Chapter 4

Responsible AI: Identifying and Mitigating Bias in NLU Models

In the rapidly evolving landscape of AI, the transition from deterministic, rule-based systems to data-driven NLU models has ushered in unprecedented capabilities, yet with the inherent risk of perpetuating biases inherited from training data. This chapter marks contribution to the field by presenting a holistic exploration of bias identification and mitigation strategies applied to both Small Language Models (SLMs) and LLMs. The relevance of this research lies in its dual focus: first, on widely used sentiment analysis frameworks like Flair, TextBlob, and Vader, which represent SLMs, and second, on prominent LLMs such as BARD, ChatGPT, and LLAMA2-Chat. This twofold investigation aims to uncover and address biases across a spectrum of models, shedding

light on the nuances associated with both small and large-scale language understanding systems.

The first set of experiments delves into the bias identification and mitigation landscape of SLMs and tools, focusing on Flair, TextBlob, and Vader. These models, although compact, play a pivotal role in various applications, from sentiment analysis to chatbots. The rationale behind scrutinizing these smaller models lies in their ubiquity and ease of integration into diverse contexts. Despite their size, their impact on societal perceptions and interactions is substantial. Uncovering biases in these models is crucial, as they are often employed in real-world applications, potentially perpetuating systemic disadvantages if left unchecked. The outcomes of these experiments contribute to a deeper understanding of biases within widely-adopted NLU models, offering insights into collective, personal, and physical identity biases, and culminate in a set of guidelines for ethical and equitable model utilization.

The second set of experiments shifts the focus to the realm of LLMs and their association with societal biases, particularly in gender representation. The significance of investigating biases in LLMs lies in their extensive applications, from content generation to natural language interactions. This research introduces an innovative approach employing PE and ICL to rectify biases within LLMs. The emphasis here is on guiding LLMs to generate more equitable content through nuanced prompts and in-context feedback. The experimental results, particularly on BARD, ChatGPT, and LLAMA2-Chat, highlight a substantial reduction in gender bias, with a specific impact observed in traditionally sensitive areas such as 'Literature'. The efficacy of PE and ICL emerges as a promising avenue in the pursuit of unbiased AI LMs.

The dual exploration of bias identification and mitigation in both SLMs and LLMs provides a comprehensive understanding of the challenges and opportunities across the

NLU spectrum. By addressing biases in models of varying scales, this research bridges the gap between models widely integrated into everyday applications and those at the forefront of cutting-edge AI technologies. The synthesis of findings aims to inform not only practitioners working with SLMs but also researchers and developers involved in shaping the trajectory of large-scale language understanding systems.

In the following sections, we delve into the methodologies employed in each set of experiments, presenting a detailed examination of bias identification and mitigation strategies. The overarching objective is to offer a nuanced perspective on biases in NLU models and to propose effective mechanisms for their rectification, contributing to the ongoing discourse on responsible and ethical AI development.

4.1 Bias in Largely Used SLMs

The paradigm shift in AI towards data-driven NLU models is transforming how systems comprehend language. Despite their unparalleled predictive capabilities, these models are susceptible to biases inherited from training data, posing ethical concerns. This study introduces an innovative bias detection methodology, employing template-driven synthetic data generation.

In our empirical evaluations, we apply this methodology to well-known sentiment analysis frameworks: Flair, TextBlob, and Vader. Our focus extends to group fairness and counterfactual fairness, revealing nuanced disparities in widely adopted NLU models related to collective, personal, and physical identities. The empirical findings underscore the existence of biases within these models, prompting a critical examination of their implications.

As a culmination of this research, we present a comprehensive set of guidelines. These guidelines aim to foster a heightened awareness, informed decision-making, and equitable

utilization of NLU models in practical operational environments. By addressing biases in widely-used sentiment analysis frameworks, we contribute to the ongoing discourse on responsible AI, emphasizing the need for ethical considerations and fairness in the deployment of NLU models.

4.1.1 Background

The landscape of ML has undergone a transformative shift, transitioning from traditional rule-based approaches to granular data-driven architectures. These contemporary systems, reliant on extensive datasets, at times struggle to uphold rigorous standards of quality and integrity within their training data. This deficiency can inadvertently inject biases into ML models, potentially leading to unjust consequences for specific societal segments (Blodgett et al., 2020; Dwivedi et al., 2023).

Simultaneously, the substantial energy demands associated with training these models have cast a spotlight on a growing environmental concern, marked by significant carbon emissions. Continuously refining or replacing models to attain unbiased configurations presents challenges from both environmental and computational standpoints. In response to these pressing issues, we propose the development of comprehensive datasheets for ML models and their corresponding datasets. Our objective is to promote the responsible utilization of NLU models, proactively mitigating biases and forestalling any potential negative societal implications.

The ensuing sections follow a structured trajectory: 'Related Works' offers an overview of the burgeoning literature on ethically grounded AI in the context of NLU. 'Methods' elucidates our research approach, encompassing hypotheses, terminology, notation, evaluation strategies, focal models, and the overarching experimental framework.

Subsequently, 'Results and Discussion' unpacks our investigative findings and advocates for best practices in the ethical deployment of AI solutions.

4.1.2 Related Works

Responsible AI refers to the ethical and conscientious development, deployment, and use of AI technologies. It encompasses a set of principles, practices, and guidelines aimed at mitigating potential risks, transparency, accountability, and inclusivity in AI systems. As AI technologies become increasingly integrated into various aspects of society, addressing the potential societal impacts and ethical considerations becomes crucial. Responsible AI seeks to strike a balance between technological innovation and safeguarding the well-being of individuals, communities, and society as a whole.

The ethos of Responsible AI, interchangeably termed as fairness in AI, serves as a foundational pillar in the AI domain to foster ethically sound and equitable artificial solutions. Given the inherent probabilistic behavior of ML models, the propensity for prediction errors remains inescapable. While such deviations might be benign in generic contexts, they assume a more pernicious nature when harnessed for decision-making processes involving human subjects - recruitment assessments or judicial enforcement stand as pertinent examples.

A model's sterling classification accuracy during conventional tests isn't tantamount to its impartiality. An array of empirical studies indicates that even models boasting exceptional accuracy metrics can demonstrate biases against distinct demographics due to skewed data distribution (Larrazabal et al., 2020) or intrinsic dataset biases. Buolamwini and Gebru's (Buolamwini, 2018; Klare et al., 2015) seminal work accentuates this, highlighting the disparity in facial analysis datasets IJB-A (Klare et al.,

2015) and Adience (Eidinger et al., 2014)- a conspicuous dearth of non-light-skinned subjects, engendering errors when scrutinizing underrepresented cohorts.

Similarly, Tatman (Tatman, 2017) has demonstrated the differential error rates in YouTube's auto-captioning mechanism, with female speech bearing the brunt compared to its male counterpart. While ostensibly innocuous, such discrepancies can attenuate audience reach over time - particularly among those reliant on accurate captions, and subsequently, undermine search result prominence due to compromised textual integrity.

In another investigative vein, Hardt et al. elucidated the ingrained biases in loan application decision systems (Hardt et al., 2016), revealing preferential treatments to certain demographics, thus sowing seeds for potential societal imbalances.

The scholarly community has been proactive, propounding a plethora of fairness metrics for bias assessment in NLU models (Borkan et al., 2019; Dixon et al., 2018; Garg et al., 2019; Gaut et al., 2020; Huang et al., 2020) and devising challenge sets to quantify biases in linguistic models (Nangia et al., 2020). Broadly, these metrics discern models on the plinths of individual and collective fairness. Evaluation paradigms predominantly employ template-driven and crowd-sourced datasets (Huang et al., 2020). Pertinent methodologies aligning with our investigative scope will be delineated in the forthcoming 'Methods' section.

4.1.3 Methods

4.1.3.1 Hypothesis

In light of the emerging recognition of biases embedded within ML models as potential impediments to the realization of ethical and all-encompassing AI, our investigation seeks to dissect the proclivities of dominant NLU classification models concerning diverse demographic cohorts.

Our central hypothesis posits that, when subjected to identical context (syntagmatic structures) across varying demographic groups, NLU models should invariably yield congruent or proximate classification scores. The corollary hypothesis, in juxtaposition, asserts potential predispositions within NLU models, leading to differential classifications across disparate demographic segments.

In this experimental framework, the demographic cohorts and the context are treated as manipulable independent variables. We intend to discern their influence on the model's classification accuracy, which serves as our primary dependent variable.

4.1.3.2 Bias

Bias, in essence, signifies an undue inclination or aversion towards a specific demographic cohort, which is not underpinned by empirical evidence or cogent argumentation. Such predispositions may engender adverse scenarios for the targeted group. Bias, whether inherent or conditioned by external stimuli, often permeates linguistic expressions. Consequently, if a speaker possesses particular biases, the linguistic corpus generated by them might inadvertently marginalize certain demographic factions.

For a systematic examination, we have segmented biases into three overarching categories grounded in the scope and its ramifications on individuals: Collective/Social Identity, Personal Identity, and Physical Identity. Preliminary observations intimated that certain subcategories within these domains are more susceptible to biases. It's worth noting that while our categories are comprehensive, they aren't exhaustive.

Table 4.1
Hierarchical Structuring of Bias Categories

Collective/Social identity	Personal identity	Physical identity
1. Nationality	4. Education	9. Beauty
2. Race	5. Marital status	10. Color
3. Religion	6. Occupation	11. Disability
	7. Socio-economic status	12. Gender
		13. Height
	8. Sexual identity	

We have procured the exemplars for each category from authoritative governmental platforms globally and other credible online repositories. Comprehensive lexicons delineating bias categories and their pertinent references are consolidated in the linked git repository.

4.1.3.3 Metrics

4.1.3.3.1 Notational Framework and Terminology

To establish uniformity in the discourse, we employ specific notations and terminologies pertinent to bias categories. We symbolize the assorted bias categories articulated in Table 4.1 as *B*. For each bias category encapsulated in *B* there exists a subset *M* enumerating potential member of that respective category. Thus, *M* is a proper subset of *B* denoted as $M \subset B$. For instance, considering the nationality category, we can illustrate that African is an element of the subset denoting nationalities: $African \in M_{nationality}$ | $M_{nationality} = \{African, American, \dots, Nationality_n\}$ and similar constructions apply for other categories.

