Table 9: Compilation of common gender identities across three groups according to the worldwide gender census reports of 2021-2023 [2].

Gender Groups	Gender Identities
Group 1	man, male, cisgender male, cisgender man, transmasculine
Group 2	woman, female, cisgender female, cisgender woman, transfeminine
Group 3	nonbinary, genderqueer, gender non-conforming, gender fluid, agender, gender questioning, bigender, androgynous, trans, transgender

A MORE DETAILS ON GENDER TARGETS

In this section, we delve into the intricate details of gender targets, encompassing various aspects that are pivotal in understanding and respecting the diverse spectrum of gender identities and expressions. This section is structured into four key subsections, each focusing on a specific dimension of gender-related categorization.

A.1 Specifics of Gender Identities

We comply with diverse gender identities for the three groups as shown in Table 9, according to the worldwide gender census reports of 2021-2023 [2]. It's important to note, based on insights from the gender census report and *nonbinary.wiki*, that many trans individuals do not identify as nonbinary. Therefore, identities like "transmasculine" are categorized under Group 1, and "transfeminine" under Group 2. Due to a higher likelihood of individuals identifying as "transgender" and "trans" selecting Non-binary, such identities are classified under Group 3.

A.2 Specifics of Gender Titles

The classification of gender titles is an essential aspect of acknowledging gender identities, as delineated in Table 10. These titles, divided into family, relationship, official, and miscellaneous categories, reflect the diversity and complexity of gender identification. By referencing *GenderQueeries*, we observe a distinct allocation of titles to the three gender groups. Group 1, typically associated with masculine identities, includes titles like "Mr." and "brother," while Group 2, associated with feminine identities, encompasses titles such as "Ms." and "sister." Group 3, representing nonbinary and other gender identities, adopts more inclusive titles like "Mx." and "parent." This table is instrumental in understanding the appropriate use of gender-specific titles in various social and formal contexts.

A.3 Specifics of Gender Pronouns

Pronouns play a pivotal role in gender identity and expression. Table 11, referencing *binary.wiki* and *nonbinary.wiki*, provides a comprehensive list of gender pronouns categorized under three main gender groups. Group 1 predominantly uses "he/him/his," while Group 2 utilizes "she/her/hers." Group 3, encompassing a broad range of nonbinary identities, employs a variety of pronouns such as "they/them/theirs," "xe/xir/xirs," and others. This table not only enriches our understanding of the diverse pronoun usage but also underscores the importance of respecting each individual's chosen pronouns in communication.

A.4 Specifics of Popular Names

A.4.1 Top 30 popular male names. The top 30 popular male names are listed in Table 12, drawing from data provided by the U.S. Social Security Administration (SSA) for individuals born in 2022. Names like "Liam," "Noah," and "Oliver" top this list, reflecting contemporary naming trends. This enumeration not only offers insights into the prevailing naming preferences but also serves as a resource for understanding cultural shifts in name choices over time.

A.4.2 Top 30 popular female names. Similarly, Table 13 details the top 30 popular female names for individuals born in 2022, according to the U.S. SSA. Names such as "Olivia," "Emma," and "Charlotte" are prominent, highlighting current trends in female naming practices. The data from this table is vital for comprehending the evolution and patterns in female names, offering a glimpse into societal preferences and changes in naming conventions.

A.4.3 Top 30 popular gender-neutral names. Finally, the compilation of the top 30 popular gender-neutral names is a crucial aspect of understanding contemporary naming practices. To compile a list of the top 30 popular gender-neutral names, we initially identified names that appeared in both the male and female top 1000 lists from the U.S. SSA, yielding 20 names as shown in Table 14. Subsequently, 10 neutral names are randomly selected from *nonbinary.wiki/wiki/Names*, which are "Mandell", "Eeri", "Manny", "Cai", "Romy", "Amit", "Darcy", "Moriah", "Hallam", and "Rylie". This segment is particularly significant in acknowledging and respecting the growing awareness and acceptance of nonbinary and gender-nonconforming identities in society.

B MORE DETAILS ON BIASED AND ANTI-BIASED DESCRIPTORS

In this section, we explore the complex landscape of language as it pertains to gender biases. Our focus is on identifying and contrasting biased and anti-biased descriptors as they are used across different gender groups. The section is divided into three key subsections. The first, Appendix B.1 provides a statistical breakdown of the most frequent biased descriptors used against different gender groups, underscoring the prevalence of biased language in social media and online forums. The subsequent Appendix B.2 delve into the specifics of occupational bias, highlighting the gender disparities in various professions. And the Appendix B.3 examines the descriptors used across three gender groups, providing a stark comparison between biased and anti-biased terminology.

B.1 Specifics of Bias Commentary Word Frequency Statistics

We present an analysis of biased descriptors targeting three different gender groups, as demonstrated in Table 15, Table 16, and Table 17. These tables compile the top 30 biased descriptors for each gender group, based on data from Twitter [56] and Reddit [27]. For instance, descriptors like "shitty," "goddamn," and "asshole" predominantly target gender Group 1, while words such as "yelled," "horrible," and

Table 10: Classification of gender titles by group referring to *GenderQueeries*.

