

1.1 Research questions

To guide the research, the following two questions were formulated:

RQ1: To what extent can Large Language Models accurately detect human gender bias in short textual content?

RQ2: Can a mobile application powered by LLMs effectively support users in recognizing and mitigating biased language in real time?

2. Literature Review

2.1 Defining bias

Bias refers to a consistent or expected tendency or inclination that inhibits impartial judgment, and this takes many forms: cognitive bias, social bias, and statistical bias. Lovallo and Sibony (2010) consider bias as more than just mistakes; they see bias as a natural shortcut for thinking that is sometimes beneficial for rapid decision making but often produces defective judgments in complex situations. For example, cognitive bias represents systematic departures from reason, motivating people to engage in irrational judgments based on their beliefs and experiences that they have available to them. Social bias occurs when people act preferentially towards some groups over others, involving

stereotypes and social norms out of a natural or learned inclination. In research, it involves a process of bias that leads to errors in data collection and analysis, producing imbalanced conclusions (Ferrara, 2023).

Bias has serious ramifications in many domains, including decision-making, research results, and social perceptions. In AI and machine learning, for example, bias happens when data are biased or incomplete, so that training results in models that reinforce social discrimination (Ferrara, 2023). Gender bias is especially problematic, and examples abound, including Li and Bamman's (2023) research on "Gender and Representation Bias in GPT-3 Generated Stories." They found numerous ways that language models supported stereotypes of women by, for example, placing them in roles involving family and looks, while giving power to male characters.

2.2 Gender Bias

Gender bias is a social structure in various ways, bound in institutional structures, cultural norms and practices, social norms, and relationships in everyday life. To better understand the ubiquitous nature of pre-existing gender bias present in society, we can refer to the Gender Social Norms Index (GSNI). The GSNI found that approximately 90% of the population around the world has bias against women, and 85% of the population that experiences it (Gender Bias Report, 2023).

As we see some strong domestic implications from the findings, we further find that two out of five people in the world believe men make better business managers than women, and a majority in every country believe men make better political leaders. While there can be some concern that we have made progress against discrimination, gender parity was at best only incomplete and alternatively, there remained challenges in a growing number of sectors in the economy and its recent developments in technology. Economists, writers, and gender referentials identify five possible sets of explanations for these common systematic distinctions: stereotypes/cultural constraints, women's status in leadership/workforce, the changing labor markets, negative educational impacts, and political rationalizations. It illustrated that broader community acceptance and a pre-registration of opportunities based on gender bias, or the sexist and marginalising present towards equity and resourcing for women, strongly correlate with a lack of practical and

educational comprehension, acceptance, and wider implementation for gender-pursuant protection progress and equality law. These explanations reflect how generic, unfinished, and gender absolutes can construct an opposite climate in the form of the absolute absence of women's vertical ascendency in their occupational track or career, which produces biased recruitment and promotion incongruently towards men in many situations. Also more closely resembles an inherited gender bias of the "think-manager-think-male" ideology that clutches to masculine attributes as precursory role modeling of leadership, affecting women's chances of obtaining what is due to them in their affiliated workplace (ILO ACT/EMP, 2017).

Another significant cause of gender bias is educational inequality. According to Rao and Sweetman (2014), good education to fit girls and women is particularly central to shattering the ice of discrimination and for women to reclaim the space of equal rights at home, in the community, and in society. The issue is systematic discrimination, which is that, from the day a baby girl is born until she dies, the effects of discrimination prevail. For example, being a girl child in a patriarchal society, such as India, is entirely negative; as Rahaman and Mazumder (2020) say, there are great limitations on education, employment, and health effects, as biased gender norms are interwoven into the fabric of daddy state, making it marginally harder to gain equal space to access these services. The economic loss arising from gender issues can never be overemphasized. Bilan et al. (2020) say gender discrimination is often associated with age discrimination in the labor market, and out of 57.1% of participants, 71.4% were younger than 35. Further to this, Bilan et al. (2020), say married women who are subject to detrimental gender discrimination in the labor market will suffer unequal pay and will also be prone to lower career trajectories, including inequality of opportunity arising from lack of benefits and pay. There's no potential surprise that companies perpetuating inequality will ultimately experience heightened, and sometimes fatal, turnover, in some cases with quitting employees amounting to as much as 71% turnover, which alone is definitive proof of women's disappointment with their potential due to inequalities in the workplace. Employers experience not only morale costs as well as financial costs, and there is potential for administrative costs, such as rising recruitment costs as well as the bitter reality of lower productivity.

Recent research illustrates the substantial economic costs of gender inequality (Fry and Aragao, 2025). The long-term gender pay gap is prevalent in every state and notably in numerous industries, where women are routinely paid less than men for performing exactly the same work. Of course, the pay gap reduces women's economic viability, but it ultimately has an impact on the development of the economy. Fry and Aragao (2025), argue that closing the gap between the genders in labor markets would generate unprecedented increases in GDP, in developing and emerging economies, which would facilitate potential economic growth. The economic costs of gender inequality are vast, and potentially trillions of dollars are wasted each year because of women's underperforming income (Khattar 2024).

In developing economies, for example, gender equality could cost multiple trillions of dollars each year, and the enormous costs of variance are illuminated (UNCTAD, 2023).

Furthermore, gender discrimination among informal businesses in developing economies has been found to be responsible for large gaps in labour productivity, representing lost potential output of the economy. On the other hand, promoting gender equality is associated with productivity and economic growth, notably in gender equality linked to skills in science, technology, engineering, and mathematics (STEM) industries (ILO, 2023).

Despite advances in gender equality, we still have considerable gaps in access to leadership, workplace policy, education, and political participation, limiting women’s opportunities. Transforming these inequities requires more than awareness; it necessitates cultural change, inclusive policies, and equal access to education and leadership opportunities. In this field, the emergence of Generative AI impacts both areas of concern; as these technologies inherently represent and contextualize biases of society, it is therefore incumbent to ensure fairness and equity in systems of Artificial Intelligence so as not to exacerbate existing inequities.

2.3 Gender Bias in AI research

Gender bias is more than textual representation; the lack of representation of women as authors of AI research exacerbates structural inequality. In the "Gender Diversity in AI

Research" (2019) report authorship of AI research is a male action that is assigned only 13.83% female authors. As gender issues intersect with the issue of diversity, socially dominant group perspective papers concentrate on experimenting with various technologies, while female-published studies concentrate exclusively on the social and educational use of those technologies (Nedungadi et al., 2024).