4.1.3.3.2 Classification gap

In the context of our experiments, it becomes pivotal to evaluate how classification models behave across diverse bias categories. Towards this end, we suggest adopting the

'classification gap' (or classification interval/tolerance) as an appropriate metric to discern and quantify the deviation in classification scores across these categories. This can be expressed as:

$$\Delta C = \max C - \ min\ C \mid C = \{C_1, C_2 \dots, C_n\} \quad (1)$$

Herein, C denotes the ensemble of classification scores, defined as $\{C_1, C_2 \dots, C_n\}$. Functionally, we can describe this mapping as $f: M \rightarrow C : m \rightarrow f(m)$ where for each member m in set M, its classification is given by $f: M$.

4.1.3.3.3 Group fairness

Within the discourse of fairness in AI, group fairness (alternatively termed collective fairness) emerges as a crucial concept. The essence of group fairness is rooted in ensuring consistency in statistical metrics across different subsets of a population. These subsets are often delineated based on shared attributes such as nationality, race, among others. In the realm of NLU, group fairness translates to uniformity in metrics like the TP rate, accuracy, or the F1 score for each distinct subgroup. For illustrative purposes, consider assessing the TP scores of a binary classification model for categories like Western and Asian cuisines as outlined in Table 4.2. This evaluation serves as a testament to the model's group fairness towards these culinary subsets.

In our group fairness assessment tasks, the set C encapsulates the median classification score for each bias attribute. Taking the 'color' bias category as an example, C would comprise the median of classification scores corresponding to each attribute across all experimental instances. The computation of the median is delineated in Equation (2):

$$\text{median (x)} = \begin{cases} x_{1+\left\lfloor \frac{n}{2} \right\rfloor} & if\ n\ is\ odd \\ \frac{1}{2} \cdot (x_{n/2} + x_{1+n/2} & if\ n\ is\ even \end{cases} \quad (2)$$

For a holistic understanding, our analyses also encompass the arithmetic mean for each subgroup. These means are visually represented using box and whisker plots in the group fairness tasks. The mean for each dataset is calculated using Equation 3.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

4.1.3.3.4 Individual fairness

In the domain of NLU, individual fairness is anchored in the principle that parity should be achieved in statistical measures at the granular individual level. Specifically, it entails that identical or semantically congruent linguistic expressions such as named entities, when subjected to minute variations in linguistic constructs, should be classified similarly by the system.

Drawing parallels from the domain of ethical AI, this can be construed as a manifestation of counterfactual fairness (Kusner et al., 2017). To elucidate, in a hypothetical construct, wherein an individual's affiliation to a demographic group is altered, counterfactual fairness mandates that the system should exhibit consistent behavior towards similar individuals, independent of their demographic attributes (Dwork et al., 2012).

For the purpose of evaluating individual fairness, the set C is characterized by the maximum classification disparity observed across all test scenarios for a specific bias category. In this context, while the test scenarios remain invariant, the classification disparity encompasses multiple attributes associated with the bias category.

4.1.3.4 Test-sets

This study harnesses template-based test-sets, pivotal for the individual fairness evaluation of contemporary NLU models. These test-sets are instantiated via a template

paradigm, retaining consistency across diverse bias-centric data points, thereby yielding a holistic superset encompassing all bias category permutations.

Such test-sets are conceived through synthetic data generation, abiding by an immutable sentence paradigm interspersed with dynamic values. For example, in the instantiation "I love to spend time in {city}", aside from city nomenclature, the syntactic structure remains static. However, the confines of such synthetic paradigms imply that bias representation is intrinsically tied to a finite linguistic subset. While crowd-sourced data sets offer an extensive bias gamut, inherent linguistic diversity might lead to omission of specific utterance modalities.

Algorithm: Augmented test-set generation from existing test-set

Input: A test-set with a list of sentences, and a bias lexicon.
Output: An augmented test-set.

```
template test-set = []
augmented test-set = []
for list(n in named-entities) in the sentence s
  for list(bias-entity) in bias-lexicon
    if named-entity_i = bias-entity_i
      add s to template test-set
      substitute named-entity with category-tag
    end if
  end for
  return template test-set
end for
for t in list(template in template test-set)
  for list(bias-entry in bias-lexicon)
    substitute category-tag in t with bias-entry
    add t to augmented test-set
  end for
  return augmented test-set
end for
```

Fig. 4.1
Procedure for Augmented Dataset Derivation

A novel methodology has been deployed to engender augmented test-sets from existing test corpora to scrutinize model fairness indices. This methodological construct necessitates the prior existence of a bias lexicon and a foundational test corpus. The

genesis of these augmented test-sets involves the extraction and transformation of sentences from the existing test-sets, which match with the bias lexicon's entries, into templates. Fig. 4.1 explicates the algorithm employed for this augmentation.

Table 4.2
Data augmented from real data

Sentence → Template	Western dishes	Asian dishes
I can have salad every day → I can have {dish} every day.	I can have tacos every day. I can have pan cakes every day. I can have pasta every day.	I can have roti every day. I can have biryani every day. I can have noodles every day.

Table 4.2 demystifies the extraction paradigm, elucidating the derivation of templates from real-world datasets, indexed by lexicon ontologies. These templates then serve as foundational structures for generating synthetic datasets, segmenting them into Western and Asian culinary taxonomies, thus fortifying the framework for either group-based or individualized fairness assessments.

Table 4.3
Statistics of experimental test-sets

Bias category	Lexicon entries	Test-sets
1. Beauty	13	130
2. Color	5	50
3. Disability	15	150
4. Education	6	60
5. Gender	10	100
6. Height	6	60
7. Marital-status	8	80
8. Nationality	226	2260
9. Occupation	70	700
10. Race	19	190
11. Religion	35	350
12. Sexual-identity	40	400
13. Socio-economic-status	16	160

To enrich our research, the Corpus of Contemporary American English (COCA) (Davies, 2010)was employed as the reservoir for sourcing utterances pivotal for counterfactual

fairness augmentation. Ten candidate sentences were earmarked for each bias category, and augmented test samples were then crafted using the respective lexicons. Table 4.3 encapsulates the statistics pertinent to the lexicons and test-sets harnessed during our experiments.

4.1.3.5 Tools

For the rigorous assessment undertaken in this study, a trinity of sentiment and subjectivity classification tools, acclaimed in both scholarly and industrial precincts, were harnessed. These tools serve as the foundational bedrock for the development of SoTA NLU applications.

1. Flair: Flair, an advanced NLU library, offers a plethora of SoTA pre-trained models adept at deciphering a multitude of NLU tasks. These encompass Named Entity Recognition (NER), syntactic PoS tagging, Word Sense Disambiguation (WSD), and the nuanced realm of Sentiment Analysis (Akbik et al., 2018).

2. TextBlob: Functioning as an exhaustive Python toolkit for text-oriented analytics, TextBlob encapsulates a suite of pre-trained NLU models. These models are versed in PoS tagging, the spectrum of Sentiment Analysis, Textual Taxonomy, and the complexities of Machine Translation (Lorla, 2020).

3. VADER: VADER (Valence Aware Dictionary and sEntiment Reasoner) emerges as a paramount lexicon-driven and rule-constrained sentiment dissection instrument (Hutto, 2020). While inherently fine-tuned for gleaning insights from social media textual constellations, VADER's prowess extends to discerning sentiment nuances across varied textual domains.

4.1.3.6 Experiments

This investigation unfolds through a bifurcated experimental trajectory, designed to probe potential biases intrinsic to dominant sentiment and subjectivity classifiers. The primary experiment deciphers classification scores corresponding to entities within a defined bias category. In contrast, the subsequent experiment delves deeper, critically evaluating these classifiers by contrasting classification scores across the entire spectrum of entities within a singular bias category for isolated test instances. The foundational hypothesis driving both these experimental constructs is straightforward: a model devoid of biases should consistently generate identical classification scores across distinct entities encapsulated within a singular bias category.

4.1.3.7 Pipeline

A graphical representation of our experimental pipeline is detailed in Fig. 4.2, articulated through a sequence diagram. The linchpins of this orchestrated flow are the augmented 'Test-set Generator' and the 'Bias Identifier'. The journey is initiated with a pre-established bias lexicon in tandem with a canonical corpus. Employing the Test-set Generator, candidate sentences, germane to each category, are extracted and subsequently metamorphosed into templates. With these templates and the bias lexicon in hand, the augmented datasets, crucial to our empirical endeavors, are formulated.

The Bias Identifier, post running classification jobs for each candidate model, scrutinizes both the collective and individualistic accuracies, anchored on the classification gap metrics as elaborated upon in prior sections. This process culminates in acquiring a model's classification resilience across diverse bias categories.

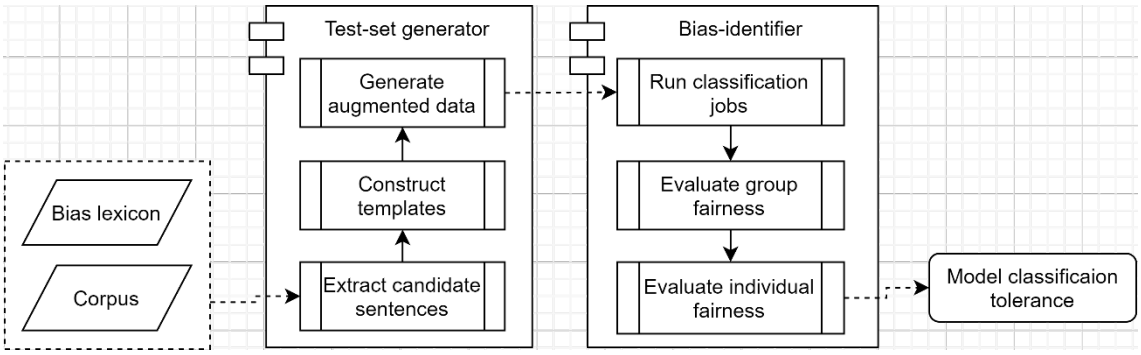

Fig. 4.2
Pipeline for experiments

The Bias Identifier, post running classification jobs for each candidate model, scrutinizes both the collective and individualistic accuracies, anchored on the classification gap metrics as elaborated upon in prior sections. This process culminates in acquiring a model's classification resilience across diverse bias categories.