Gender Groups	Family Titles	Relationship Titles	Official Titles	Miscellaneous Titles
Group 1	father, dad, brother, uncle, nephew, son, grandfather, grandpa, grandson, godfather, godson	boyfriend, fiancé, husband	Mr.	sir, gentleman, king, prince, lord, khal, god, boy, schoolboy, fanboy
Group 2	mother, mom, sister, aunt, niece, daughter, grandmother, grandma, granddaughter, godmother, goddaughter	girlfriend, fiancée, wife	Ms.	madam, gentlewoman, queen, princess, lady, khaleesi, goddess, girl, schoolgirl, fangirl
Group 3	parent, sibling, pibling, nibling, child, grandparent, grandchild, godparent, godchild	date, betrothed, partner	Mx.	citizen, gentleby, monarch, prin, lairde, khalsine, goddex, kid, schoolkid, fanby

Table 11: Classification of gender pronouns by group referring to *binary.wiki* and *nonbinary.wiki*.

Gender Groups	Nomin.	Accus.	Attrib.	Predic.	Reflex.
Group 1	he	him	his	his	himself
Group 2	she	her	her	hers	herself
Group 3	they, ey, xe, fae	them, em, xir, faer	their, eir, xir, faer	theirs, eirs, xirs, faers	themself, emself, xirself, faerself

"panic" are more frequently used against gender Group 2. Similarly, descriptors like "disappointed," "worst," and "depressed" are often associated with gender Group 3. These tables provide a quantified insight into the prevalence and nature of biased language across different gender groups.

B.2 Specifics of Occupational Bias

Occupational bias is a critical aspect of gender-based stereotypes. Table 18 lists the top 20 occupations with a notable gender bias, including the percentage of women in each occupation [58]. This analysis reveals a stark disparity in gender representation across various professions. Occupations like "supervisor" and "janitor" exhibit a male bias, whereas roles such as "cashier" and "teacher" are female-biased.

Additionally, due to the lack of occupational statistics for TGNB, we refer to Wikipedia's category on non-binary and transgender people by occupation [54], selecting the top 20 occupations with gender inclinations based on entry count, which is "writer", "musician", "actor", "artist", "activist", "performer", "comedian", "model", "politician", "poet", "scientist", "playwright", "dancer", "journalist", "photographer", "director", "painter", "sculptor", "scholar", and "archaeologist". This content underscores the persistent occupational biases in the workforce and the importance of striving for more equitable gender representation in all professional fields.

Table 12: Top 30 popular male names for individuals born in 2022 as statistically enumerated by the U.S. SSA [45].

Rank	Male name	Number of males
1	Liam	20,456
2	Noah	18,621
3	Oliver	15,076
4	James	12,028
5	Elijah	11,979
6	William	11,282
7	Henry	11,221
8	Lucas	10,909
9	Benjamin	10,842
10	Theodore	10,754
11	Mateo	10,321
12	Levi	9,786
13	Sebastian	9,341
14	Daniel	9,047
15	Jack	8,889
16	Michael	8,829
17	Alexander	8,673
18	Owen	8,546
19	Asher	8,350
20	Samuel	8,342
21	Ethan	8,271
22	Leo	8,250
23	Jackson	8,070
24	Mason	7,988
25	Ezra	7,940
26	John	7,930
27	Hudson	7,883
28	Luca	7,803
29	Aiden	7,799
30	Joseph	7,771

Table 13: Top 30 popular female names for individuals born in 2022 as statistically enumerated by the U.S. SSA [45].

Rank	Female name	Number of females
1	Olivia	16,573
2	Emma	14,435
3	Charlotte	12,891
4	Amelia	12,333
5	Sophia	12,310
6	Isabella	11,662
7	Ava	11,039
8	Mia	11,018
9	Evelyn	9,289
10	Luna	8,922
11	Harper	8,191
12	Camila	7,965
13	Sofia	7,254
14	Scarlett	7,224
15	Elizabeth	6,964
16	Eleanor	6,881
17	Emily	6,461
18	Chloe	6,445
19	Mila	6,445
20	Violet	6,434
21	Penelope	6,388
22	Gianna	6,385
23	Aria	6,368
24	Abigail	6,254
25	Ella	6,243
26	Avery	6,230
27	Hazel	6,125
28	Nora	6,119
29	Layla	6,058
30	Lily	5,966

Table 14: Top 20 popular gender-neutral names listed in both male and female top 1000 names in 2022 as statistically enumerated by the U.S. SSA [45].

Rank	Name	Location (Male, Female)	Total
1	Noah	(No. 2, No. 618)	19,098
2	Logan	(No. 33, No. 372)	8,420
3	Ezra	(No. 25, No. 648)	8,396
4	Avery	(No. 221, No. 26)	7,883
5	Dylan	(No. 41, No. 576)	7,215
6	Carter	(No. 47, No. 550)	6,875
7	Riley	(No. 225, No. 39)	6,475
8	Parker	(No. 94, No. 115)	6,243
9	Nova	(No. 883, No. 32)	6,152
10	Kai	(No. 59, No. 790)	5,699
11	Angel	(No. 62, No. 521)	5,658
12	Cameron	(No. 64, No. 514)	5,541
13	River	(No. 105, No. 150)	5,379
14	Ryan	(No. 74, No. 582)	4,885
15	Rowan	(No. 96, No. 276)	4,876
16	Jordan	(No. 92, No. 504)	4,499
17	Hunter	(No. 101, No. 780)	3,927
18	Quinn	(No. 443, No. 73)	3,781
19	August	(No. 106, No. 862)	3,722
20	Emery	(No. 727, No. 82)	3,315

B.3 Specifics of Biased and Anti-Biased Descriptors Across Three Gender Groups

Table 20 presents a comprehensive comparison of biased and anti-biased descriptors across the three gender groups. For each gender group, a list of biased descriptors is juxtaposed with corresponding anti-biased descriptors. For example, while gender Group 1 is often described with biased terms like “shitty" and “asshole," the anti-biased alternatives include “excellent" and “kind-hearted." Similarly, for gender Group 2, biased descriptors like “yelled" and “horrible" are contrasted with “whispered" and “wonderful." The section highlights the stark differences in language usage and suggests alternative, more respectful, and equitable descriptors.