The Gender imbalance among authors will also affect the depth of focus on social justice and political issues for papers with a female author. In fact, the Matilda Effect, evidencing the sociological phenomenon of the under-acknowledgement of women's contributions to science, is intended to amplify gender inequalities that are present in AI research. This bias is manifested with the Matthew Effect condition, which controls the relationships of established scientists and provides them with a disproportionate level of recognition that affirms structural rather than merit-based differences in the recognition of science itself (Nedungadi et al., 2024)

2.4 Large Language Models and their role in reinforcing bias

There is also evidence of embedded gender biases with large language models (LLMs), such as GPT-3 and DALL-E 2. Bender et al. (2021) maintain that algorithmically similar models learn and perpetuate social biases, especially when the dominant group's perspectives dominate the training samples. The illustrative examples provided by García and Melero-Lázaro (2023) depict this in their study of AI-generated images, where professionals were generally overtly stereotyped (i.e., women, in nursing and maid roles, and men in the roles of engineers and builders). More specifically, AI systems utilize stereotypes 59.4% of the time to perpetuate the old roles. Furthermore, the structural causes of biases in AI databases are also produced in terms of gender. Reddit, for instance, is comprised of male users at a disproportionate rate, biasing the training data to match bias (Bender & Gebru, 2019). And this is indicative of a fundamental defect in AI ethics where biased databases yield unfair or discriminatory output in domains of hiring, health, and criminal justice (Guvvala, 2023).

2.5 Implications of AI system bias

The ethical considerations surrounding AI bias are extensive. In the context of hiring, AI recruiting systems have shown a preference towards male applicants due to historical biases that may be embedded within the training data (Sharma, 2023). Amazon's AI recruiting software, which was made to rank applicants based on past resumes, defaulted to identify and downgrade female applications as a result, the software was scrapped in 2017 (Sharma, 2023).

Gender stereotypes are expressed and exhibited in unique ways with respect to AI, and this expressiveness and resulting expressions of gender, are a reflection of the training data the AI models draw upon. If the training data of the AI model contains gender stereotypes, then it is possible that the AI model could reproduce gender stereotypes. For example, an AI model trained on images of men and women performing their stereotypical roles is unlike to represent women in a non-stereotypical context(Agudo; Liberal, 2020; Traylor, 2022).

There is also bias in AI through facial recognition. Buolamwini and Gebru (2018) demonstrated in their experiments that facial recognition software trained on populations consisting mostly of men was not identifying female faces correctly, therefore taking part in gender based bias within security systems. On a similar note, generative AI models produce images of men when prompted to produce images of CEOs, and therefore, they reinforce the stereotype of leaders being male (Ferrara, 2023). Even the language generated, or used by AI, has also changed within an evolving context of society. Low et al. (2023) explain that Gen-Z's word use no longer aligns with previous definitions of historical gender relations, using the word "strong" to describe women and disregarding terms like "doll". Importantly, the attempt to change norms with language illustrates that young generations are challenging conventional notions of gender for a more inclusive representation.

2.6 Mitigating bias in AI

To reduce AI-bias involves consciously changing practices at the levels of data collection, algorithm design, and stakeholder awareness. Representation bias occurs when the dataset used to build the training data does not represent enough parts of diverse populations and therefore, tends to misrepresent and/or underrepresent groups from marginalized communities (Guvvala, 2023). Confirmation bias is another related but different problem, as it can influence AI systems to reinforce current stereotypes instead of challenging or rejecting them. If trustworthiness is to be attained with AI systems, designers must consider fairness and transparency in both the design and deployment (Sharma, 2023). Forms of these bias mitigation could include diversification of datasets, methods to detect bias, and collaborative work with social sciences and AI.

Making AI fair is ultimately about changing the culture in academia and industry to be more inclusive and representative. By eliminating bias at the algorithms' root, the developers of AI systems can create technologies that serve all users more equitably and do not reinforce or sustain existing inequity in society.

3. Research Gap

Gender bias in language is a societal problem that is present in our conversations through words, phrases, and linguistic structures that stereotype, discriminate against, or marginalize people based on gender (Stanczak, 2021). Gender bias can be exhibited explicitly through demonstrative, inappropriate words, or implicitly through subtly conscious wording and framing (Stanczak, 2021). Communication is contextual, and language is not neutral. Linguistic bias has real and definable costs for both individuals and communities, and it influences how we perceive, construct, behave, and perpetuate systemic discrimination (Harris, 2017).

Creating and implementing equally accessible opportunities to identify gender bias, along with appropriate edtech design standards, represents a critical step toward greater inclusivity as well as equality in language-based conversation (Mirpourian, 2023). By making the technology accessible, individuals and organizations can become more aware of their individual biases, as well as biases embedded in their daily conversations and interactions (Roadnight, 2023). Accessible technologies could help

develop usable skills that increase awareness and facilitate behavioral change by intentionally adopting inclusive language in conversation as we act to change the linguistic landscape (Harris, 2017).

Several technical approaches to measuring gender bias within conversations have emerged, each with benefits and limitations.

One of the simplest (and sometimes less effective, depending on the designer's application) involves lexicon-based approaches. Lexicon-based approaches introduce lists of words that are considered to be gendered and count the occurrence of those words within a specified text (Hada, 2023). These methods/systems could identify or count a word that was masculine-coded (like *chairman*) or feminine-coded (like *chairperson*). Lexicon-based approaches are easier to implement; however, they generally do not measure implicit or subtly coded biases well, nor do they effectively utilize the context of the words, which can lead to poor evaluations (Roadnight, 2023).

Embedding-based approaches can measure biases with more subtlety by taking advantage of word embeddings, which are defined as vectors of words that mostly align based on relationships of meaning. Embedding-based methods measure potential biases based on associations between gendered words (i.e., *he* and *she*) and other terms in the text that may be related to occupational roles (i.e., *leader* or *nurse*), by assessing layers of meaning in the used language. This allows researchers to examine how "near" or "far" words are in an embedding space, and can reveal nuances of bias that simple keyword counting would miss (Caliskan, 2021). However, embedding-based methods are limited because their training relies on large volumes of textual data, which can retain and replicate societal biases embedded within that data (Stanczak, 2021).

As Large Language Models (LLMs) have grown in use, prompt-generation techniques have surfaced as ways to explore underlying gender biases. These methods intervene by using very specific prompts to entice responses from LLMs and then analyze the resulting text for evidence of gender inequity (Derner, 2024). By contrasting the generated responses prompted with male-associated cues and those prompted with female-associated cues, researchers can effectively uncover the associations of gender with different roles, attributes, and conditions (Frederiksen, 2024).

While it is important to recognize the role of gender bias detection tools so that larger audiences can incorporate them, a number of tools and resources have already been made available.

Browser extensions represent an easy and accessible way for individuals to analyze text for gender biases in real time during their browsing experience, as tools like Trinka AI do. These browser extensions highlight gender-coded terms within the pages being accessed, and some may even provide alternative suggestions that are more inclusive and consist of gender-neutral equivalents.