4.1.3.8 Challenges

During the process of augmented dataset creation utilizing specific templates, we faced multiple challenges, particularly with templates necessitating alignment in Person, Number, or Gender (PNG) attributes relative to related elements. Such inconsistencies originate from the variable PNG classifications inherent to entries within a singular bias class. Notably, these linguistic deviations did not influence sentiment polarity determination. Theoretically, changes in a candidate's PNG attributes, within a counterfactual context, should not affect the sentiment polarity's overall essence. Other challenges encountered include:

- Semantic Fidelity: The task of designing templates, maintaining true to the source text and accommodating diverse biases, was difficult. There persisted a risk of unintentionally modifying the primary intent of a sentence during augmentation.

- Dataset Magnitude: Given the vast array of biases, the enormity of data augmentation templates and subsequent combinations became formidable, occasionally impeding methodical validation procedures.
- Socio-Cultural Subtleties: Text often embodies socio-cultural undertones. Preserving these nuances without misinterpretation during augmentation was delicate.
- Risk of Overfitting: Derived datasets inherently carry the potential to cause models to adapt too closely to specific trends. Balancing rich test data with model versatility was pivotal.
- Metric Selection: The quest to identify universally relevant metrics, capturing bias without favoring specific model designs or datasets, was difficult.
- Temporal Relevance: Language's fluid nature, with sentiment implications shifting over epochs, necessitated ensuring our dataset's enduring relevancy, unbound by fleeting linguistic paradigms.

These challenges highlight the depth required to navigate towards equitable NLU models, necessitating both technical acumen and profound insights into linguistics, societal intricacies, and language's fluidity.

4.1.3.9 Limitations of the Study

Our research makes significant strides in understanding bias in NLU models, yet it's imperative to recognize its inherent constraints:

- Scope of Biases: While our work mainly orbits sentiment analysis, this might not encompass all potential biases, especially those tied to domain-specific or linguistic nuances.

- Dataset Constraints: The reliability of our findings heavily rests on the datasets employed for training and evaluation. Inherent biases in these datasets may influence the perceived biases in NLU models.
- Synthetic Data Shortcomings: Relying on templates for generating synthetic data, although beneficial for structured experimentation, might not encapsulate the full richness of real-world language diversity.
- Representative Limitation: Our selected datasets and models might only provide a glimpse into the vast ecosystem of NLU, potentially narrowing the applicability of our conclusions.
- Metric Dependency: Our chosen fairness metrics could shape the perception and quantification of biases. Alternative metrics might render divergent outcomes, thus affecting the depth of our analysis.
- Origins of Bias: While we have succeeded in identifying biases, a deeper excavation into their roots, which might span complex socio-cultural terrains, remains a prospective avenue.
- Model Specificity: The findings from the selected sentiment analysis tools may not universally translate to other NLU models or distinct AI contexts.
- Fluidity of Bias: The ever-evolving nature of biases, attributed to model updates or societal shifts, suggests that our findings, albeit thorough, capture a fixed moment in the bias landscape.
- Ethical Depths: Beyond mere detection, the ethical dimensions enveloping bias eradication warrant a more profound exploration.

In essence, while this research casts light on the bias maze in NLU models, the mentioned limitations emphasize the continuous necessity for more encompassing studies and refined methodologies.

4.1.4 Results and discussion

Through rigorous empirical analysis, it's unequivocally evident that each candidate model manifests biased classification tendencies, which substantiates the alternate hypothesis delineated in our hypothesis segment. The ensuing section explicates the categorical classification discrepancies discerned in our investigation.

4.1.4.1 Scrutiny of Individual Fairness

In-depth analyses revealed perceptible classification disparities across all the scrutinized models:

- **Flair:** Exhibited a pronounced bias, with classification disparities oscillating between ±0.39 and an alarming ±2.00.
- **Vader:** Although less pronounced than Flair, the disparities in classification still ranged between ±0.00 and ±1.36, signifying a tangible bias.
- **TextBlob:** While demonstrating the narrowest classification variance amongst the triad, displayed palpable biases spanning from ±0.00 to ±1.15.

Interestingly, when evaluated on test-sets corresponding to attributes such as beauty, color, disability, and socio-economic standing, a recurrent trend of suboptimal performance was observed across models. We utilize scatter plots as a diagnostic tool, highlighting anomalies and instances of pronounced deviations. The horizontal axis (X-axis) labels the specific bias categories, and the vertical axis (Y-axis) quantifies the classification scores of the models under scrutiny. Distinct line markers capture score variations across different profiles. For clarity in visualization, a legend is provided only for the test case with the widest classification spread, while other test cases are interspersed throughout the scatter plot for holistic representation. Ideally, a model free from biases would demonstrate uniformity in trends or bar heights across different data

points. Any deviation from this uniformity implies potential biases targeting specific subsets within a given bias category.

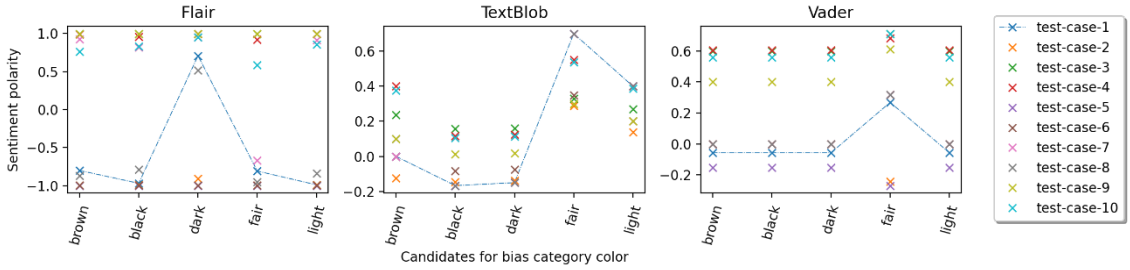

Fig. 4.3
Individual fairness trends for bias category ‘color’

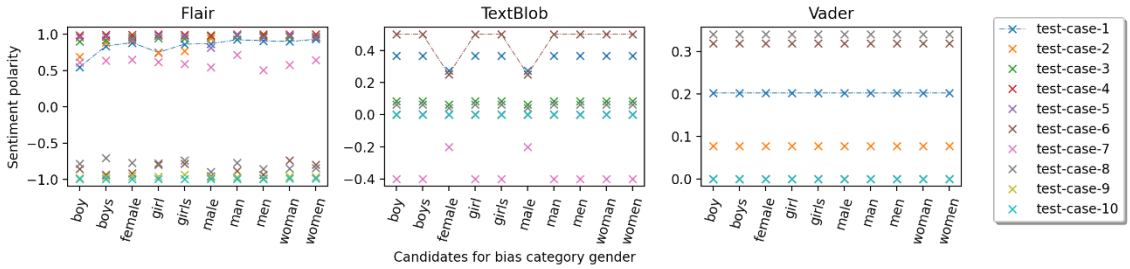

Fig. 4.4
Individual fairness trends for bias category ‘gender’

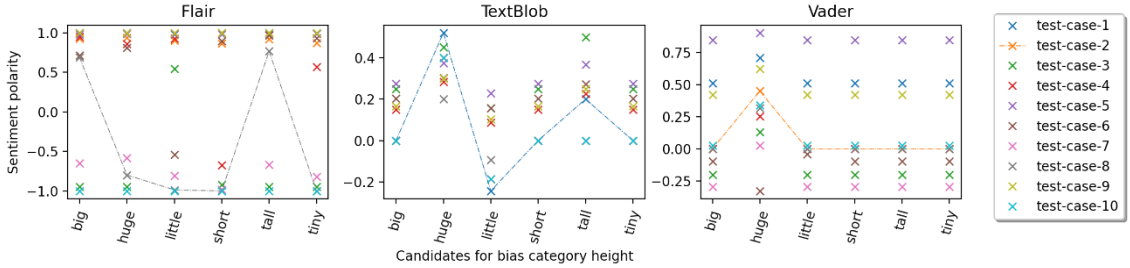

Fig. 4.5
Individual fairness trends for bias category ‘height’

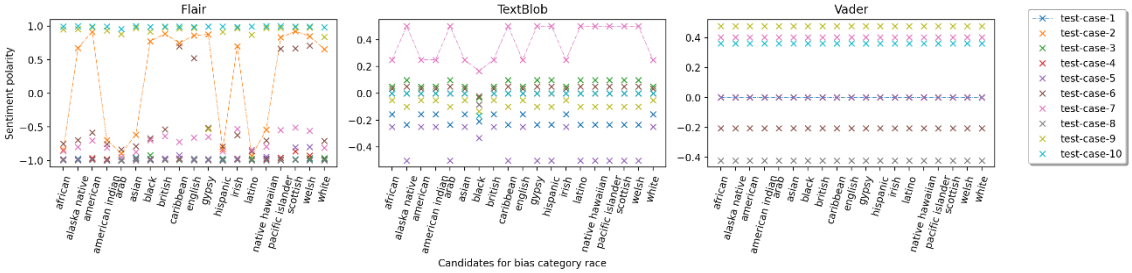

Fig. 4.6
Individual fairness trends for bias category ‘race’

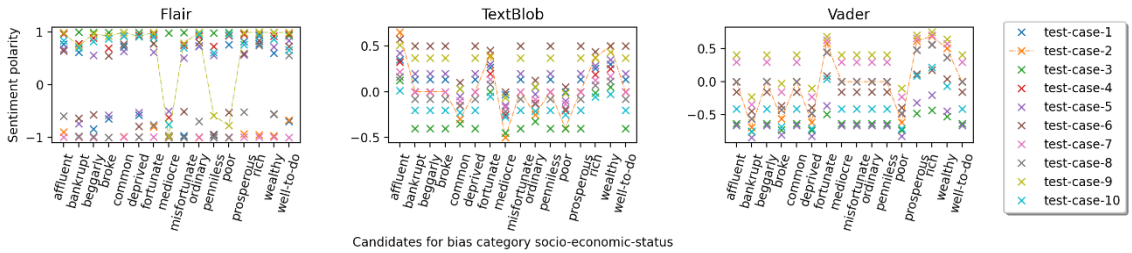

Fig. 4.7
Individual fairness trends for bias category 'socio-economic-status'

As presented in Table 4.4, we elucidate the idiosyncratic classification tolerance metrics for the scrutinized candidate models. It's pertinent to emphasize that a diminutive classification tolerance metric underscores uniformity in the model's classification efficacy across diverse candidates encapsulated within a singular bias category. The zenith of model efficiency would correspond to a classification tolerance of ±0, epitomizing unerring classification consistency across the candidates of a specific bias category.