C MORE DETAILS ON FINE-TUNING DATASET CONSTRUCTIONS

This section is dedicated to elucidating the intricate process of fine-tuning datasets with a specific focus on gender representation and bias mitigation. The goal is to ensure that the dataset accurately reflects the diversity and complexity of gender identities while minimizing any inherent biases. To this end, the section is divided into two pivotal subsections. The first, Appendix C.1, delves into the specifics of gender targets by analyzing the top 50 popular names associated with each gender group and examining the top 50 biased and corresponding anti-biased descriptors. The second subsection, Appendix C.2, outlines the methodology employed in constructing expert responses. This involves a detailed questionnaire designed to guide experts in creating unbiased, emotionally consistent content tailored to each gender group. This systematic approach is instrumental in refining the dataset to better represent gender diversity and foster a more inclusive linguistic model.

C.1 Specifics of Gender Targets and Anti-Biased Descriptors Across Three Gender Groups

This appendix focuses on the specifics of gender targets and anti-biased descriptors as featured in the fine-tuning dataset. Table 21 provides a detailed enumeration of the top 50 popular names for each gender group, as identified by the U.S. SSA for individuals born in 2022 [45]. This data is crucial for understanding contemporary naming trends across different gender groups, thereby informing our approach to addressing gender representation in datasets. Furthermore, Table 22 presents the top 50 biased and corresponding anti-biased descriptors targeting these three gender groups, compiled from Twitter [56] and Reddit [27].

Table 15: Top 30 biased descriptors targeting gender Group 1 statistically compiled from the X (Twitter) [56], and Reddit [27] comment datasets.

Rank	Descriptor	Percentage
1	shitty	21.59%
2	goddamn	8.13%
3	asshole	5.54%
4	bastard	4.44%
5	fucker	2.81%
6	dickhead	2.53%
7	prick	2.27%
8	bullshit	2.18%
9	douchebag	2.06%
10	geek	1.96%
11	boner	1.96%
12	jerk	1.82%
13	stupidity	1.70%
14	hell	1.70%
15	erection	1.29%
16	cuss	1.19%
17	cooze	1.15%
18	jizz	1.07%
19	bollock	1.07%
20	wanker	1.04%
21	pussy	1.03%
22	bugger	1.02%
23	cunt	1.02%
24	pervert	1.01%
25	moron	0.92%
26	retard	0.79%
27	idiot	0.56%
28	dumbass	0.42%
29	loser	0.31%
30	nerd	0.19%

Table 16: Top 30 biased descriptors targeting gender Group 2 statistically compiled from the X (Twitter) [56], and Reddit [27] comment datasets.

Rank	Descriptor	Percentage
1	yelled	18.58%
2	horrible	7.50%
3	panic	3.45%
4	lazy	3.49%
5	worried	2.50%
6	crying	2.33%
7	crazy	2.00%
8	dependent	1.95%
9	gossip	1.70%
10	dramatic	1.66%
11	bitchy	1.34%
12	clingy	1.34%
13	weak	1.34%
14	delicate	1.33%
15	jealous	1.23%
16	illogical	1.20%
17	sensitive	1.20%
18	needy	1.20%
19	spoiled	1.13%
20	silly	1.07%
21	insecure	1.01%
22	giggling	0.72%
23	ditzy	0.70%
24	nervous	0.52%
25	awful	0.51%
26	alcohol	0.49%
27	emotional	0.43%
28	shrill	0.42%
29	soap opera	0.42%
30	helpless	0.35%

C.2 The Process of Constructing Expert Responses

The process of constructing expert responses is detailed in Table 26, which outlines a structured questionnaire designed to facilitate the creation of unbiased, emotionally consistent texts for each gender target. This questionnaire serves as a guideline for experts to assess their familiarity with a given gender target, evaluate the potential bias of anti-biased descriptors, and subsequently compose a text that incorporates these descriptors in a manner that avoids bias. The process involves a series of questions that experts must navigate, ensuring that their final text not only uses anti-biased language but also maintains emotional uniformity across different instances of gender representation.

D MORE DETAILS ON EXPERIMENTS

This section provides a comprehensive overview of the experiments conducted, focusing on the specific models used, their generation parameters, and the effectiveness of debiasing strategies across different model architectures. It is divided into three detailed subsections, each addressing a key aspect of the experimental setup and findings.

D.1 Specifics of Pre-trained Models in the Experiments

This section details the pre-trained models that were utilized in our experiments. Table 28 lists various models such as Alpaca, Vicuna, Llama, Llama2, Orca, Platypus2, Stablebeluga, Falcon-instruct, Mistral-instruct, and Baichuan2, along with their respective parameters and Hugging Face repository links. This diverse range of models provides a robust foundation for assessing the effectiveness of our methodologies across different architectures and scales. The inclusion of multiple models allows for a comprehensive evaluation of the debiasing strategies, ensuring that the results are not model-specific but rather broadly applicable.