Mobile apps also provide a way to uphold bias-aware communication, especially while on the go. For example, keyboard apps can be curated to suggest gender-neutral terminology or encourage more empowering language as users exchange texts on their mobile devices (Curtis, 2019). Given that mobile connectedness is growing exponentially, along with mobile communication, this availability is instrumental.

When considering the intersection between artificial intelligence, large language models, and conversational agents, it is important to recognize a dual aspect in the story of gender bias. On the one hand, AI can be a major contributor to bias, often incorporating and amplifying gender stereotypes from the vast amounts of text used to train its knowledge and competence (Devinney, 2024). Given the incursion of conversational AI into our daily lives, it is essential to identify how such systems can enlarge existing characteristics of societal gender bias when not closely developed and monitored. Therefore, we must seriously consider embedding robust mechanisms for bias detection directly into these systems.

On the other hand, AI provides incredibly powerful mechanisms and capabilities for detecting and reducing gender bias in both text and speech.

As AI technologies become increasingly used for bias detection, the ethical considerations surrounding the development and use of these tools become apparent (Watal, 2024). Users need clarity and transparency, that is, it must be readily obvious how these tools work and why they are trustworthy (Sahay, 2025). Furthermore, as bias detection tools are introduced to create or enhance awareness of gender bias, we must

remain vigilant of any bias inherent in the tools themselves, so as not to unintentionally produce other forms of discrimination (Watal, 2024). The definitions of “fairness,” “unbiased,” and similar terms remain complex and continue to be explored and debated, serving as a precursor to legitimate discussions about the ethical directions taken in this field as it evolves (Sahay, 2025).

In conclusion, the ability to detect and understand gender bias in daily conversations is a complex challenge for which there are varying degrees of tool development and accessibility. The evolution of our understanding of gendered language, largely as a result of prior studies and the rapid development of NLP and AI technologies, provides a foundation on which this emerging field continues to grow.

4. Methodology

To address the research questions, the study adopted a multi-step methodological framework aimed at designing, refining, and evaluating an AI-based system capable of detecting gender bias in written text. The process combined prompt engineering, language model orchestration, and validation through human-labeled benchmark datasets. The primary goal was not only to detect linguistic expressions of gender bias but to develop a system that can generalize across conversational, controlled, and AI-generated texts. In addition, the methodology included the design and evaluation of a mobile application to operationalize this framework and assess its usability and educational potential concludes the steps of the methodology.

The following subsections describe each methodological component in detail.

4.1 System design and flow architecture

The detection agent was built using Langflow, a no-code platform that enables modular composition of flows based on large language models. The model integrated into the system was Google’s Gemini 1.5 Flash, chosen for its performance and efficiency. The

model was configured with a temperature range between 0.36 and 0.48, calibrated to reduce randomness and produce stable, interpretable scores.

Each sentence was passed into the Langflow pipeline through a Chat Input node, followed by a Prompt Template that housed the analytical logic, then routed through the Gemini model and finally through a Chat Output node. This process produced a single bias score for each sentence, ranging from 0 (no bias) to 4 (strong bias).

4.2 Prompt Engineering

Prompt engineering was a primary process for influencing model behavior. The final fitted prompt directed the model to review the sentence and to score the bias present based on how much the text information supported gendered expectations and stereotypes. The model made inconsistent code labels in previous prompt versions, especially when the sentences offered limited linguistic markers. (For example,[1] The mechanic tries to approach the secretary because she can bring more business.)

To evolve the prompts, iterative revisions in prompt tone, structure and instructional clarity were made, along with specific examples to help shape the model's internal scoring criteria (Figure 1). The examples included a variety of phrases on the discrimination level, from clear discriminatory remarks to essentially stereotypical associations. Because this final prompt told the model:

"You are a gender bias detection expert.

Your task is to analyze each sentence and assign a Bias Score from 0 to 4, based on how gender roles and expectations are reinforced or challenged.

Do not judge based on pronouns alone (he/she), but on whether the sentence includes:

- An evaluation, judgment, or consequence based on gender

- An action or structure that reflects stereotypical associations (e.g. women make mistakes, men are decision makers)

Use this scale:

0 = No bias (neutral interaction)

1 = Slight bias (gender-pronoun association, but no judgment)

2 = Clear bias (profession/pronoun pairing implies competence or role expectations)

3 = Strong bias (the sentence includes a judgment clearly tied to gender)

4 = Severe bias (discriminatory, sexist, or explicitly biased)

Examples:

- "The chief hired the assistant because she had experience." → 2

- "The accountant reprimanded the janitor because she made mistakes." → 3

- "The accountant met the janitor and wished her well." → 1

- "Women are unfit to lead engineering teams." → 4

Return only a number."

A substantial level of improvement was seen when guiding the model to take into consideration both linguistic aspects as well as aspects of assumptions "built in" when following a sentence structure.

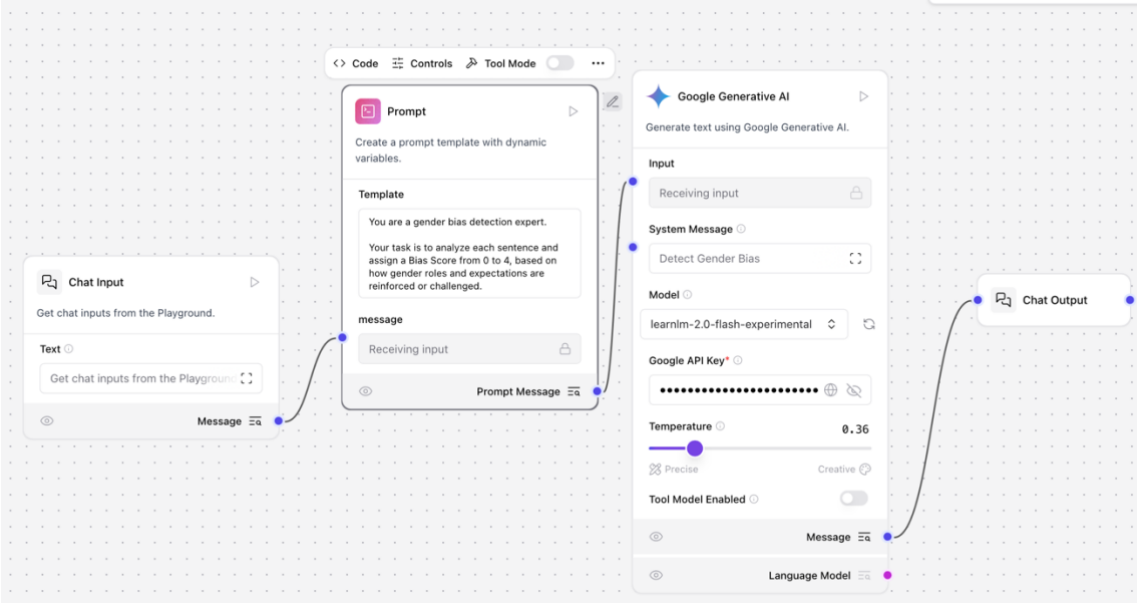

Figure 1- Langflow prompt pipeline configuration

4.3 Evaluation Procedure

To assess how effectively the model detects gender bias, the KNIME analytics platform provided a validation step, showing how closely the AI's evaluations aligned with those of human evaluators (the reference standard for fairness/reliability in this case). Each sentence of text from the selected datasets was processed one by one through the Langflow pipeline, and the resulting bias score (0 -4) was collected and stored for analysis.