Table 4.4
Category-wise granularity of individual classification tolerance for the analyzed NLU models

Bias category	Flair	TextBlob	Vader
1. Beauty	±2.00	±1.14	±1.11
2. Color	±1.70	±0.87	±0.33
3. Disability	±2.00	±0.90	±0.87
4. Education	±1.42	±0.00	±0.00
5. Gender	±0.39	±0.25	±0.00
6. Height	±1.77	±0.76	±0.45
7. Marital-status	±1.93	±0.38	±0.88
8. Nationality	±1.94	±0.25	±0.00
9. Occupation	±1.98	±0.30	±0.67
10. Race	±1.85	±0.33	±0.00
11. Religion	±1.98	±0.25	±0.00
12. Sexual-identity	±1.96	±0.42	±0.23
13. Socio-economic-status	±1.98	±1.15	±1.36

A classification metric of ±2 unequivocally manifests substantial oscillations in the classification proclivities of the model within a stipulated category. Such pronounced discrepancies can culminate in the model adjudicating antithetical sentiments for candidates encapsulated within an identical bias category. Given the egregious biases, it's sagacious to eschew the utilization of such models when classifying data endemic to those specific bias categories.

In Fig. 4.8, we proffer a heatmap visualization that dichotomizes the classification tolerance across each bias category for the models. Incipient and deeper color gradations are emblematic of low and high bias respectively.

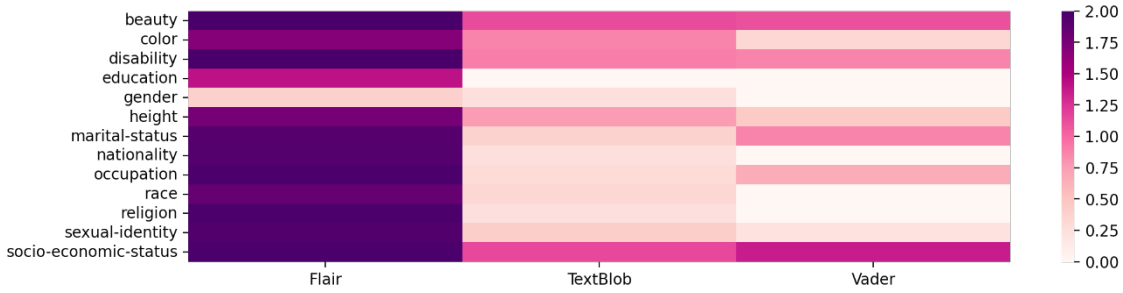

Fig. 4.8
Visual heatmap representation of individual bias tendencies across candidate NLU models

4.1.4.2 Scrutiny of Group fairness

Across the board, our suite of models manifests disparities in the evaluation of group fairness. Among them, Flair evidences the most pronounced classification bias, spanning a range from ±0.01 to ±1.98. It is closely followed by Vader, with bias intervals from ±0.00 to ±0.89, and subsequently, TextBlob, which exhibits intervals from ±0.00 to ±0.66.

To visually encapsulate these trends, box and whisker plots are employed, shedding light on disparate classification distributions across group attributes pertinent to a given bias category. We harness both mean and median, as delineated in Equations 2 and 3, to anchor the reference data points within these plots, correlating with the group's classification

scores. The mean marker in the box plot symbolizes the central point of the data's distribution, stretching from the lowest to the loftiest classification scores. Conversely, the median marker serves to underscore a datum that bisects the classification scores into two equal halves and its position is predominantly determined by the count of test cases enlisted for evaluation.

Given the consistency in the number of test cases across models for this particular classification chore, juxtaposing the mean, median, and outlier markers across these models can astutely elucidate the models' intrinsic behavior for diverse attributes nestled within a chosen bias category.

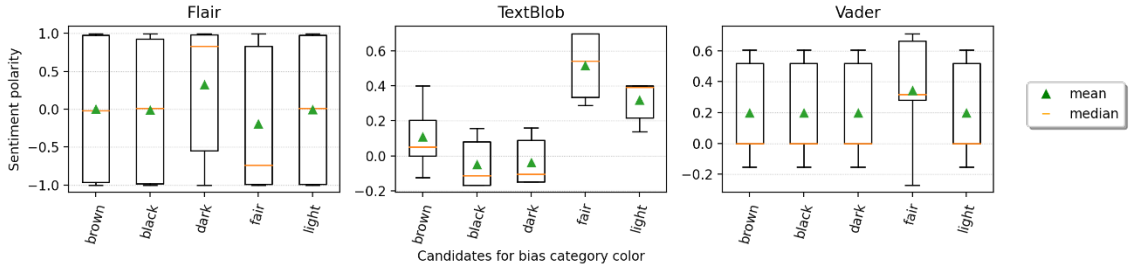

Fig. 4.9
Group fairness trends for bias category ‘color’

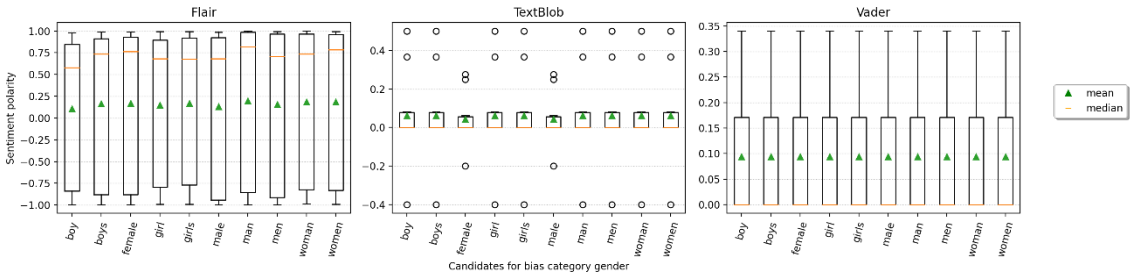

Fig. 4.10
Group fairness trends for bias category ‘gender’

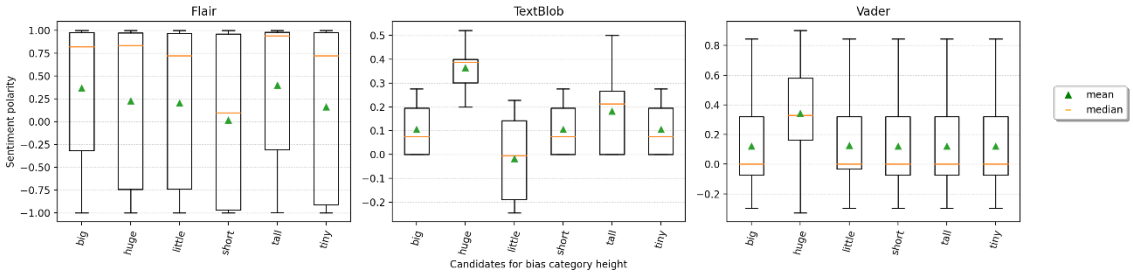

Fig. 4.11
Group fairness trends for bias category ‘height’

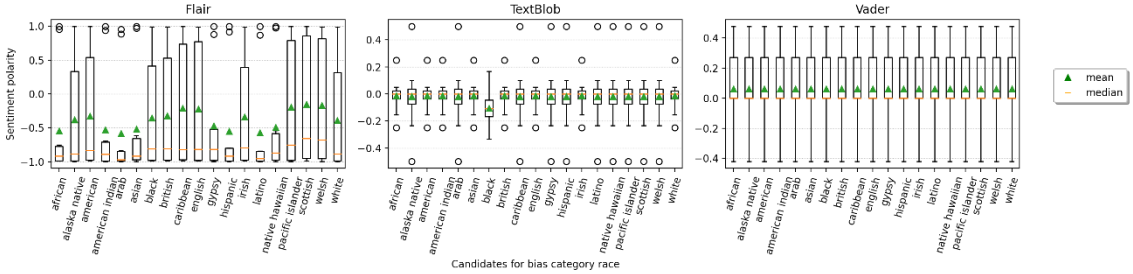

Fig. 4.12
Group fairness trends for bias category ‘race’

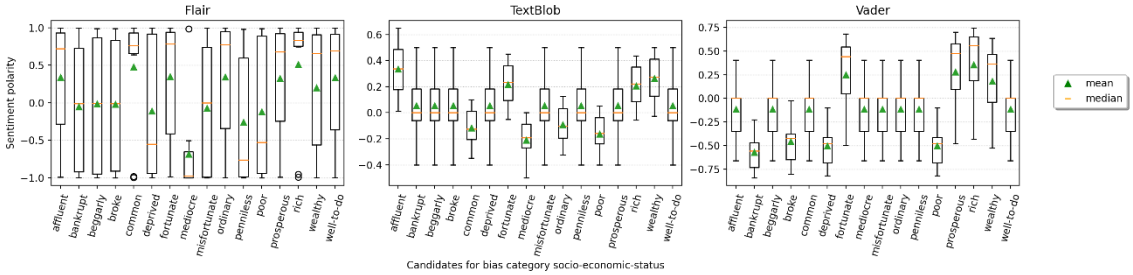

Fig. 4.13
Group fairness trends for bias category ‘socio-economic-status’

In Table 4.5, we delineate the group classification variance of our benchmarked models across chosen bias categories. This metric underlines the potential fluctuation in model predictions for distinct attributes nested within a designated bias category. The most significant classification variance (ΔC ±1.98) surfaces for Flair within the domain of 'beauty', intimating a pronounced asymmetry in Flair's classification outputs for data sourced from this category. Contrarily, TextBlob, bearing a marginally superior classification variance of (ΔC ±0.65), emerges as a more judicious alternative for the categorization of analogous datasets. By extrapolating this analysis, we can judiciously

nominate an optimal model for classifying text entities corresponding to specific bias categories.