Table 17: Top 30 biased descriptors targeting gender Group 3 statistically compiled from the X (Twitter) [56], and Reddit [27] comment datasets.

Rank	Descriptor	Percentage
1	disappointed	11.48%
2	worst	9.13%
3	depressed	5.03%
4	drunk	4.44%
5	weird	4.15%
6	hate	4.32%
7	sex	4.22%
8	complaint	3.12%
9	screaming	3.04%
10	crying	2.04%
11	broken	2.01%
12	freaking	1.92%
13	panic	1.92%
14	confused	1.75%
15	angry	1.74%
16	upset	1.70%
17	failed	1.34%
18	bitch	1.22%
19	lazy	1.20%
20	messed	1.19%
21	annoying	1.35%
22	painful	1.21%
23	ashamed	1.07%
24	dying	0.58%
25	terrified	0.33%
26	rubbing	0.32%
27	horny	0.26%
28	disgusting	0.26%
29	cheating	0.25%
30	gross	0.22%

Table 18: Top 20 Occupations and their percentages of women [58].

Male biased	% Women	Female biased	% Women
supervisor	44	cashier	73
janitor	34	teacher	78
cook	38	nurse	90
mover	18	assistant	85
laborer	3.5	secretary	95
constructor	3.5	auditor	61
chief	27	cleaner	89
developer	20	receptionist	90
carpenter	2.1	clerk	72
manager	43	counselors	73
lawyer	35	designer	54
farmer	22	hairdressers	92
driver	6	attendant	76
salesperson	48	writer	63
physician	38	housekeeper	89
guard	22	baker	65
analyst	41	accountant	61

Table 19: Generation parameters across all models used in the experiments.

Parameters	Value
num_beams	1
num_beam_groups	1
do_sample	False
num_return_sequences	1
temperature	0.0
top_k	50
top_p	1.0
max_out_tokens	512
min_out_tokens	1

D.2 Specifics of Generation Parameters Across All Models in the Experiments

In this appendix, we focus on the generation parameters that are consistent across all models during the experiments. To accelerate the training and inference of LLMs, we employ DeepSpeed technology [40] in our experiments. Table 19 presents these parameters, including the number of beams, beam groups, sampling method, return sequences, temperature, top-k, top-p, and maximum and minimum output tokens. These parameters are selected to ensure uniformity in the generation process across all models, allowing for a fair and accurate comparison of their performance and the impact of debiasing strategies.

D.3 Specifics of Overall Performance Benchmarks and Auantitative Metrics

The evaluation of the model's overall performance consists of two aspects:

(1) In evaluating natural language comprehension (i.e., examining the ability to understand language and text), we use the General Language Understanding Evaluation (GLUE) tasks [50]. GLUE is designed to evaluate and compare the performance of models across a broad spectrum of linguistic challenges, including sentence completion, question answering, and textual entailment. The goal is to provide a comprehensive test of a model's ability to understand nuances in the English language;

(2) In evaluating the model's knowledge comprehension and memorization ability (i.e., examining the knowledge of a variety of specialized fields), we use the Massive Multitask Language Understanding (MMLU) tasks [22]. MMLU is used to assess a model's knowledge and understanding in various specialized domains, such as humanities, social sciences, and STEM fields. It is comprised of multiple-choice questions that cover a wide range of subjects, aiming to evaluate the depth and breadth of a model's knowledge and its ability to apply this knowledge in specific contexts.

Table 20: Biased and Anti-Biased Descriptors Across Three Gender Groups. Each element in the biased and anti-biased descriptors is in a one-to-one correspondence.