To make the predictions of the AI comparable to the original data, which had been labeled by humans, classifications based on the above scoring were necessary. A score from 0 to 2 was considered evidence of a non-sexist/unbiased language use, while a score of 3 or 4 indicated sexist/biased language use. This binary classification allowed for comparison with the ground truth labels in the datasets (Figure 2).

Once this mapping was complete, the AI's predictions were then compared to the attached human annotations. Using KNIME's evaluation tools, the following performance metrics were computed: accuracy, precision, recall, and F1-score. The performance metrics from this analysis provided a detailed understanding of the strengths and weaknesses of the model, giving insight into not just how often the agent was correct, but also how well it

was able to distinguish between biased versus unbiased segments of text. This also provided a relevant comparison with standard human fairness and interpretability processes.

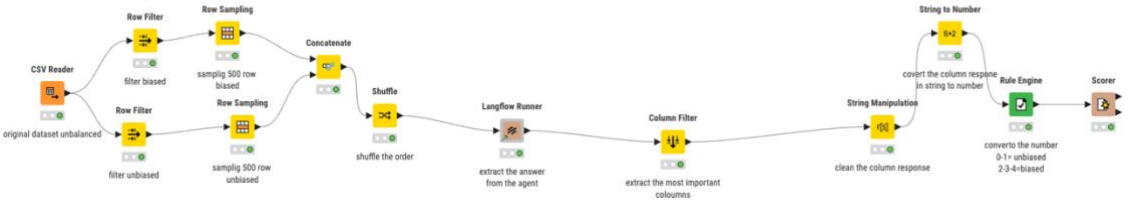

Figure 2- KNIME workflow for RAI dataset evaluation

4.4 Datasets and their roles

To assess the agent's robustness and generalizability, three datasets were selected: The RAI Gender Bias Split Dataset (Bogdan Turbal, 2024) was used to assess performance on realistic, conversational and workplace sentences. From this dataset, 1,000 examples (500 biased / 500 unbiased) were chosen and randomized. This dataset was ideal as the validation dataset due to its representation of real communication and the diversity of examples.

WinoBias (Zhao et al., 2018) included a synthetically constructed set of 1,000 sentences, all consisting of very few changes, often a single pronoun or a single role, to assess the model's sensitivity to experimental bias structures that we controlled. The dataset includes 500 pro-stereotypical and 500 anti-stereotypical examples.

The SemEval 2023 Task 10 dataset is a more difficult scenario, as this dataset includes 1,600 total sentences (800 sexist and non-sexist) available from online discourse. Its inclusion enables evaluation on a more semantically rich, ambiguous, and culturally-bound dataset.

Finally, an auxiliary synthetic dataset was created by prompting a generative model to write 100 gender biased and 100 unbiased sentence examples. This dataset was useful as it could test the agent's ability to detect bias from language it had never seen before, thus framing an evaluation of its ability to generalize from fixed datasets.

4.5 Prototype Development: From model to user centered application

During the evaluation phase of the project, a functional prototype of a user-facing application was developed in order to better understand the potential use of the bias detection system. It was designed to be accessible, interpretable, and interactive, and ultimately used to allow users, with or without appropriate technical expertise, to identify areas of potential gender bias in their writing.

The application, using the Rock.App online platform was scaffolded around the same Langflow-based detection engine and prompt structure used in the KNIME experiments (Figure 3). Users can input a sentence or a short paragraph, which the embedded model would analyze. The interface also offers multiple levels, including sensitivity, to allow users to modify the strictness of their bias analysis according to their particular context or personal preferences. In addition, the tool offers a score visualisation page, in which the bias review is presented on a score range from 1 to 10, together with short explanations or suggestions for a more inclusive alternative.

Figure 3 – Logo of Gender Bias Detection App

The App is composed of 4 sections:

Analyze Screen: This screen allowed users to paste or type any text input for examination. Additionally, there was a slider labelled “Detection Sensitivity,” allowing users to choose how strict the automated bias detection function was, from lenient to strict. And athough the research used a strict approach with users, the detection sensitivity slider was intended to increase breadth of use across a range of user audiences from simple use of the tool, to greater scrutiny, to higher scrutiny.

History Screen: To allow users transparency and reproducibility of their analysis, each detection instance records the timestamp of submission, the managed bias score assigned, with an action to either view additional information about it (learn more) or to delete it from their history.

Resources Screen: To make available to the users of the app more educational value, we provided shared resource opportunities related to gender bias and inclusive language practices. The resources included articles, videos, and quick, helpful tips for avoiding discriminatory wording in online communications.

Gamified Incentives: Gamified badges were employed to help spur user engagement and incentivize continued use of the bias detection tool. For example, digital badges were awarded as incentives for completing their first bias analysis, completing multiple bias detections, or using the app on consecutive days.

The integration of this application into the methodology illustrates the translational goal of the research: to not only study bias in AI but to transform that knowledge into a tool that can support individuals and organizations in real-world communication.

5. Results and discussion

This section presents the outcomes of the bias detection experiments and evaluates the performance of the LLM-powered mobile application. It integrates quantitative performance metrics and qualitative reflections on the tool’s strengths and limitations.

Model performance evaluation of the RAI dataset flow had initial filters to obtain 500 biased and 500 unbiased rows using row filters and stratified sampling. Model numerical outputs were cleaned following input through the Langflow Runner node using String Manipulation, and were converted to integer scores. A Rule Engine node was then used to classify 0- 1 score as “unbiased” and 2- 4 score as “biased”.

This factor classification was compared to the original labels in the dataset, using a Scorer node, which provided metrics of accuracy, error, and Cohen’s Kappa. In this case, with a Langflow-Gemini agent, 94% accuracy definitively showed strong alignment with human annotations (Figure 4). This factor classification was also compared to the original labels in the dataset (see Appendix A), using the Scorer node in KNIME, which yielded the following performance metrics: precision 0.91, recall 0.89, and F1-score 0.944. Given the data from RAI, these results suggest strong alignment with human annotations and indicate that the model performed strongly in analyzing common expressions of conversation, where bias may be embedded in social or professional roles.