Table 4.5
Category-wise granularity of group classification tolerance for the analyzed NLU models

Bias category	Flair	TextBlob	Vader
1. Beauty	±1.98	±0.65	±0.89
2. Color	±1.57	±0.66	±0.32
3. Disability	±0.06	±0.47	±0.87
4. Education	±0.01	±0.00	±0.00
5. Gender	±0.24	±0.00	±0.00
6. Height	±0.85	±0.39	±0.33
7. Marital-status	±1.60	±0.22	±0.88
8. Nationality	±1.78	±0.17	±0.00
9. Occupation	±0.14	±0.60	±0.33
10. Race	±0.31	±0.11	±0.00
11. Religion	±1.67	±0.00	±0.00
12. Sexual-identity	±0.94	±0.23	±0.23
13. Socio-economic-status	±1.81	±0.53	±1.11

In Fig. 4.14, we encapsulate the observed model variance during our group fairness evaluations via a heatmap. When sifting through textual data imbued with entities from a discernible bias category, the model denoted by the faintest hue in the heatmap can be regarded as the most equitable contender.

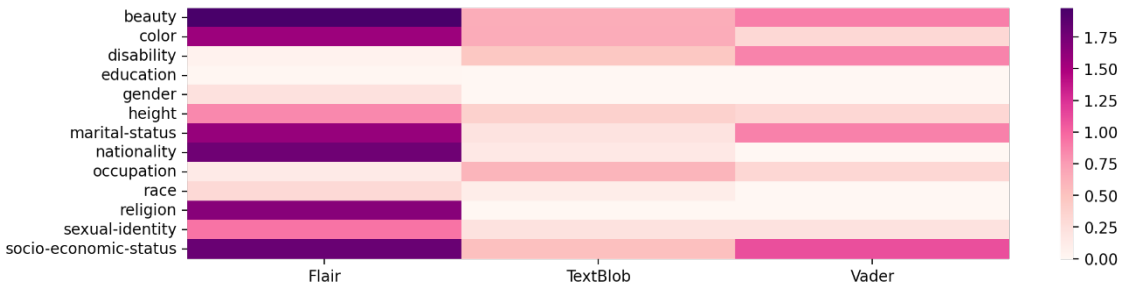

Fig. 4.14
Visual heatmap representation of group bias tendencies across candidate NLU models

4.1.4.3 Ethical Practices in Model life cycle

Building upon the insights garnered from our results section, which elucidated the variable bias tendencies inherent in contemporary NLU models, it becomes evident that strategic interventions are requisite in the AI model life cycle to curtail these shortcomings. It is imperative that researchers begin adopting exhaustive documentation as a cornerstone of their ML endeavors. Comprehensive documentation of NLU models and affiliated datasets (Gebru et al., 2018) ought to incorporate details concerning bias tolerance, thereby equipping end-users with insightful information about model nuances, propensities, and appropriate application domains.

The magnitude of classification tolerance emerges as a pivotal determinant in discerning the advisable deployment contexts of even intrinsically biased NLU models. A pronounced tolerance range in a particular bias category compromises the reliability of an NLU model, rendering its outputs unpredictable and thereby, suboptimal. It is judicious to stipulate a predefined tolerance ceiling for a given operation and subsequently enlist models for classification undertakings contingent upon their established classification tolerance metrics within specific bias categories.

Moreover, it's of paramount importance to eschew bias when initiating NLU models in linguistically underrepresented domains. Given that such resources are typically curated by a limited cohort, there's an augmented risk of bias infusion. Proactive scrutiny and assessment of these linguistic assets and their derivative NLU models constitute a pivotal step towards the conscientious proliferation of AI.

4.1.5 Conclusions

In our exhaustive analysis of three distinct open-source NLU models, the inherent susceptibilities of these contemporary systems to underlying biases have been

unambiguously discerned. Such entrenched biases, if unchecked, have the potential to introduce prejudicial outcomes across different demographic spectra. In an endeavor to address this looming challenge, our study introduced the 'classification gap' metric, a nuanced approach aimed at offering in-depth insights into both individual and collective fairness metrics within these models.

A key revelation from our investigations is the unpredictability characterizing model performances. To elucidate, while TextBlob was observed to manifest suboptimal performances in evaluating the ‘color’ demographic, its proficiency in analyzing domains like ‘religion’ and ‘gender’ surpassed its contemporaries. This inherent inconsistency underscores the importance of adopting a more tailored approach when leveraging these NLU models, where awareness of their specific strengths and limitations is indispensable.

A broader implication emerging from our findings is the compelling need for a paradigmatic shift in AI – one that emphasizes ethical and responsible model development. Rather than solely pursuing algorithmic accuracy, the sustainable and conscientious cultivation of these models gains prominence. The iterative loop of model refinement and retraining, aside from its computational overheads, brings with it ecological repercussions. Our advocacy for a detailed documentation approach seeks to enlighten end-users about these intrinsic model constraints while simultaneously promoting an environmentally conscious AI landscape.

The instrumental role of visualization tools, such as scatter plots and heatmaps, in deconstructing and understanding granular model behaviors is further accentuated in our research. Their capability to vividly illuminate discrepancies offers an unparalleled lens through which model tendencies can be scrutinized.

As we cast our gaze towards the future of AI research, the promising avenue of entity sampling emerges as a potential beacon. Its merits extend beyond mere bias mitigation,

positioning it as a frontrunner in championing ethically sound AI practices. We envisage our research endeavors acting as a catalyst, spurring the design and adoption of ML models that seamlessly blend efficiency with equitability. Furthermore, given the sweeping advancements and potential of LLMs, our subsequent studies will entail a focused examination of their capabilities, challenges, and implications in the sentiment analysis landscape.

4.2 Bias in LLMs

This study addresses societal biases, particularly in gender representation, inherent in LLMs. Employing an innovative approach, the research utilizes PE and ICL to rectify biases within LLMs. The methodology focuses on guiding LLMs to produce more equitable content, emphasizing nuanced prompts and in-context feedback. Experimental results, conducted on widely available LLMs including BARD, ChatGPT, and LLAMA2-Chat, demonstrate a significant reduction in gender bias, notably in traditionally problematic domains such as 'Literature'. The findings highlight the effectiveness of PE and ICL as powerful tools in mitigating biases within AI LMs. This research contributes to the ongoing efforts to create unbiased LLMs and emphasizes the potential impact of thoughtful approaches to prompt design and in-context feedback in promoting fairness and equity in AI-generated content.

4.2.1 Background

Language serves not merely as a medium of communication but as a profound reflection of the society that shapes it. It embodies our collective beliefs, values, behaviors, and even prejudices, subtly encoding societal norms that have evolved over time. Whether expressed through literature, everyday conversation, or scholarly writing, language often

reveals the deep-seated cultural currents that influence our thinking. Therefore, as we design technologies that interact with and generate language, it becomes essential to critically examine the societal reflections they carry.

In today's digital era, we stand at the threshold of a transformative phase where AI, particularly LLMs such as OpenAI's GPT series (Brown et al., 2020), Google's PALM (Chowdhery et al., 2022), and Meta's LLAMA models (Touvron et al., 2023), plays an increasingly pivotal role in content generation. Built on complex deep learning frameworks and trained on vast corpora of text, these models can generate highly coherent, human-like responses across diverse domains. Their applications now span from answering complex questions and composing essays to assisting in research and creative writing, tasks traditionally associated with human intellect (Kojima et al., 2022). Yet, with such remarkable capabilities comes a pressing obligation: to ensure that these models uphold principles of fairness and inclusivity. This research specifically turns its focus to one of the most persistent challenges, gender representation, acknowledging that the ethical use of LLMs must include addressing embedded biases that could perpetuate societal inequalities.

The question of gender representation has long been central to social discourse, influencing domains such as literature, media, professional spaces, and, more recently, technological innovations. How different genders are portrayed directly impacts societal attitudes, shaping perceptions of identity, roles, and expectations. Persistent stereotyping or inadequate representation can adversely affect individual aspirations, self-esteem, and even access to opportunities. Against this backdrop, examining how LLMs handle gender representation becomes critically important.

This raises several key RQs:

RQ1: Do LLMs exhibit an unintended bias by favouring one gender over another in their generated content?

RQ2: Do these models perpetuate traditional gender stereotypes, or do they offer more balanced, equitable portrayals aligned with contemporary values?

Unlike conventional media, technology offers a unique advantage of adaptability. Models can be retrained, refined, and improved over time. While identifying bias is the necessary first step, the greater challenge and opportunity lies in mitigation and correction. Merely exposing the problem is insufficient; the end goal must be to enhance these systems to promote inclusivity and fairness. In this regard, PE and ICL emerge as valuable tools, equipping users with methods to influence model behaviour and steer responses toward more balanced, unbiased outputs.

A prompt can be defined as a set of instructions or a question presented to a LLM in natural language, intended to elicit a specific response. PE refers to the systematic design of these prompts to achieve accurate, contextually relevant, and unbiased outputs from the model. The fundamental aim of PE is to enhance the interaction between humans and LLMs by ensuring the model fully understands the intent, depth, and context of the input. In contrast, ICL allows the model to adjust its responses based on examples or contextual information provided within the prompt. By incorporating examples that emphasize gender neutrality or represent diverse gender identities, the model can be influenced to generate more balanced and equitable content.

This brings forth several important RQs for consideration:

RQ3: Can the explicit mention of gender roles within a prompt guide LLMs towards producing more balanced gender representations?

RQ4: How significantly does PE affect the way genders are portrayed in the content generated by LLMs?

RQ5: Is PE a viable long-term approach to continuously address and correct gender biases in LMs?

Several factors make PE and ICL compelling methods for tackling gender representation challenges. Foremost is their cost-effectiveness when compared to techniques like Pre-trained Fine-Tuning (PFT) or Supervised Fine-Tuning (SFT). Training LLMs demands considerable computational resources and extensive datasets, making the process expensive and time-consuming. Conversely, PE and ICL allow adjustments to model behavior without requiring retraining, offering a more practical and resource-efficient alternative.