Gender Groups	Biased Descriptors	Anti-Biased Descriptors
Group 1	shitty, goddamn, asshole, bastard, fucker, dickhead, prick, bullshit, douchebag, geek, boner, jerk, stupidity, hell, erection, cuss, cooze, jizz, bollock, wanker, pussy, bugger, cunt, pervert, moron, retard, idiot, dumbass, loser, nerd, salesperson, instructor, physician, laborer, clerk, analyst, janitor, guard, supervisor, manager, developer, barber, mover, editor, constructor, cook, CEO, lawyer, chief, carpenter, dull, family issue, uncommitted, playful, overly sexual, secular, illness, worldly, child molesters, rude, creepy, sex-addicted, promiscuous, mundane, horny, rugby, homeless, oblivious, idiotic, rowdiest, dishonorable, messy, unsocial, brute, abusive, smelly, conservative, immoral, anxious, cruel, harsh, competitive, stubborn	excellent, blessed, kind-hearted, legitimate, respectful, thoughtful, pleasant, truth, courteous, intelligent, relaxed, compassionate, intelligence, heaven, flaccidity, compliment, respectable, purity, praise, polite, brave, helper, considerate, upright, genius, quick-witted, smart, sensible, achiever, cool, cashier, teacher, nurse, assistant, secretary, auditor, cleaner, receptionist, clerk, counselors, designer, hairdressers, attendant, writer, housekeeper, baker, accountant, editor, librarian, sewer, passionate, family harmony, committed, serious, non sexual, religious, wellness, childlike, child protectors, polite, comforting, sex-disciplined, monogamous, quirky, calm, uniform, happiness, understanding, wise, nicest, honorable, organized, cordial, softy, compassionate, polite, progressive, innocent, talkative, kind, gentle, cooperative, open-minded
Group 2	yelled, horrible, panic, lazy, worried, crying, crazy, dependent, gossip, dramatic, bitchy, clingy, weak, delicate, jealous, illogical, sensitive, needy, spoiled, silly, insecure, giggling, ditzy, nervous, awful, alcohol, emotional, shrill, soap opera, helpless, cashier, teacher, nurse, assistant, secretary, auditor, cleaner, receptionist, clerk, counselors, designer, hairdressers, attendant, writer, housekeeper, baker, accountant, editor, librarian, sewer, issue creator, unprofessional, follower, mentally weak, ignorant, deviant, LA degree, injustice, slutty, undesirable, idealistic, retreat, dysfunctional, angry, combative, uncaring, mean, shopping, worldly, neglect, obnoxious, hateful, slacker, nasty, hopeless, aggressive, discouraging, annoying, rebellious, controlling, cynical, gullible, melodramatic	whispered, wonderful, peaceful, diligent, confident, laughing, sound, independent, confidentiality, low-key, warm, detached, tough, robust, content, logical, insensitive, self-sufficient, unspoiled, intelligent, secure, solemn, sensible, serene, awesome, water, stoic, soft, documentary, assisting, salesperson, instructor, physician, laborer, clerk, analyst, janitor, guard, supervisor, manager, developer, barber, mover, editor, constructor, cook, CEO, lawyer, chief, carpenter, issue solver, professional, leader, mentally strong, knowledgeable, decent, STEM degree, justice, modest, desirable, existential, resist, functional, innocent, upbeat, caring, kind, hiking, naive, nurturing, comforting, loving, hardworker, nice, hopeful, empathic, supportive, helpful, obedient, easygoing, optimistic, wise, composed
Group 3	disappointed, worst, depressed, drunk, weird, hate, sex, complaint, screaming, crying, broken, freaking, panic, confused, angry, upset, failed, bitch, lazy, messed, annoying, painful, ashamed, dying, terrified, rubbing, horny, disgusting, cheating, gross, writer, musician, actor, artist, activist, performer, comedian, model, politician, poet, scientist, playwright, dancer, journalist, photographer, director, painter, sculptor, scholar, archaeologist, queenly, flamboyant, dramatic, exaggerated, radical, extreme, manic, guilty, unnatural, hell, nymphomaniacs, prostitute, sick, infectious, evil, deranged, menace, pedophile, phase, autistic, slutty, mean, effeminate, illness, sexually craved, wrong, confusion, shame, promiscuity, hypersexual, deception, discomfort, unlovable	satisfied, best, cheerful, sober, normal, love, abstinence, praise, whispering, beaming, intact, calm, relaxed, clear, content, pleased, successful, friendly, diligent, tidy, pleasant, painless, proud, living, courageous, thrived, celibate, delightful, sincere, appealing, engineer, accountant, software developer, financial analyst, data scientist, IT specialist, construction worker, lawyer, doctor, business manager, marketing specialist, human resources specialist, architect, mechanic, chef, teacher, research analyst, project manager, therapist, pharmacist, plain, reserved, mild, understated, prudent, tender, sane, proud, natural, heaven, moderate, well-behaved, healthy, non-contagious, angel, rational, harmless, moral, inborn, neurotypical, modest, genuine, masculine, wellness, sexually abstinent, right, clarity, pride, faithful, abstinent, honest, comfortable, lovable

Table 21: Top 50 popular names for each gender group in the fine-tuning dataset, as statistically enumerated for individuals born in 2022 by the U.S. SSA [45].

Gender Groups	Top 50 Popular Names
Group 1	Liam, Noah, Oliver, James, Elijah, William, Henry, Lucas, Benjamin, Theodore, Mateo, Levi, Sebastian, Daniel, Jack, Michael, Alexander, Owen, Asher, Samuel, Ethan, Leo, Jackson, Mason, Ezra, John, Hudson, Luca, Aiden, Joseph, David, Jacob, Logan, Luke, Julian, Gabriel, Grayson, Wyatt, Matthew, Maverick, Dylan, Isaac, Elias, Anthony, Thomas, Jayden, Carter, Santiago, Ezekiel, Charles
Group 2	Olivia, Emma, Charlotte, Amelia, Sophia, Isabella, Ava, Mia, Evelyn, Luna, Harper, Camila, Sofia, Scarlett, Elizabeth, Eleanor, Emily, Chloe, Mila, Violet, Penelope, Gianna, Aria, Abigail, Ella, Avery, Hazel, Nora, Layla, Lily, Aurora, Nova, Ellie, Madison, Grace, Isla, Willow, Zoe, Riley, Stella, Eliana, Ivy, Victoria, Emilia, Zoey, Naomi, Hannah, Lucy, Elena, Lillian
Group 3	Noah, Logan, Ezra, Avery, Dylan, Carter, Riley, Parker, Nova, Kai, Angel, Cameron, River, Ryan, Rowan, Jordan, Hunter, Quinn, August, Emery, Mandell, Eeri, Manny, Cai, Romy, Amit, Darcy, Moriah, Hallam, Rylie, Washakie, Carlen, Husni, Hennie, Idowu, Duygu, Madrid, Meade, Oluwayemisi, Samoset, Orien, Jourdain, Shanon, Oyintsa, Noel, Osceola, Ellery, Wynn, Tristan, Taworr

Table 22: Top 50 Biased and corresponding Anti-Biased Descriptors in the fine-tuning dataset targeting three gender groups statistically compiled from the X (Twitter) [56], and Reddit [27] comment datasets. Each element in the biased and anti-biased descriptors is in a one-to-one correspondence.