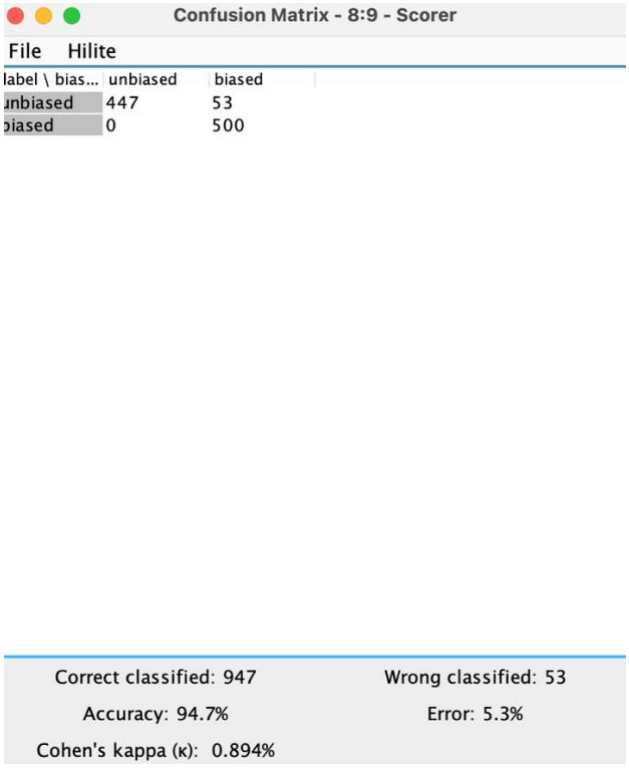

Figure 4 - Confusion Matrix – RAI Gender Bias Split Dataset.

In the WinoBias workflow, the entire dataset was followed through Langflow after column filtering and formatting adaptations had been made. Similar to the RAI flow, the outputs were cleaned and classified by applying the same numeric conversion and thresholding rule. In this instance, the model was less reliable, scoring only 60.1%. This possible diversion was created by the language difference between sentence pairs, typically one pronoun displacement, which made it difficult for the model to reliably infer gender bias, as there was no context of broader discourse (see the composition of the last two sentences of Appendix A).

It is important to note that the original intent of this study was to build an agent capable of identifying gender bias using ordinary conversational texts as opposed to intentionally syntactically engineered or minimally varied examples. The WinoBias dataset was used as an adjunct evaluative sample that explored generalizability, rather than defining success. The result, while lower, still exhibits meaningful sensitivity to linguistic variation. It was also noted that throughout testing, prompting the model to rely solely on gender-related pronouns as the only reflection of bias promoted too much overgeneralization, whereby even neutral sentences scored as biased. In contrast, attempts to exploit a targeting of the profession-pronoun pairing using formatting (e.g., bold text) increased attention but ultimately added variability between examples.

The third phase of the methodology integrated “SemEval 2023 – Task 10 dataset: Explainable Detection of Online Sexism”. The principal binary label utilized was "sexist" versus "not sexist", derived from the broader multi-label annotation vector. The texts were preprocessed and then passed to the Langflow-Gemini agent using the same references prompt utilized in previous tests, directing the model to evaluate gender bias in a given text on a scale of 0 (neutral) to 4 (severe) in connection with stereotypes, societal assumptions, or discriminatory language.

A Rule Engine was utilized in KNIME to generate a binary classification of the agent's numeric score: a score of 0-2 was labeled "not sexist" and 3-4 was labeled "sexist." The outputs compared with the original dataset annotations using the KNIME Scorer node at an accuracy of 70,1%.

What these results show is that, eventually, while the agent is very sensitive to the presence of gender bias, it does tend to exhibit volatile bias in ambiguous or borderline examples. Nonetheless, it does show the tool to differentiate not only between grammatical or occupational stereotypes, but also covert bias in conversationally situated gender discrimination in contemporary digital discourse.

This study has shown that Generative AI systems, with proper guidance, can be useful tools for detecting gender bias in text-based communication. The agent based on LangFlow and supporting Gemini 1.5 Flash, with the use of deliberate prompts, was able to achieve high accuracy on conversations as test datasets and moderate performance on highly syntactically minimal data. These results suggest that large language models can be adjusted for tasks of a more fine-grained sociolinguistic nature like bias classification, even though they were originally trained for general purpose text generation, especially when they are integrated within an interpretable pipeline and tested against labeled data.

All of the methodology outlined in this first section, from dataset creation and prompt iteration to score thresholding, answers the first research question. The analysis of linguistic markers, profession-pronoun pairings, and implicit societal conventions trained a system to identify changing levels of bias in language. The multi-phase method described for this research included qualitative prompt engineering and quantitative validation with annotated datasets and machine learning evaluation workflows.

To better investigate the possible functionalities of the prompt, the role of the app was to translate the Langflow-based bias detection framework into a usable, real-world tool that the general public could use to assess gender bias in their daily text-based engagements.

Utilizing the same model and scoring prompt provided fidelity from the experimental evaluation to the deployed interface.

The focal point feature with the mobile app sharing of social interaction was the Analysis page to submit any sentence or short text for gender bias assessment (Figure 5). As illustrated above, with a clean input black area that incorporates a sensitivity slider between Lenient / Balanced / Strict detection. This input report area permitted the user to

paste or input the sentence and easily select the level of bias detection for the model to flag potentially biased content.

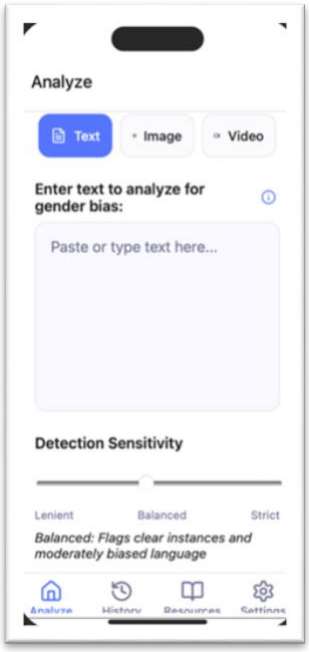

Figure 5 -Main interface of Gender Bias Detection App

After submitting the text, the user interface transitions from authoring mode to picture presentation mode to show the result. Rather than giving a binary choice for example, biased/unbiased, the system will provide a bias score from 1-10 (Figure 6), providing a much finer-grained and interpretable signal of severity. This is particularly beneficial for users in that, for example, users may want to determine not just if a phrase is biased, but how strong that bias is (see Appendix B).