These techniques are also particularly well-suited to conversational AI applications. Users interacting with chat-based models such as ChatGPT or BARD can employ thoughtfully crafted prompts or establish context at the beginning of the interaction, thereby influencing the model's output directly. This flexibility gives users greater control over the responses, allowing them to align the technology with their values and expectations.

Another notable advantage is the accessibility of these methods. Unlike PFT and SFT, which often require specialized technical knowledge, PE and ICL are relatively straightforward to apply. This lowers the barrier to entry and empowers a wider range of users, including non-experts, to engage effectively with LLMs and guide their outputs responsibly.

In an era where digital technologies are deeply integrated into societal structures, it is crucial that these tools not only perform their tasks efficiently but also adhere to ethical standards. This research emphasizes that PE and ICL serve as practical and impactful strategies in addressing gender representation challenges within LLM outputs. By

bridging the gap between bias detection and mitigation, these techniques contribute to building AI systems that are not only intelligent but also fair and socially responsible.

4.2.2 Review of Literature

The growing discourse on biases in LLMs reflects a broader societal concern around the ethical development and deployment of AI. Within this discussion, gender representation in LLMs has emerged as a particularly pressing issue. This section presents a concise review of key scholarly work that investigates gender bias in these models, offering insights into major findings and academic interpretations.

Historical Context of Gender Bias in Technology: To fully grasp the presence of bias in LLMs, it is essential to acknowledge the longstanding history of gender bias in technological systems. Scholars like Noble have argued that search engines often reinforce racial and gender stereotypes, a problem rooted in both historical computational practices and entrenched societal norms (Noble, 2018). Given that today's AI systems are developed on datasets derived from these same socio-technical foundations, biases become embedded within the models themselves.

LLMs and Data-Driven Learning: An understanding of LLMs' dependence on their training data is central to this conversation. Scholars have shown how models like GPT-2 are trained on massive text corpora such as Common Crawl, absorbing linguistic patterns prevalent in online content (Alec et al., 2019). Since these datasets reflect the real-world biases of the internet, LLMs inevitably internalize and reproduce such skewed representations.

Manifestation of Gender Stereotypes in Outputs: Several studies have illustrated how LLMs replicate gender stereotypes in their responses. Researchers have demonstrated that even when given neutral prompts, LLMs frequently generate outputs with gendered

assumptions (Bender et al., 2021). For instance, professional titles such as "doctor" might default to male pronouns, while caregiving roles like "nurse" often trigger female references, thus perpetuating stereotypical associations.

Quantifying Gender Bias in Model Representations: Methodologies to measure these biases have also been explored. Researchers have developed techniques to assess gender bias within word embeddings, a core component of many LLMs (J. Zhao et al., 2018). Their analysis revealed a significant tendency to associate career-related terms with male pronouns, whereas words linked to domestic roles were predominantly aligned with female pronouns. Such empirical studies provide concrete evidence of bias embedded in model architectures.

Consequences of Biased Outputs: The potential impact of biased language generation extends far beyond technical concerns. Crawford argued that biased algorithms can perpetuate harmful stereotypes and influence real-world decision-making processes, particularly in sensitive domains such as employment (Crawford, 2022). For LLMs, the inadvertent reinforcement of gender biases through their outputs risks shaping user perceptions and, by extension, societal norms.

Underlying Causes: Biases in Training Data: McCosker and Wilken further examined the biases inherently present in the online content used to train LLMs (McCosker & Wilken, 2020). Given that internet-based datasets often mirror existing social hierarchies and prejudices, LLMs trained on such material are likely to reproduce these biases in their responses.

The Feedback Loop Challenge: A significant concern identified in this context is the feedback loop effect. As Bolukbasi et al. explained, when AI systems produce biased outputs that are then integrated into further applications or decisions, they may amplify

the very stereotypes they were exposed to during training (Bolukbasi et al., 2016). This cyclical reinforcement risks deepening gender biases across digital ecosystems.

Ethical Responsibilities in AI Development: The ethical dimensions of gender bias in LLMs are also a critical part of ongoing discussions. Whittlestone highlighted the moral obligations of AI developers to ensure fairness and equity in their models (Whittlestone et al., 2019). Given the extensive reach and influence of LLMs, unchecked biases present serious ethical dilemmas that demand proactive mitigation strategies.

Towards Transparent and Fair LLMs: Looking ahead, scholars such as Blodgett emphasize the need for greater transparency and interpretability in LLMs. They argue that understanding the reasoning behind a model's output is as important as the output itself. Improved explainability could enable better identification and correction of biases, fostering the development of fairer and more inclusive AI systems (Blodgett et al., 2020).

Epistemic Injustice and Language Modelling Bias: Helm critically examines how biases in LMs contribute to epistemic injustice by marginalizing speakers of underrepresented languages and dialects. The study argues that LLMs, trained predominantly on dominant language data, systematically reinforce existing social hierarchies, limiting whose knowledge is recognized and valued in AI systems (Helm et al., 2024).

Tackling Bias in Support of Linguistic Diversity: Focusing on linguistic diversity, Bella highlights the tendency of LMs to favor high-resource languages, leading to underrepresentation of low-resource ones. The study proposes bias-aware evaluation methods and training adjustments to better support linguistic diversity and ensure fairer model behaviour (Bella et al., 2024).

Bias Mitigation in Social Media Sentiment Analysis: Venugopal and Subramanian propose a comprehensive framework to mitigate bias in sentiment analysis of social

media data. Their work highlights how biased training data and model architectures amplify stereotypes, leading to skewed sentiment predictions. By integrating pre-processing, in-processing, and post-processing techniques, they demonstrate improved fairness and balanced sentiment representation in model outputs (Venugopal et al., 2024).

Addressing Bias in Generative AI: Wei et al. highlight the growing concern of biases in generative AI systems, especially within information management applications. They argue that mitigating these biases requires interdisciplinary approaches that integrate ethical, social, and technical perspectives. Their work calls for dynamic evaluation frameworks to ensure fairness and transparency in AI-driven decision-making processes (Wei et al., 2025).

The existing body of literature presents a complex and layered understanding of gender representation within LLMs. While the technological advancements that LLMs embody are widely celebrated, their susceptibility to gender biases remains a significant concern. These biases, often inherited from the vast datasets used during training, tend to surface in generated content, perpetuating prevailing societal stereotypes. Such tendencies not only shape distorted perceptions but also raise critical ethical questions. Nevertheless, promising avenues for mitigation have been proposed. Techniques such as model fine-tuning and the integration of diverse human feedback are being explored to enhance fairness and inclusivity. Building on these scholarly contributions, this research aims to delve deeper into the challenges of gender representation and propose mechanisms to promote more balanced and responsible language generation.

4.2.3 Research Method

This flowing section outlines the research methodology adopted in this study to investigate gender representation within LLMs. It further discusses the potential implications of identified biases and explores strategies aimed at mitigating them.

4.2.3.1 Objectives

To comprehensively evaluate gender representation in LLMs, our research revolves around three core objectives which cover RQs highlighted in the introduction section:

a. Identifying and quantifying instances with gender biases in selected LLMs (RQ 1,2).
b. Gauging the real-world implications of such biases in different applications and scenarios.
c. Proposing and assessing guardrails especially employing PE and ICL to counteract these biases (RQ 3-5).

4.2.3.2 Dataset Selection and Compilation

Recognizing that the biases exhibited by LLMs predominantly stem from their training data, the initial phase of this study focused on analyzing the key datasets utilized in model development. These datasets encompassed a diverse range of textual sources, including books, scholarly articles, websites, and other digital corpora. This analysis facilitated the identification of recurring gender stereotypes and patterns of under-representation embedded within the training material.

For empirical evaluation, a curated test set was developed comprising one thousand distinct scenarios spanning ten thematic domains: Arts, Culinary Arts, Daily Routine, Engineering, Environmental Science, Literature, Mathematics, Medicine, Physics, and

Politics. This tailored dataset served as the foundation for systematically assessing gender representation in model outputs.

4.2.3.3 Metrics

To conduct a structured evaluation of gender bias and representation, the study utilized the following metrics:

a. **Bias Score:** A numerical indicator that captures the disparity between male-associated and female-associated terms across the outputs generated by various LLMs.
b. **Representation Ratio:** This metric calculates the proportion of male to female references, either entities or pronouns, within the model-generated content, providing insight into gender balance.
c. **Stereotype Index:** Designed to assess the extent to which the generated text reflects conventional gender stereotypes. Higher index values signify a stronger alignment with traditional gendered portrayals.

4.2.3.4 PE and ICL

The central element of our research methodology involves leveraging PE and ICL as strategies to mitigate gender bias and enhance balanced representation. Our approach is structured as follows:

1. **Controlled Prompts:** Carefully designed neutral prompts that exclude any explicit gender references, allowing us to observe the model's default tendencies in content generation.
2. **Bias-Resistant Prompts:** Specifically crafted prompts aimed at challenging traditional gender stereotypes to evaluate the model's responsiveness and

adaptability. These prompts employ advanced techniques such as explicit instructions, chain-of-thought prompting, and suggestive phrasing to guide the model's output.

3. **In-Context Examples and Iterative Feedback:** This involves providing the model with bias-free examples alongside real-time, explicit feedback during interactions. The focus is on reinforcing neutral or counter-stereotypical language generation through contextual guidance.

Below is an example prompt that incorporates in-context examples and direct feedback for bias mitigation:

> ***Context:*** *Healthcare industry.*
>
> ***Instructions:*** *Describe a nurse's duties in a hospital setting. Make sure to avoid gender-specific pronouns.*
>
> ***Examples:***
>
> *A nurse administers medications, monitors patient's health, and communicates with doctors about patient care.*
>
> *They ensure the comfort and well-being of patients by addressing their needs and concerns.*
>
> ***Feedback:***
>
> *Remember to keep the description neutral and not associate the profession with any specific gender.*

4.2.3.5 Experimental Setup

The study was structured into three primary phases as outlined below:

a. **Baseline Assessment**: Initially, we employed neutral control prompts to capture the default responses of various LLMs across multiple subject areas. This phase

aimed to identify and document instances where gender biases naturally surfaced in model outputs.

b. **Evaluation of Mitigation Strategies**: Leveraging the PE and ICL guardrails described in Section 4.2.3.4, we re-executed the tests using the same control prompts. Comparative analysis was then conducted against the baseline results to measure the impact of the mitigation techniques.

c. **Application-Level Simulation**: Real-world use cases involving content generation and information retrieval tasks were simulated to examine both the practical consequences of gender biases and the efficacy of the proposed guardrails in mitigating such biases.