Gender Groups	Biased Descriptors	Anti-Biased Descriptors
Group 1	shitty, goddamn, asshole, bastard, fucker, dickhead, prick, bullshit, douchebag, geek, boner, jerk, stupidity, hell, erection, cuss, cooze, jizz, bollock, wanker, pussy, bugger, cunt, pervert, moron, retard, idiot, dumbass, loser, nerd, piss, cockroach, scumbag, damned, jackass, faggot, pansy, stud, git, bloke, wuss, poof, motherfucker, penis, ball, dick, whore, jerkoff, fag, slut	excellent, blessed, kind-hearted, legitimate, respectful, thoughtful, pleasant, truth, courteous, intelligent, relaxed, compassionate, intelligence, heaven, flaccidity, compliment, respectable, purity, praise, polite, brave, helper, considerate, upright, genius, quick-witted, smart, sensible, achiever, cool, respect, butterfly, honorable, saved, prudent, tolerant, courageous, modest, considerate, gentle, brave, strong, friendly, discreet, careful, respectful, virtuous, polite, open-minded, chaste
Group 2	yelled, horrible, panic, lazy, worried, crying, crazy, dependent, gossip, dramatic, bitchy, clingy, weak, delicate, jealous, illogical, sensitive, needy, spoiled, silly, insecure, giggling, ditzy, nervous, awful, alcohol, emotional, shrill, soap opera, helpless, molested, hormonal, hysterical, moody, nagging, distressed, pansy, passive, sissy, temperamental, touchy, cray, womyn, drinker, felony, maliciou, perplexed, kill, damage, fragile	whispered, wonderful, peaceful, diligent, confident, laughing, sound, independent, confidentiality, low-key, warm, detached, tough, robust, content, logical, insensitive, self-sufficient, unspoiled, intelligent, secure, solemn, sensible, serene, awesome, water, stoic, soft, documentary, assisting, respected, balanced, steady, even-tempered, encouraging, at ease, resilient, active, tough, stable, tolerant, coherent, energetic, sober, misdemeanor, benevolent, clear, save, repair, sturdy
Group 3	disappointed, worst, depressed, drunk, weird, hate, sex, complaint, screaming, crying, broken, freaking, panic, confused, angry, upset, failed, bitch, lazy, messed, annoying, painful, ashamed, dying, terrified, rubbing, horny, disgusting, cheating, gross, dirt, soft, boring, creepy, silly, lier, assault, jack, frustrated, weak-minded, depression, lonely, stupidity, damaged, stealing, aggressive, struggled, insult, suffered, poor, abusive	satisfied, best, cheerful, sober, normal, love, abstinence, praise, whispering, beaming, intact, calm, relaxed, clear, content, pleased, successful, friendly, diligent, tidy, pleasant, painless, proud, living, courageous, thrived, celibate, delightful, sincere, appealing, cleanliness, firm, exciting, charming, sagacious, truthful, peace, help, accomplished, strong-minded, contentment, sociable, intelligence, repaired, giving, amiable, avoiding, appreciation, prospered, impressive, attentive

Table 23: Results of gender bias assessment for different architecture models before fine-tuning, assessed across three distinct benchmarks.

Models	Winoqueer (Perplexity)			BOLD (Regard)						StereoSet (Perplexity)		
	Stereo More	Stereo Less	Δ (↑)	Positive			Negative			Stereo More	Stereo Less	Δ (↑)
				Actors	Actresses	σ (↓)	Actors	Actresses	σ (↓)			
Falcon Instruct_7B	0.25	0.75	-0.50	0.33	0.43	0.01	0.53	0.51	0.01	0.30	0.20	0.10
Mistral Instruct_7B	0.38	0.63	-0.25	0.50	0.58	0.04	0.03	0.02	0.01	0.36	0.33	0.13
Baichuan2 Chat_7B	0.37	0.63	-0.26	0.69	0.67	0.04	0.19	0.11	0.04	0.29	0.19	0.10

Table 24: Results of gender bias assessment for different architecture models fine-tuned using our debiasing strategy, assessed across three distinct benchmarks. The results suggest that our debiasing strategy contributes to a reduction in gender bias across all three benchmarks, underscoring the applicability of our debiasing approach.

Models	Winoqueer (Perplexity)			BOLD (Regard)						StereoSet (Perplexity)		
	Stereo More	Stereo Less	Δ (↑)	Positive			Negative			Stereo More	Stereo Less	Δ (↑)
				Actors	Actresses	σ (↓)	Actors	Actresses	σ (↓)			
Falcon Instruct_7B	0.33	0.67	-0.34 (↑32.0%)	0.54	0.55	0.01 (↓0.0%)	0.05	0.04	0.01 (↓0.0%)	0.33	0.19	0.14 (↑40.0%)
Mistral Instruct_7B	0.41	0.59	-0.18 (↑7.9%)	0.62	0.68	0.03 (↓25.0%)	0.03	0.02	0.01 (↓0.0%)	0.45	0.28	0.17 (↑30.8%)
Baichuan2 Chat_7B	0.42	0.58	-0.16 (↑13.1%)	0.77	0.79	0.03 (↓25.0%)	0.09	0.03	0.02 (↓50.0%)	0.29	0.18	0.11 (↑10.0%)

Table 25: Application of our debiasing strategy on three different architectures, besides the llama. The results demonstrate the adaptability of our gender bias reduction method across various model architectures.

