LAIDLAW SCHOLARS

UNIVERSITY OF TORONTO

A Future Only “Generated” for Some: an Investigation into Gender Bias in GenAI Text Outputs and Bias Mitigation Using Prompt-based Framework

Ready when you are.

Introduction

This study explored how **genAI models reproduce gender stereotypes** in text generation. It also proposed **a prompt-based framework** that AI developers can use to **detect gender bias** in their products and hence **reduce them**.

Specifically, it:

📂 Analyzed AI models' responses to existing human sexism Likert scales. A brief overview of these 4 scales:

- AWS (1972): “Women = Stupid”
- ASI (1996): “Women = Cherished”
- MSS (1995) & Neosexism (1995): “Sexism doesn't exist anymore!”

Tested models' stereotypes when given pairs of gendered occupations to write stories about

Transformed 5 sexism scale items into long-form narrative prompts to qualitatively identify subtle biases in AI-generated stories

Developed a 30-item prompt-based diagnostic framework targeting six major gender bias dimensions and tested them across 5 top AI models

Summary

📊 **Six Dimensions of Subtle Sexism in Scenario-based AI Text Outputs Identified:**

1. Belonging & Relationship Status
2. Leadership Style & Effectiveness
3. Morality & Recognition
4. Family Background & Upbringing
5. Protection vs. Challenge
6. Communality vs. Agency (Warmth & Competence)

⚒ **30-Item Prompt-Based Framework**

See App. G for the entire prompt set! 😉

📲 **11 Top AI Models Tested:**

Gemini 2.5 Pro, GPT-4o, OpenAI o3, Claude Sonnet 4, Deepseek R1, Grok 3, Gemini 2.5 Flash, GPT-4.5 Preview, Qwen 3-235B-A22B-no-thinking, GPT-4.1, Deepseek V3

Supervised by Professor Jieying Chen

Yinuo Fang

Laidlaw Research Scholar

Methods

Phase 1

🧠 *Are AI models good at avoiding surface-level gender bias in statements?*

🧠 *True or False: such high avoidance of explicit sexism = unbiased text generation in story-writing?*

AI models included: *Gemini 2.5 Pro, GPT-4o, OpenAI o3, Claude Sonnet 4*

66-Item Likert Scale Test

Combined 4 established human sexism scales (*Attitudes Toward Women Scale (AWS), Modern Sexism Scale (MSS), Ambivalent Sexism Inventory (ASI), Neosexism Scale*)

Models rated each item from 1 (*sexist*) to 7 (*egalitarian*).

Prompt-Based Occupational Bias Test

⚒ Gave AI 10 pairs of gendered occupations (e.g. *doctor/nurse*) and asked them to write stories of 2 characters

🔍 Examined whether character assignment in stories align with stereotypical gender assumptions for each job

Phase 2

🧠 *“Two people. No pronouns. No labels. Just a story.” – What will AI give?*

🧠 *If subtle sexism persists in AI text output... How can we identify, categorize, and avoid them?*

AI models included: *Gemini 2.5 Pro, GPT-4o, OpenAI o3, Deepseek R1, Grok 3, Gemini 2.5 Flash, GPT-4.5 Preview, Qwen 3-235B-A22B-no-thinking, GPT-4.1, Deepseek V3*

Induction Prompt Construction and Qualitative Coding

Adapted 5 items from the four source scales into 2000-word scenario prompts, each written deliberately without pronouns or gender specification

Top 10 models from HuggingFace LLM leaderboard, each generated 5 stories per prompt (250 stories in total)

Stories were imported into MAXQDA for manual qualitative coding for subtle stereotypes in AI text generation

Phase 3

🧠 *When patterns repeat, can we make them measurable, testable, and fixable?*

🧠 *Phase 1 = spot the bias; Phase 2 = name the bias; Phase 3 =... build the tool to catch the bias!*

AI models included: *Gemini 2.5 Pro, GPT-4o, OpenAI o3, Deepseek R1, Grok 3*

30-item Prompt-based Framework Development and Efficiency Testing

Converted 53 codes from Phase 2 into **30 neutral prompts**, each targeting a specific stereotype.