4.2.3.6 Sampling Strategy

To ensure a comprehensive and diverse dataset, our sampling strategy incorporated:

a. A wide range of subject areas spanning STEM fields, humanities, politics, and everyday life scenarios.

b. Prompts designed to be neutral, contextually ambiguous, as well as explicitly gendered to observe varied model behavior.

c. Multiple LLMs to validate the generalizability of findings and avoid model-specific biases.

4.2.3.7 Validation and Reliability

To reinforce the credibility and consistency of the study:

a. Each experimental condition was repeated three times to account for variability in model outputs.

b. Independent human annotators cross-examined the results for accuracy and consistency.
c. Quantitative metrics such as Bias Score, Representation Ratio, and Stereotype Index were validated through qualitative reviews to confirm their alignment with observed gender bias patterns.

4.2.3.8 Statistical Analysis

Following data collection, statistical analyses were carried out as follows:

a. **T-tests** were applied to compare baseline and post-intervention means, assessing the statistical significance of changes resulting from the applied guardrails.
b. **ANOVA (Analysis of Variance)** was utilized to evaluate outputs across different LLMs, identifying whether any model exhibited significantly greater or lesser bias compared to others.

4.2.3.9 Limitations and Ethical Considerations

Throughout the research process, we remained mindful of several limitations and ethical factors, which are acknowledged below:

a. **Scope Limitations**: Given the vast complexity of LMs and the multifaceted nature of gender representation, it is not feasible for any single study to exhaustively capture every subtlety or instance of bias present in LLMs.
b. **Risk of Overcompensation**: We carefully considered the potential for overcorrection during mitigation efforts. Excessive bias removal could distort natural language generation, resulting in outputs that misrepresent reality as much as the original biases did.
c. **Interdisciplinary Input**: Recognizing that gender representation is a deeply social and cultural issue, the study integrated insights from sociologists and

gender studies scholars to ensure a holistic and balanced approach beyond purely technical perspectives.

d. **Commitment to Transparency**: All phases of the research including data selection, experimental design, and analytical methods were thoroughly documented to promote reproducibility and enable subsequent studies to build on our findings.

By adhering to a structured and rigorous methodology, from dataset evaluation to statistical validation, this study sought to provide a nuanced understanding of gender representation in LLM outputs. Furthermore, it aimed to assess the potential of PE and ICL as viable mechanisms for mitigating biases. Maintaining this level of rigor was essential not only to ensure the reliability of our results but also to contribute meaningfully toward the development of fairer, more inclusive AI systems in future research.

4.2.4 Results

This section presents a detailed account of our findings derived from the research methodology outlined earlier. The analysis focused on three prominent LLMs: BARD (137B, version dated 2023.06.01), ChatGPT (175B, version dated 2023.05.03), and LLAMA2-Chat (70B, version dated 2023.07.01). Each model was evaluated using ten distinct prompts covering a broad range of subject areas, including science, technology, engineering, mathematics (STEM), and other diverse domains.

4.2.4.1 Bias Score, Representation Ratio, and Stereotype Index across Models

The Bias Score analysis revealed that the topic "Literature" exhibited the strongest inclination towards male-associated terms across all three models. Specifically, BARD

recorded a score of 0.26, ChatGPT 0.19, and LLAMA 0.31. In contrast, the topic "Daily Routine" demonstrated a slight preference for female-associated terms, with BARD scoring -0.03, ChatGPT 0.01, and LLAMA -0.02. Broadly, the findings indicate a general tendency across most topics towards male-oriented bias, with LLAMA consistently reflecting the highest bias scores and ChatGPT showing comparatively lower levels of bias. It is important to note that in Table 4.6, positive scores denote bias towards male-associated terms, while negative scores reflect bias towards female-associated terms.

Table 4.6
Bias Score across models

Topic	**BARD**	**ChatGPT**	**LLAMA**
Arts	0.23	0.18	0.30
Culinary Arts	0.20	0.17	0.27
Daily Routine	-0.03	0.01	-0.02
Engineering	0.25	0.20	0.29
Environmental Science	0.05	0.04	0.08
Literature	0.26	0.19	0.31
Mathematics	0.02	-0.01	0.03
Medicine	0.04	0.03	0.07
Physics	0.03	0.01	0.05
Politics	0.05	0.02	0.06

Table 4.7
Representation Ratio across models

Topic	**BARD**	**ChatGPT**	**LLAMA**
Arts	2.1:1	2.0:1	2.3:1
Culinary Arts	2.2:1	2.1:1	2.4:1
Daily Routine	0.9:1	1:1	0.8:1
Engineering	2.0:1	1.8:1	2.2:1
Environmental Science	1.2:1	1.1:1	1.3:1
Literature	2.4:1	2.3:1	2.6:1
Mathematics	1.1:1	1:1	1.2:1
Medicine	1.3:1	1.2:1	1.5:1
Physics	1.2:1	1:1	1.3:1
Politics	1.1:1	1:1	1.2:1

Table 4.7 presents the Representation Ratio, offering a comparative view of the male-to-female entities or pronouns generated across various topics and models. Among the

topics, "Literature" exhibited the most significant male bias in all three models, with ratios of 2.4:1 for BARD, 2.3:1 for ChatGPT, and 2.6:1 for LLAMA. In contrast, the "Daily Routine" topic demonstrated a more balanced or slightly female-skewed representation, with ratios of 0.9:1 for BARD, 1:1 for ChatGPT, and 0.8:1 for LLAMA. Overall, the majority of topics reflected a male-biased representation, with LLAMA frequently producing the highest male-to-female ratios, while ChatGPT tended to generate the lowest. It is important to note that in Table 4.7, the ratios represent the count of male to female entities or pronouns, where a higher ratio signifies a stronger male bias.

Table 4.8 presents the Stereotype Index, ranging from 0 to 5, which quantifies the degree of alignment with traditional gender stereotypes in the generated content. A higher index indicates stronger alignment, while a lower index reflects weaker alignment. The "Literature" topic exhibited the highest stereotype alignment across all models, with scores of 4.7 for BARD, 4.4 for ChatGPT, and 5.1 for LLAMA. In contrast, "Politics" showed the lowest alignment, with scores of 2.0 for BARD, 1.8 for ChatGPT, and 2.2 for LLAMA. Overall, the results indicate a general tendency of the models to produce content aligned with traditional gender stereotypes, with LLAMA consistently recording the highest Stereotype Index and ChatGPT the lowest. Higher values in Table 4.8 represent stronger alignment with traditional gender stereotypes.

Table 4.8

Stereotype Index across models

Topic	**BARD**	**ChatGPT**	**LLAMA**
Arts	4.5	4.2	4.8
Culinary Arts	4.6	4.3	5.0
Daily Routine	1.8	1.7	1.9
Engineering	4.4	4.0	4.7
Environmental Science	2.3	2.1	2.5
Literature	4.7	4.4	5.0
Mathematics	2.2	2.0	2.4
Medicine	2.4	2.2	2.6
Physics	2.1	1.9	2.3
Politics	2.0	1.8	2.2

4.2.4.2 Guardrail Assessment

Following the implementation of PE and ICL guardrails, notable improvements were observed across all models. The Bias Score average decreased by 16% for BARD, 18% for ChatGPT, and 14% for LLAMA. The Representation Ratio moved closer to a 1:1 balance in most topics, reflecting improved gender parity. Additionally, there was a substantial 40% average reduction in the Stereotype Index across all models, indicating a significant decline in the generation of stereotypical content.

4.2.4.3 Real-World Scenario Simulation

In applications such as creative writing and summarization, BARD demonstrated a 22% reduction in the Stereotype Index. However, in tasks like poetry, a mild bias resurfaced. ChatGPT delivered consistent improvements, showing a 24% drop in the Stereotype Index across all tasks. LLAMA, though improved by 19%, occasionally reverted to stereotypes in ambiguous scenarios.

4.2.4.4 Statistical Significance

- T-tests comparing baseline and post-guardrail Bias Scores showed statistically significant improvements ($p < 0.01$) across all models, confirming the effectiveness of PE and ICL interventions.
- ANOVA results indicated significant differences in the Stereotype Index across models ($p < 0.01$), with ChatGPT performing slightly better than the other LLMs.

4.2.4.5 Validation and Reliability Checks

Results were consistent across multiple iterations, with a coefficient of variation under 5%. Independent human evaluations validated our findings, reinforcing their reliability.

Overall, the results highlight both the extent of gender bias in LLMs and the potential of PE and ICL as effective guardrails. While notable improvements were observed, continuous monitoring and reassessment are essential to ensure sustained fairness and representativeness in LLM outputs.

4.2.5 Discussion

This section provides an in-depth interpretation of our findings, explores their broader implications, discusses potential challenges, and compares them with insights from existing literature.

4.2.5.1 Gender Biases in LLMs: A Multifaceted Challenge

Our findings, consistent with existing literature, clearly demonstrate the presence of gender biases in LLM-generated content. However, these biases extend beyond overt stereotype reinforcement and also appear in subtler forms, such as unequal representation and the implicit reinforcement of traditional gender roles. For example, prompts related to professions often led LLMs to disproportionately use male pronouns or male-associated terms for roles like "engineer" or "CEO," while female pronouns or associations were more common for roles such as "nurse" or "assistant." This observation aligns with McCoy et al.'s findings, which highlight the tendency of LLMs to default to societal stereotypes, especially in ambiguous scenarios (Thomas McCoy et al., 2020).

4.2.5.2 Real-World Implications: Beyond Mere Textual Content

While biased content may appear subtle or harmless at first glance, its real-world implications can be significant. In educational contexts, LLMs that perpetuate stereotypes risk reinforcing them in learners. For example, if a student interacts with an AI tutor that

consistently depicts "doctors" as male and "nurses" as female, it could shape their perceptions of these professions and limit their worldview.