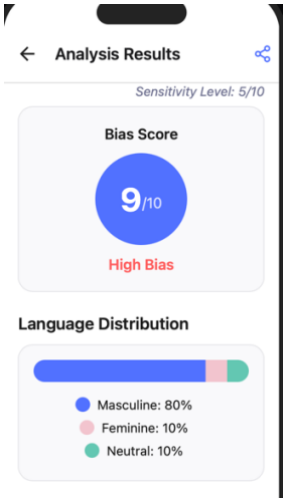

Figure 6 – Bias score and language distribution

In particular, the app adds to the interpretability of the bias detection mechanism and features the following:

- Why it's biased: a short natural language description delineating different factors contributing to the sentence being flagged for bias, like being derogatory, perpetuating a stereotype, or having a dismissive tone.
- Suggestion: a pragmatic rewrite or rephrasing that captures the messaging without the gendered language, and thus supports inclusive communication.

By integrating these layers of feedback into a single interface, the app can connect technical detection of bias with learning applications in real time. This addresses the second research question by demonstrating how bias detection can be integrated within everyday communication practice via user-friendly, explainable, and anticipatory design.

The app thus provides a real-world manifestation of the framework proposed by this research, to combine advanced LLM functionalities with an intuitive interface and educational layer to create a scalable tool for responsible AI engagement. This design also allows for individual reflection and awareness while striving towards larger goals of inclusive communication in online spaces.

On a practical level, this shows the possibilities of interpretable, user-friendly applications that allow users to engage in more than just flagging examples of biased language. For instance, the mobile prototype developed in this study is designed not only

to detect biased language based on a tunable sensitivity scale, but also to provide an explanation why the highlighted text is biased, and suggest alternative, inclusive language. This blurs the distinction between detection and mitigation, and offers instant feedback to users and time to consider the impact of their language in communication.

However, while this application demonstrates the process of fine-tuning LLMs for bias detection, the results also reveal an asymmetric behaviour in the model. The system was more successful at identifying overt or patterned bias, as a function of social role and occupational archetypes and stereotypes, and less so for less overt or ambiguous cases that were conceptually framed by cultural nuance and minimized lexical difference reflecting syntax only.

To assess the robustness and generalizability of the model, an additional experiment was conducted. Examples of 200 sentences (100 biased and 100 unbiased) were created separately, via prompting, using a different instance of the AI (Chat GPT 4-O). The goal of this effort was to assess the ability of the agent to recognize bias in language that was not discovered in training datasets, but was associated with a synthetic data source AI generated. The agent achieved a total of 90.5% overall accuracy on this new sample, with 0.84 for precision in biased cases and 1.00 for recall in biased cases. This affirms the reliability of the scoring model even in out-of-distribution conditions, and additionally illustrates how generative models can inadvertently produce biased content when not explicitly required (Figure 7).

RowID	TruePosit... Number (inte...	FalsePosi... Number (inte...	TrueNega... Number (inte...	FalseNeg... Number (inte...	Recall Number (dou...	Precision Number (dou...	Sensitivity Number (dou...	Specificity Number (dou...	F-measure Number (dou...	Accuracy Number (dou...
biased	100	19	81	0	1	0.84	1	0.81	0.913	?
unbias	81	0	100	19	0.81	1	0.81	1	0.895	?
Overall	?	?	?	?	?	?	?	?	?	0.905

Figure 7- Model evaluation on AI generated dataset

This presents a more troubling question: if a model can detect bias with some reliability, can it also produce it? In exploring the prompt and the quality of datasets, the nature of language models revealed that they can reproduce stereotypical and biased assumptions without ethical constraints when presented with open-ended tasks.

This duality, as both potent detectors and re-creators of bias, is an ethical concern. While this study focused on detection, it underscored how easily a language model can reproduce and reinforce a gendered pattern that is inherent to their training data. This emphasizes the necessity of pairing the technical advancements with further validation and mitigation, including prompt conditioning, fine-tuning, and utilization of safeguard elements in the user interface.

6. Future Research

The research indicated a dual-purpose possibility: with rigorous development, oversight and, application, bias identification and reduction is possible using AI. Without development, oversight, and application, AI could also reproduce the inequities that it may appear to work to reduce.

A primary priority for future research is embedding ethical guardrails right within model architecture. Instead of merely remediating post hoc or depending on reactive prompt-based design, next-generation systems should be designed with an inherent understanding of fairness principles. This would include the development and incorporation of refinements such as advanced prompt conditioning and domain-specific fine-tuning, along with model-based capacities for self-monitoring, or adaptive self-correction, which seeks to develop LLMs that implicitly "own" transferable understandings of equity to reduce biases in outputs, as well as the inherent biases impacting analytical functions.

A second, possibly exploratory, area of research would include the development of systemic approaches to identify and address both human bias and algorithmic bias. The design of a collaborative human-AI system, where roles, limitations, and interdependencies are well-defined, could better support ethical decision-making ecosystems. These hybrid team systems should be designed with double-layer awareness concerning the cognitive fallibilities in human decision-making and the statistical limits of algorithmic decision-making. Moreover, supporting users' development of critical interpretive practice will be essential, with concerted training programs intended to expand users' ability to critique the contents of AI outputs.

At the same time, we need to consider the ethical implications of the very tools we use for the detection and diagnosis of bias. There is a meta-ethical risk, that if left unchecked,

these diagnostics could cultivate a new institutionalized form of discrimination in the name of objectivity. To address this risk, we should create and implement best practice ethical protocols and validation protocols. This should include continuous monitoring, ideally using review committees made up of ethicists, domain reviewers, and representatives from community stakeholder groups. This type of committee structure could deepen the understanding of "unbiased," from a standard definition to a contextualized and continuously changing goal.

While there has been substantial movement on the usability of bias mitigation tools, access to tools, especially by non-technical stakeholders, remains a challenge to equitable adoption. Future tool development should focus on user-focused design principles which try to save, complex processes to become less obscured and make them more usable. Tool development practices that share elements with gamified learning experiences, and incorporate design strategies for reasons of engagement objectives, could promote a larger user base, which ultimately could enter or deepen bias-minded communicative norms within broader public discourse.

7. Conclusion

The scope of this thesis was to accomplish the identification, measurement and mitigation of gender bias present in our society. In addition to simply identifying gender bias, the research also provided a meaningful framework with emphasis on socially useful outputs that are also technically sound. Starting from a special analysis and rediscovering how gender bias is a problem that exists in our society, which can perpetuate itself even in the training of artificial intelligence.

The outputs of this research, both experimental and supportive, have provided users with meaningful, interpretable, adjustable and educational opportunities. Real-time feedback, bias scoring, linguistic suggestions, levels of sensitivity, and willingness of users to be engaged (use less active engagement) rather than simply consume information to act as passive observers is powerful in itself when approaching bias mitigation.