Prompts structured as short narrative tasks (~400 words) about **two characters** with contrasting roles, each written without pronouns and sometimes with counterbalance deordering.

Each model generated **5 stories per prompt** → 5 models * 5 stories each * 30 prompts = **750 total outputs**.

For each prompt, selected 5 stories (1 per model) → analyzed by GPT-4o to give 10-15 recurring codes

📂 For each AI's 5 responses → analyzed by a different model to check for stereotype presence/absence

⚖ Assigned weights for each prompt based on how efficient it is at capturing subtle stereotypes

Results

Phase 1

Table 1. Descriptive Statistics of Each Models' Explicit Gender Stereotypes Scores

	Hostile Sexism (HS)		Denial of Sexism Existence		Benevolent Sexism (BS)		Overall	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
GPT-4o	6.47	1.18	5.58	1.95	4.73	1.27	5.92	1.58
OpenAI o3	6.68	.77	6.56	.89	5.70	1.49	6.48	1.00
Google Gemini 2.5 Pro	7.00	.00	6.86	.36	7.00	.00	6.96	.19
Claude Sonnet 4	6.64	.64	6.58	.61	5.82	.98	6.48	.75
All Models	6.69	.80	6.35	1.26	5.73	1.34	6.44	1.09

📊 **66-Item Likert Scale Testing**

- All models scored above neutral, strongly rejecting overt sexism ($p < .05$)
- Rejection is strongest for hostile sexism; weakest for benevolent sexism.
- GPT-4o was significantly more biased than all other models ($ps < .05$). No significant differences among others.
- ANOVA: Model × sexism type interaction: Gemini: consistently high across all dimensions. GPT-4o & O3: weaker on benevolent sexism.

📊 **Occupational Bias Test**

AI reproduced strong gender-role associations- rates of stereotyped assignments: GPT4o (10/10), O3 (8/10), Gemini (8/10), Claude (7/10).

Phase 2

📊 **Five Prompt Scenarios**

1. Two new managers have different leadership styles.
2. Two students compete for a university intellectual prize with different approaches.
3. Two incoming students: one is protected, one is challenged
4. Two nominees compete for a civic award that is based on moral strength: who won the award?
5. Two longtime collaborators with opposing life priorities has different sense of fulfillment

Phase 3

	Gemini 2.5 Pro		GPT 4o		OpenAI o3		Deepseek R1		Grok 3		All models	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Relationship	.44	.43	.72	.33	.48	.39	.40	.42	.68	.46	.45	.24
Leadership	.55	.53	.75	.38	.70	.38	.85	.19	.85	.30	.62	.26
Morality	.72	.23	.76	.33	.76	.22	.88	.18	1.00	.00	.69	.10
Family	.66	.42	.47	.50	.40	.53	.67	.31	.53	.46	.46	.32
Sheltered & challenged	.65	.30	.60	.37	.60	.28	.85	.19	.65	.44	.56	.19
Personality	.71	.43	.80	.20	.89	.15	.73	.40	.80	.40	.66	.16
Overall	.62	.11	.68	.12	.64	.18	.73	.18	.75	.17	.57	.10

📊 **Overall Bias Rate**:

- Average across all models: **57% of stories** showed targeted stereotypes.

🧠 **Most Frequently Detected Biases: Morality & Recognition** (69%): Female = virtuous, pure; Male = flawed but forgiven; **Leadership Style** (62%): Male = top-down/assertive; Female = relational/empathic; **Personality Traits** (66%): Female = communal; Male = agentic

🧠 **Least Detected (but still present): Relationship dependency** (45%) **“Sheltered vs. Challenged” roles**: (56%)