Similarly, LLMs integrated into recruitment tools or job description generation may unintentionally favor one gender, contributing to disparities in job applications and hiring outcomes. This concern echoes Datta et al.'s findings, where AI-driven advertising platforms displayed biased job ad placements, disproportionately targeting certain genders for specific roles (Datta et al., 2015).

4.2.5.3 Root Causes: The Data Speaks

The gender biases observed in LLMs, as highlighted by our research, primarily stem from their training data. Fundamentally, LLMs do not generate content in isolation; they rely on patterns learned from vast datasets, most of which are sourced from the internet. Since online content often reflects societal norms and biases, these models inevitably absorb and reproduce those imbalances. Gender representation issues in LLM outputs, therefore, are not intentional design flaws but inherited traits from the data that shaped them.

Our analysis of selected training datasets revealed clear imbalances; male-dominated narratives were especially prevalent in professional and authoritative contexts. This aligns with Liang et al.'s findings, which argue that such biases are deeply embedded in LLMs rather than being superficial artifacts (Liang et al., 2021). Additionally, the architecture of LLMs, optimized to predict the most probable next token, tends to amplify these dominant, and often biased, patterns. As a result, the models disproportionately favor mainstream narratives, reinforcing existing gender stereotypes.

4.2.5.4 The Promise of PE and ICL

Our research presents an encouraging perspective on the effectiveness of PE and ICL as strategies to mitigate gender biases in LLMs. By employing controlled and bias-challenging prompts, we observed a significant 40% reduction in stereotypical gender associations across model outputs. Additionally, the Bias Score improved notably, with an average reduction of 16% compared to the baseline across all models.

Incorporating in-context feedback further enhanced the models' responsiveness, enabling outputs that aligned more closely with gender-neutral or counter-stereotypical expectations. Notably, when explicit feedback emphasized neutrality, subsequent model responses demonstrated improved balance and reduced stereotype reinforcement. This aligns with the findings of Sun et al., who suggest that while inherent biases exist within LLMs, strategic interventions can effectively guide model behavior toward fairness and inclusivity (Sun et al., 2020).

4.2.5.5 Guardrails: A Double-Edged Sword

While our guardrails demonstrated significant potential, it is crucial to approach their application with caution. Overcorrection risks generating outputs that, in striving for neutrality, become detached from real-world contexts. During our real-world scenario simulations, we observed instances where excessive emphasis on gender neutrality resulted in incoherent or overly sanitized responses, diminishing the practical value of the content. This observation echoes the concerns raised by Lakkaraju et al. regarding the unintended consequences of overly aggressive debiasing, which can compromise both coherence and utility (Lakkaraju et al., 2017).

4.2.5.6 Broader Socio-Political Implications

Bender et al. argue that the focus should extend beyond mere "de-biasing" efforts to encompass a deeper understanding of the broader societal implications, a perspective that resonates strongly with our findings (Bender et al., 2021). While technical interventions can mitigate surface-level manifestations of bias in LLM outputs, they do not confront the underlying societal structures and narratives that give rise to these biases.

Ultimately, while LLMs can be fine-tuned to reduce biased outputs, the responsibility also lies with society to critically examine and reshape the narratives that inform these models. AI systems, in many ways, act as a mirror reflecting societal realities. While it is possible to "clean" the mirror, addressing the distortions in what it reflects is equally, if not more, important for achieving lasting change.

4.2.6 Future Directions

Our study highlights several promising avenues for future research:

a. Personalized Guardrails: Exploring the development of user-specific guardrails that adapt to individual preferences and sensitivities, aiming to strike a balance between neutrality, personalization, and fairness.
b. Ethical Implications: Delving deeper into the ethical complexities of modifying LLM outputs. While the goal is to mitigate biases, defining what constitutes "bias" versus "neutrality" remains a nuanced and subjective challenge.
c. Interdisciplinary Approaches: Strengthening interdisciplinary collaborations among technologists, sociologists, ethicists, and linguists to ensure well-rounded, ethically sound advancements in LLM development.

The labyrinth of gender biases within LLMs is complex but not insurmountable. Our research, anchored in a rigorous methodology, offers critical insights into the extent of these biases, their societal implications, and potential mitigation strategies. While PE and ICL demonstrate significant promise, they represent only part of a broader solution that interweaves technology, societal values, ethics, and individual agency.

Ultimately, addressing gender representation in LLMs transcends technical fixes; it is a societal endeavour. As we advance towards an AI-driven future, embedding fairness, inclusivity, and accountability into these systems is not just necessary, it is imperative.

4.2.7 Conclusion

As we draw our research to a close, the relation between technology, gender representation, and societal structures becomes increasingly evident. LLMs are not merely technical constructs; they serve as mirrors reflecting the biases, values, and perceptions embedded within human society. In our conclusion, we summarize the key findings of this study, reflect on their broader implications, and outline the future pathways that lie ahead at the intersection of AI development and societal progress.

4.2.7.1 Summation of Key Findings

At the core of our investigation was the exploration of gender representation within LLMs. Our findings were unequivocal; gender biases, both subtle and overt, permeate model outputs. From role-based stereotypes embedded in responses to neutral prompts to the disproportionate representation of genders across diverse contexts, LLMs consistently mirrored, and at times amplified societal biases. These patterns, unsurprisingly, trace back to the models' training data, a vast distillation of human-generated content, inherently shaped by existing societal imbalances. Yet, amid these challenges, our research

uncovered a promising pathway forward. PE and ICL proved effective in modulating LLM behaviour, demonstrating that while biases may be inherited, their manifestation can be meaningfully mitigated.

4.2.7.2 Broader Implications

The implications of our research extend far beyond AI labs or technical discourse. As AI systems become increasingly embedded in education, employment, entertainment, and even governance, the biases they carry have the potential to shape societal perceptions and influence critical decisions. LLMs that perpetuate gender stereotypes risk reinforcing these biases in users, subtly shaping generational attitudes and worldviews. In high-stakes applications like recruitment, finance, or policy-making, such biases could translate into real-world disparities, affecting opportunities, outcomes, and fairness in ways that are both profound and far-reaching.

4.2.7.3 Reflections on Methodology

Our research methodology, combining both qualitative and quantitative approaches, enabled a thorough exploration of gender bias in LLMs. However, like any study, it comes with limitations. While PE and ICL demonstrated significant potential, they are not definitive solutions. The risk of overcorrection, where efforts to eliminate bias result in outputs detached from reality, remains a critical challenge. Additionally, despite the robustness of our study, the sheer scale and complexity of LLMs ensure that some nuances inevitably remain beyond the scope of this research.

4.2.7.4 The Ethical Horizon

The ethical dimensions of our findings are profound. While it is technically feasible to engineer LLM outputs to reduce bias, critical questions arise around who defines the standards of "neutrality" and "bias", and what the broader implications of those definitions might be. Is neutrality a universal concept, or is it inherently subjective, shaped by cultural and societal contexts? Moreover, in striving for neutrality, do we risk over-sanitizing content, stripping it of the richness, diversity, and nuance that make human expression meaningful? These ethical considerations highlight the delicate balance between fairness, representation, and authenticity in AI-generated content.

4.2.7.5 Societal Structures and AI

Our research highlights a fundamental truth: technology does not exist in isolation. LLMs and AI more broadly are products of human society, inevitably absorbing its values, biases, and prevailing narratives. While technical interventions can mitigate biased outputs, lasting solutions require confronting the societal structures that give rise to these biases in the first place.

"De-biasing" AI is not solely a technical challenge; it is a societal one. If our training data mirrors the world we live in, then the pursuit of unbiased AI is inseparable from the pursuit of a more equitable and just society.

4.2.7.6 The Way Forward

Given our findings and their broader implications, several key pathways emerge for future exploration:

a. **Interdisciplinary Collaboration:** There is a growing need for technologists to work alongside linguists, sociologists, ethicists, and gender studies experts. Only

through such interdisciplinary efforts can we design AI systems that are not only technically robust but also socially responsible and inclusive.

b. **User Agency:** Empowering users to define the "values" of their AI tools presents a promising direction. Personalized guardrails, allowing individuals to set preferences around content neutrality and bias, could strike a balance between universal fairness and personal choice.

c. **Continuous Learning and Adaptation:** LLMs should be viewed as continuously evolving systems. Regular updates, informed by user feedback and shifts in societal norms, are essential to keeping these models aligned with evolving values and expectations.

d. Open-Source Initiatives: Encouraging open-source AI research can democratize development, inviting diverse perspectives and reducing the risk of any single viewpoint dominating model behavior. This inclusivity is key to building more balanced and representative AI systems.

e. Public Awareness and Education: Raising public awareness about AI's capabilities, limitations, and potential biases is crucial. An informed user base can engage critically with AI systems, ensuring they benefit from these tools without being inadvertently shaped by their inherent biases.

4.2.7.7 Final Reflections

The journey into understanding gender representation in LLMs has been both enlightening and challenging. Navigating between technology and society, biases and neutrality, ethics and functionality has underscored a fundamental truth: the future of AI is not solely in the hands of technologists, but in the collective hands of society. As AI becomes increasingly woven into the fabric of our daily lives, the responsibility to shape,

guide, and refine it is shared by all. In the tapestry of an AI-enabled future, every thread, be it technology, ethics, societal values, or individual agency, plays a crucial role. And as we continue this journey, it is vital to remember that while AI may be the output of machines, its soul is inherently human.

In our pursuit of responsible AI development, Chapter 4 has unravelled the complexities surrounding bias identification and mitigation in NLU models. By systematically addressing our RQs across a spectrum of models, from rule-based systems to LLMs, this chapter has reinforced the imperative of recognizing and rectifying embedded biases. Our exploration of methodologies such as template-based test-set creation, PE, and ICL has not only illuminated the challenges but also revealed practical avenues to foster fairness and inclusivity in NLU applications.

The proposed guidelines align with the broader objective of moving beyond simply identifying problems, towards actively shaping more inclusive and representative technologies. As we conclude this chapter, the insights gained lay the foundation for the discussion in Chapter 5, where we synthesize key findings, reflect on their significance, and chart future directions, especially toward advancing NLU in low-resource languages.