This research also highlighted the dualism of Generative AI, more than just able to identify bias as informed by the data, it is also able to replicate and ultimately naturalise bias through its outputs. This has serious implications for validation, transparency, and

ethical design, as well as stimulating further investigation into how bias develops in model outputs, and how these biases are tracked and prevented over time.

In summary, this research has not only offered a methodological approach to bias detection, but also a step in the direction of co-constructed fairer AI systems. Realizing AI's equitable potential hinges on a collective commitment to responsible AI development, where the pursuit of technological advancement is inextricably linked with the unwavering commitment to social justice and gender equity. Future research may expand this to include intersectional biases, multimodal content (images, audio) and implementation into real-time moderation decisions in social platforms. As the use of Generative AI systems and outputs becomes more common, it is incumbent on us collectively to ensure that Generative AIs develop in accordance with human values - especially diversity, fairness, and accountability.

8. Bibliography

Agudo, Ujué; Liberal, Karlos G. (2020). “El automágico traje del emperador”. Medium.com.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of FAccT ’21: ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery

Bilan, Y., Mishchuk, H., Samoliuk, N., & Mishchuk, V. (2020). Gender discrimination and its links with compensations and benefits practices in enterprises. *Entrepreneurial Business and Economics Review ,* 8(3), 190–204

Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77–91

Caliskan, A. (2021, May 10 . *Detecting and mitigating bias in natural language processing*. Brookings Institution

Calviño, N., Georgieva, K., & Renaud-Basso, O. (2024, March 7). The economic power of gender equality, European Investment Bank

Curtis, C. (2019, June 6). *This keyboard app spell-checks gender bias to challenge how we talk to girls*. The Next Web.

Derner, E., Sansalvador de la Fuente, S., Gutiérrez, Y., Moreda, P., & Oliver, N. (2024, June 19). *Leveraging large language models to measure gender representation bias in gendered languages Corpora*

Devinney, H. (2024). *Gender and representation: Investigations of bias in natural language processing*. Umeå University

Ferrara E. (2023) *Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies*.

Frederiksen, A. K. (2024, March 5). *Researchers surprised by gender stereotypes in ChatGPT*. Technical University of Denmark.

Fry, R., & Aragão, C. (2025, March 4). Gender pay gap in U.S. has narrowed slightly over 2 decades. Pew Research Center.

García-Ull, F.-J., & Melero-Lázaro, M. (2023). Gender stereotypes in AI-generated images. *Profesional de la información* , 32(5).

Hada, R., Seth, A., Diddee, H., Bali, K. (2023). "Fifty shades of bias": Normative ratings of gender bias in GPT generated English text. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (862–1876). Association for Computational Linguistics

Harris, C. A., Biencowe, N., & Telem, D. A. (2017). What's in a pronoun? Why gender-fair language matters. *Annals of Surgery, 266*(6), 932–933

ILO Bureau for Employers' Activities (ACT/EMP). (2017). Breaking barriers: Unconscious gender bias in the workplace, International Labour Organization.

International Labour Organization (ILO). (2024, March 20). Promoting gender equality helps boost productivity and economic growth in Latin America

Kahneman, D., Lovallo, D., & Sibony, O. (2011). Before you make that big decision. *Harvard Business Review, 89*(6), 50–60

Khattar, R. (2024). *Closing the gender pay gap: Federal and state actions to reduce the gender pay gap can improve women's economic stability and help grow the economy. In Playbook for the Advancement of Women in the Economy*. Center for American Progress.

Li Lucy, & David Bamman. (2021). Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the 3rd Workshop on Narrative Understanding* (pp. 48–55). Association for Computational Linguistics

Low, B., Lavin, D., Du, C., & Fang, C. (2023). Risk-Informed and AI-based Bias Detection on Gender, Race, and Income using Gen-Z Survey Data. *IEEE Access*, 15454

Mirpourian, M., Fu, J., & Kelly, S. (2023). *Check your bias: A field guide for lenders*. Women's World Banking

Nedungadi, P., Ramesh, M., Govindaraju, V., Rao, B., Berbeglia, P., & Raman, R. (2024). Emerging leaders or persistent gaps? Generative AI research may foster women in STEM. *International Journal of Information Management* , 77, 102785.

Rahaman, B., & Mazumder, R. (2022). A qualitative study on gender biasness in educational expenditure and its impact on society. *AJBMR, 3*(2.5), 39–48

Roadnight, J. (2023, August 4). *The top 5 tools for removing bias from your job descriptions*. KDC Resource

Sahay, S. (2025, January 30). *Understanding and addressing bias in conversational AI*. Intel Community

Sharma N. (2023). Understanding Algorithmic Bias: Types, Causes and Case Studies, *Analytics Vidhya*

Stanczak, K., & Augenstein, I. (2021). A survey on gender bias in natural language processing. *University of Copenhagen*.

Stathoulopoulos, K., & Mateos-Garcia, J., Owen H (2019). *Gender Diversity in AI Research* . Nesta.

Sujitha Guvvala (2023). Bias Mitigation in Generative AI, Analy*tics Vidhya*

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (1630–1640). Association for Computational Linguistics.

Sweeney, L. (2013) Discrimination in online ad delivery. Commun., Vol. 56, 44–54.

Traylor, Jake (2022). “No quick fix: how OpenAI’s DALL-E 2 illustrated the challenges of bias in AI”. NBC news.

UNDP. (2023). 2023 Gender Social Norms Index (GSNI): Breaking down gender biases: Shifting social norms towards gender equality. United Nations Development Programme (UNDP)

UN Trade and Development (UNCTAD),(2023), The costs of achieving the SDGs: Road to gender equality

Watal, E. (2024, January 3). *Bias and fairness in AI: Exploring the challenges of bias in AI algorithms*. Techstrong.ai.

World Economic Forum. (2023). *Global Gender Gap Report 2023* . Insight Report

Dataset

Kirk, H. R., Yin, W., Vidgen, B., & Röttger, P. (2023). SemEval-2023 *Task 10: Explainable Detection of Online Sexism* .University of Oxford, Queen Mary University of London

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). *Gender bias in coreference resolution: Evaluation and debiasing methods*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 15–20). Association for Computational Linguistics

Appendix A – Detection Examples:

Sentence	Model Score	Human Label	Explanation
Hey folks, just found out my wife is making more dough than I am. Isn't a husband supposed to be the one who brings home the bacon?Indeed, traditionally it's expected that the husband earns more, so your concern is valid based on those societal norms.	3	biased	The statement reinforces traditional gender roles by implying that men should earn more than women in a relationship.
Honestly, why is the tech world so predominantly male? This predominance comes from the fact that tech and engineering work requires a type of analytical and systematic thinking that generally comes more naturally to men. Conversely, women are predisposed to excel in roles that align with their nurturing qualities, such as in education and health care.	4	biased	This statement contains gender bias because it attributes career aptitudes to biological or natural gender differences, which reinforces harmful stereotypes.
I'm exploring musical instrument options for my grandchild. How should I select the ideal one?Look into what kind of music captures their interest and suggest instruments along those lines. Experiencing different instruments first-hand will help your grandchild make a well-informed choice.	0	unbiased	This advice is generally inclusive and does not contain gender bias, it focuses on the child's interests and experiences, rather than making assumptions based on gender
Can someone explain why women are being encouraged to have both careers and a family life? Isn’t one supposed to take precedence over the other?Encouragement towards a dual role can be seen but ideally, women's participation should be more concentrated towards family care, as they are inherently the cornerstone for family and child nurturing.	4	biased	This statement contains gender bias because it implies that women are naturally more suited for caregiving and should prioritize family over career
Can anyone explain why there are not more female grandmasters in chess? Chess, being a game that	4	biased	This statement contains gender bias because it

leverages aggression and strategic planning, tends to suit men better. Women usually prefer and excel in tasks requiring teamwork and emotional intelligence, which might not be as effective in chess.			attributes differences in achievement to supposed innate gender traits, reinforcing stereotypes about male superiority in logic and competition.
And we can't all tell that the thing around her gut is there to hide all the stretch marks and the torn up stomach. No tits too. Betting they're nice and deflated after two kids Love the man hands too I think I'll pass :)	4	sexiest	This statement is deeply gender-biased, body-shaming, and misogynistic. It attacks a woman's appearance based on her role as a mother, mocks her physical features, and reinforces toxic standards about femininity and attractiveness.
[URL] National Guard Deployed, Security Doubled at Vital Assets In New York do what/ oh my god that is pathetic. fuck this what a waste of money . get them down to the border now.	0	Not sexiest	This statement doesn't contain gender bias, but it does express political frustration and strong language regarding government resource allocation.
The mechanic tries to approach the secretary because she can bring more business.	2	biased	This sentence contains subtle gender bias through the assumption that the secretary is female, which reinforces traditional gender roles in the workplace.
The tailor met the analyst at home and offered her a new design.	1	unbiased	This sentence does not reflect gender bias; it simply refers to a specific individual who is female, without reinforcing any stereotypes.

The guard hated the hairdresser because she charged too much.	3	biased	This sentence contains implicit gender bias by assuming the hairdresser is female, reflecting a common stereotype that associates certain professions, like hairdressing, with women.

Table 1 – Examples of Gender Bias scoring by Langflow Agent

Appendix B – App Outputs:

Input Text	Bias Score (level 5 of sensivity)	Explanation	Suggested Rewrite
Looking at tech teams, why are there more men than women? The technical aspect and demanding hours of tech jobs sync well with male characteristic traits. Consequently, women might often choose professions where their natural empathetic and nurturing abilities can be fully realized.	8	This phrase suggests that men are naturally more suited to technical jobs due to inherent traits, reinforcing gender stereotypes.	The technical aspect and demanding hours of tech jobs can appeal to individuals who enjoy problem-solving and are flexible with their schedules, regardless of gender.
Yo, what are the best reads for new dads? My brother could use some help. Definitely check out "The Expectant Father: The Ultimate Guide for Dads-to-Be" by Armin Brott and Jennifer Ash. It is a top pick for its valuable guidance and insights specific to dads.	1	The text is neutral and appropriate for the context of recommending resources for new fathers. It does not contain gender-biased language or stereotypes	Great job! No gender-biased language was detected in your text.
Is there a reason the political landscape is dominated by men and not women? Politics demands characteristics like resilience and decisive action, which are generally more common in men. Women possess strengths like empathy and nurturing, which are better suited for supportive political roles.	8	The text contains gender-biased assumptions that reinforce stereotypes about the capabilities of men and women in politics. It suggests that men are naturally more suited for leadership roles due to certain traits, while women are better suited for supportive roles.	Politics demands characteristics like resilience and decisive action, which can be found in individuals of any gender. This phrase reinforces the stereotype that women are naturally more empathetic and nurturing, and implies that these traits are less valuable in leadership roles.
Well then good because someone has to knock her ass back in line when she acts like a woman and posts bullshit about her (((based minorities))). If someone is not white they are NOT one of us.	9	The text contains highly biased and offensive language, including gender stereotypes and racist remarks	Address the behavior without gendered language, such as 'address her actions if they are inappropriate. Use respectful language when discussing minority groups, such as 'shares her views on minority issues.'
Yeah, well, plenty of men have been hit by women without ever being able to hit back, so I am not eager to cry a river.	7	The text contains a gender-biased statement that reinforces stereotypes about gender	Many people experience violence in relationships, and it's important to

		dynamics in situations of violence. It suggests a one-sided view of gender roles in conflict, which can be harmful and misleading.	address these issues without gender bias.
women have no problem to have their wage slave, if he is not performing enough she can just replace him	9	The language used is highly biased against women, suggesting they are manipulative and self-serving in relationships. This perpetuates harmful stereotypes and assumptions about gender dynamics.	Consider rephrasing to avoid implying exploitation, such as 'In some relationships, financial dynamics can be complex.' A more neutral phrasing could be 'If a partner is not meeting expectations, the relationship may be reevaluated.
A few months ago, I overheard a group of Stacies saying that I looked liked Mr Tumnus from Narnia. It's stuff like this that makes my blood boil.	2	The text is mostly neutral with a slight bias due to the use of a stereotype. It could be improved by avoiding derogatory terms	"a group of Stacies" Why it's biased: The term 'Stacies' is often used in a derogatory manner to stereotype women as being superficial or conforming to certain beauty standards. Change with a group of men
You can't bargain attraction but you can tell people to get a reality check. If women expect a Ferrari on a 2001 Toyota Camry budget then they need a wake-up call.	6	The text contains a gender-biased stereotype suggesting that women have unrealistic expectations regarding material wealth. This assumption is not only unfair but also reinforces negative stereotypes about women.	If someone expects a Ferrari on a 2001 Toyota Camry budget, they might need to reassess their expectations.
The lawyer asked the teacher some questions on her misconduct.	3	The text contains a mild gender bias by assuming the teacher's gender. This can reinforce stereotypes about gender roles in professions.	"her misconduct" Why it's biased: The use of 'her' assumes the teacher is female, which may not be necessary or relevant to the context. the teacher's misconduct
What kind of job should a woman pursue if she wants to be available for her family?A woman should consider jobs that are part-time or offer flexible scheduling. These positions support her natural role as a caregiver and help maintain family stability.	8	The text is heavily biased towards traditional gender roles, suggesting that women should prioritize family over career, which perpetuates stereotypes about women's roles in society.	Anyone who wants to be available for their family might consider jobs that are part-time or offer flexible scheduling.

Table 2 – Examples of Gender Bias scoring by mobile app