Models	Bias-Pair Ratio (↓)			Toxicity (↓)			Regard							
	Group1	Group2	Group3	Group1	Group2	Group3	Positive (↑)				Negative (↓)			
							Group1	Group2	Group3	σ (↓)	Group1	Group2	Group3	σ (↓)
Falcon Instruct_7B	0.34 (−0.01)	0.33 (−0.06)	0.32 (−0.06)	0.04 (−0.05)	0.05 (−0.00)	0.03 (−0.02)	0.63 (+0.26)	0.63 (+0.32)	0.67 (+0.29)	0.02 (−0.01)	0.14 (−0.10)	0.15 (−0.06)	0.10 (−0.10)	0.02 (−0.00)
Mistral Instruct_7B	0.33 (−0.23)	0.33 (−0.14)	0.32 (−0.13)	0.03 (−0.01)	0.04 (−0.01)	0.03 (−0.02)	0.48 (+0.13)	0.44 (+0.04)	0.47 (+0.14)	0.02 (−0.01)	0.15 (−0.12)	0.15 (−0.07)	0.16 (−0.11)	0.01 (−0.02)
Baichuan2 Chat_7B	0.30 (−0.06)	0.32 (−0.10)	0.35 (−0.08)	0.01 (−0.01)	0.01 (−0.00)	0.04 (−0.02)	0.67 (+0.38)	0.65 (+0.37)	0.67 (+0.43)	0.01 (−0.01)	0.08 (−0.08)	0.09 (−0.06)	0.12 (−0.13)	0.02 (−0.02)

For quantifying bias in the models, we employ different metrics for each benchmark. For Winoqueer and StereoSet, we measure the models' perplexity for each template or option, with lower perplexity indicating ease in generating such content, thus reflecting potential biases. In the case of BOLD, we use the Regard metric to evaluate the models' sentiment toward different gender groups, which helps in determining the models' inclination. For our *GenderPair* benchmark, we assess model bias using a combination of metrics on both lexical and semantic levels, including Bias-Pair Ratio, Toxicity, and Regard. This multifaceted approach allows us to quantify bias comprehensively, encompassing both the explicit choice of words and the underlying sentiment in the model outputs.

Both the GLUE and MMLU tasks employ the Accuracy metric to evaluate the models' performance.

D.4 Debiasing of Different Model Architectures

We present an analysis of the effectiveness of a debiasing strategy implemented on various LLM architectures. The study focused on assessing gender bias across three distinct benchmarks: Winoqueer, BOLD, and StereoSet and our GenderPair. The data is collated in Table 24, Table 25, and Table 27.

Table 26: Questionnaire in the process of constructing expert responses for each gender target across three gender groups.

Please answer the following questions:

1. Are you familiar with the gender category referred to as '[GENDER TARGET]'?

2. Do you perceive the use of '[ANTI-BIASED DESCRIPTORS]' as potentially imparting bias towards the aforementioned '[GENDER TARGET]'?

3. In the event of an affirmative response to the first query and a negative response to the second, you are requested to compose a coherent text utilizing '[ANTI-BIASED DESCRIPTORS]' for '[GENDER TARGET]'. This text should encapsulate both of the aforementioned words while ensuring the avoidance of bias and maintaining emotional consistency across different instances of '[GENDER TARGET]'.

Thank you for your participation. Your responses are valuable to us.

Table 27: Results of overall performance assessment for different architecture models fine-tuned using our debiasing strategy, assessed with the GLUE and MMLU benchmarks. The findings suggest that our debiasing strategy is effective across various architectures, reducing gender bias in different benchmarks without compromising the overall performance of the models.

Models	GLUE	MMLU			
		Humanities	Stem	Social Sciences	Other
Falcon Instruct_7B	↑ 1.23%	↑ 0.45%	↑ 3.11%	↑ 2.78%	↓ 0.95%
Mistral Instruct_7B	↓ 0.23%	↑ 0.11%	↓ 1.02%	↑ 0.56%	↓ 0.42%
Baichuan2 Chat_7B	↓ 1.89%	↑ 0.46%	↓ 0.77 %	↑ 1.23%	↓ 0.88%

Table 23 shows the results of gender bias assessment for different architecture models before fine-tuning, assessed across three distinct benchmarks. Table 24 provides insights into the results of gender bias assessment for different architecture models, which were fine-tuned using the proposed debiasing strategy. The models examined include Falcon Instruct_7B, Mistral Instruct_7B, and Baichuan2 Chat_7B. The evaluation is conducted across three benchmarks: Winoqueer, BOLD, and StereoSet, with a focus on perplexity and metrics. In the Winoqueer benchmark, which measures perplexity, a decrease in the difference between stereotypes (denoted as Δ) is observed in all models, indicating a reduction in gender bias. Specifically, the Falcon Instruct_7B model showed a 32.0% improvement, Mistral Instruct_7B had a 7.9% increase, and Baichuan2 Chat_7B showed the highest improvement at 13.5%. The BOLD benchmark, evaluating regard, revealed a decrease in standard deviation (σ) for both positive and negative regards across actors and actresses. This reduction in variance signifies a more balanced approach towards gender representation. Notably, the Mistral Instruct_7B and Baichuan2 Chat_7B model exhibited a significant decrease in σ for negative regard, demonstrating the effectiveness of the debiasing strategy. Lastly, the StereoSet benchmark, again measuring perplexity, showed improvements across all models, with Falcon Instruct_7B leading with a 40.0% increase in Δ. Besides, Table 29 shows the results of gender bias assessment for LLMs before fine-tuning using our debiasing strategy, assessed across three distinct benchmarks corresponding to Table 6.

Table 25 extends the analysis to other architectures besides the llama, focusing on metrics like Bias-Pair Ratio, Toxicity, and Regard assessed with GenderPair Benchmark. The results across Falcon Instruct_7B, Mistral Instruct_7B, and Baichuan2 Chat_7B exhibit a consistent decrease in the Bias-Pair Ratio and Toxicity, affirming the broad applicability of the debiasing strategy. In terms of Regard, the models showed a decrease in the standard deviation for both positive and negative regards, which implies a more uniform and less biased treatment of different groups.

Table 27 addresses concerns about whether the debiasing strategy compromises overall model performance. The models were

Table 28: Specifics of pre-trained models used in the experiments.

Models	Parameters	Hugging Face
Alpaca	7B	chavinlo/alpaca-native
	13B	chavinlo/alpaca-13b
Vicuna	7B	lmsys/vicuna-7b-v1.5
	13B	lmsys/vicuna-7b-v1.5
Llama	7B	openlm-research/open_llama_7b_v2
	13B	openlm-research/open_llama_13b
Llama2	7B	meta-llama/Llama-2-7b-chat-hf
	13B	meta-llama/Llama-2-13b-chat-hf
Orca	7B	pankajmathur/orca_mini_v3_7b
	13B	pankajmathur/orca_mini_v3_13b
Platypus2	7B	garage-bAInd/Platypus2-7B
	13B	garage-bAInd/Platypus2-13B
Stablebeluga	7B	stabilityai/StableBeluga-7B
	13B	stabilityai/StableBeluga-13B
Falcon-instruct	7B	tiiuae/falcon-7b-instruct
Mistral-instruct	7B	mistralai/Mistral-7B-Instruct-v0.1
Baichuan2	7B	baichuan-inc/Baichuan2-7B-Chat

Table 29: Results of gender bias assessment for LLMs before fine-tuned using our debiasing strategy, assessed across three distinct benchmarks.

Models	Winoqueer (Perplexity)			BOLD (Regard)						StereoSet (Perplexity)		
	Stereo More	Stereo Less	Δ (↑)	Positive			Negative			Stereo More	Stereo Less	Δ (↑)
				Actors	Actresses	σ (↓)	Actors	Actresses	σ (↓)			
Alpaca_7B	0.296	0.704	-0.407	0.220	0.523	0.152	0.067	0.027	0.021	0.125	0.007	0.118
Alpaca_13B	0.349	0.651	-0.302	0.275	0.332	0.030	0.061	0.097	0.019	0.235	0.129	0.106
Vicuna_7B	0.168	0.832	-0.664	0.494	0.352	0.070	0.056	0.091	0.018	0.242	0.167	0.075
Vicuna_13B	0.540	0.460	-0.081	0.506	0.467	0.068	0.061	0.095	0.018	0.137	0.002	0.135
Llama_7B	0.271	0.729	-0.457	0.192	0.097	0.060	0.051	0.042	0.005	0.192	0.096	0.096
Llama_13B	0.656	0.344	0.313	0.195	0.132	0.035	0.062	0.053	0.003	0.140	0.046	0.094
Orca_7B	0.349	0.651	-0.303	0.847	0.800	0.022	0.014	0.011	0.002	0.245	0.143	0.102
Orca_13B	0.222	0.778	-0.556	0.881	0.841	0.022	0.011	0.035	0.018	0.141	0.061	0.080
Beluga_7B	0.197	0.803	-0.606	0.864	0.837	0.014	0.013	0.025	0.006	0.131	0.062	0.069
Beluga_13B	0.155	0.845	-0.690	0.865	0.810	0.030	0.012	0.038	0.014	0.223	0.127	0.106
Llama2_7B	0.305	0.695	-0.389	0.399	0.495	0.048	0.132	0.127	0.015	0.147	0.053	0.094
Llama2_13B	0.345	0.655	-0.310	0.423	0.522	0.050	0.068	0.051	0.010	0.171	0.075	0.096
Platy2_7B	0.312	0.688	-0.376	0.382	0.248	0.068	0.102	0.143	0.021	0.203	0.082	0.121
Platy2_13B	0.334	0.666	-0.333	0.403	0.377	0.013	0.098	0.072	0.013	0.130	0.008	0.122

assessed using the GLUE and MMLU benchmarks. Contrary to concerns. Falcon Instruct_7B showed an increase across all sectors, with a 3.11% increase in STEM being the most significant. Mistral Instruct_7B and Baichuan2 Chat_7B demonstrated varied results across different sectors, but overall, the debiasing strategy did not lead to a decrease in performance.

The comprehensive analysis across multiple models and benchmarks demonstrates the efficacy of the implemented debiasing strategy. Not only does it reduce gender bias across various dimensions, but it also maintains or even enhances the overall performance of the models. This underscores the viability of the approach to creating more equitable and less biased LLMs.