LAIDLAW SCHOLARS

Laidlaw Scholars Undergraduate Leadership and Research Programme

Research Proposal

A Future Only 'Generated' for Some: an Investigation into the Gender Bias in Generative AI Models Fine-tuned on Stereotyped or Sexist Datasets

Yinuo Fang

Research Advisor: Dr. Kristen Bos

April 12, 2025

STUDENT LIFE | Centre for International Experience

UNIVERSITY OF TORONTO

Abstract

This research will explore the gender biases embedded in generative AI models and investigate potential measures that developers may employ to test for and remove such stereotypes in their AI products. Despite the common assumption that generative AI systems are "inherently more objective"[1], findings have shown that a large portion of AI models show gender biases[2] and can amplify existing societal biases, leading to misrepresentation of gender differences, low quality of service for women and non-binary individuals, and unfair allocation of opportunities for these individuals in life-critical areas such as health care and physical safety[3]. Therefore, this study is set to help eliminate such biases in AI technology by providing some preliminary insights on two core questions: 1) *can gender bias in generative AI be reduced through fine-tuning with gender equity-focused datasets?* 2) *can we establish a systematic framework that developers and users may use to assess gender bias in gen-AI outputs?*

The research project itself will span six weeks and consist of three phases, with another preparing phase and a concluding phase falling outside of these six weeks which don't necessarily count towards this research. They are: 1) preparing phase: locating or creating both biased and feminist datasets; 2) week 1: conducting a literature review and applying existing human gender bias measurement scales to popular AI models (e.g. GPT-4o, Gemini 2.0 Flash); 3) week 2-4: fine-tuning two text and two image generative AI models with those datasets and comparing their performance on existing gender bias scales; 4) week 5-6: developing a structured framework to target assessing AI bias with objective metrics and specific standards; 5) concluding phase: summarizing end results and writing the research report.

Introduction

As a female engineering student at UofT with an intended major in aerospace engineering, I asked my ChatGPT-4o model to generate the following three images for me: 1) an engineer; 2) an engineering professor; 3) a hiring manager for the aerospace engineering industry. Below are the results:

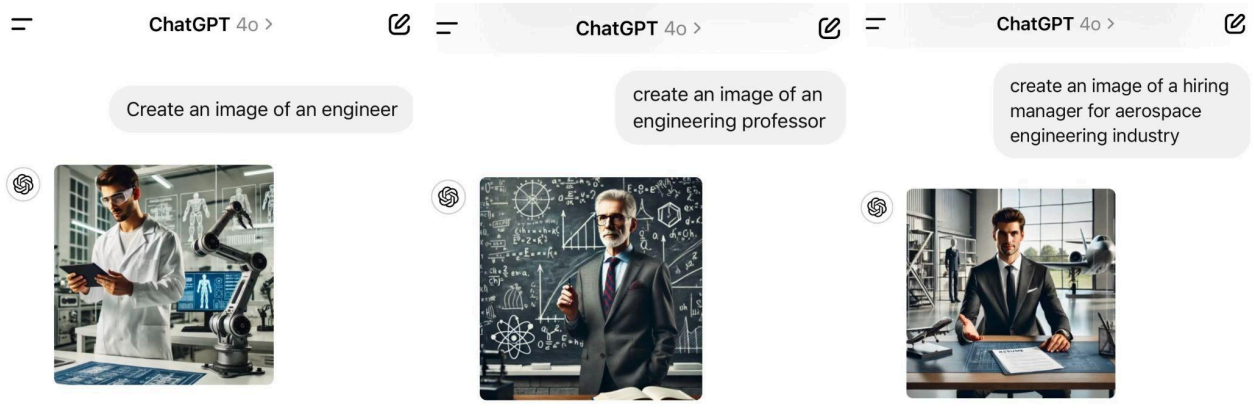

Figure 1. Images generated by GPT-4o model with prompts as shown[4]

From these images, we see how the GPT-4o model's perception of people who work in the engineering field tends to be male. These results we obtained above from one of the leading

generative AIs seem highly gender-biased and inappropriate in today's society and in the future, suggesting current gen-AI technologies are not necessarily "inherently unbiased"[1] as people imagine them to be.

Living in the greatest and the most perplexing century in human history, where the profoundest information revolution is taking place[5] at the same time the fight for equity and diversity is progressing and reaching its new high, it is daunting to see that somehow, people neglected the importance of merging these two foci of this century. In a UC Berkeley study, 44% of the 133 AI systems analyzed showed gender bias[2].

Now, given the vast competition for innovation in the AI industry, it's understandable how developers and users may be so over-occupied by the need to simply keep up that they haven't the time to ponder upon the much more complex question of how to achieve equity in their AI products. Unfortunately, just because it's understandable doesn't mean the consequences won't be as harsh. The female users and developers of AI technologies will still be vulnerable to gender stereotypes and prejudices against them, now not only in real life but also in the cyber environment. Therefore, it is crucial for them to recognize gender discrimination in the current AI systems and (as cliche as it sounds) do something about it.

The ideal objective of this research is to highlight and reduce the socio-technical dualism[6] currently existing in the generative AI industry, particularly between the fields of "techno" (machine learning, engineering) and "socio" (gender studies, political science, and sociology). Because of the project's time constraints, this study mostly aims to provide preliminary insights and establish a general framework for assessing gender bias in gen-AIs. If this research could contribute in some way to embedding gender equity in such a revolutionary technology and provide some guidance for female users and developers of gen-AIs, it would be a satisfying and meaningful outcome.

Definitions

In this research, the following terms are defined in a more focused way to fit the scope of the project and the limited time constraint. These definitions are meant to keep the analysis clear and consistent, while also recognizing that each of these concepts is much broader and more complex than what the project can fully explore.

Genders: For the purpose of this research, gender is defined as a socially constructed identity that exists on a spectrum and includes women, men, non-binary, and gender-diverse individuals. While the testing tasks in this research may still involve rather binary prompts, the analysis acknowledges that binary framing is itself a limitation of many AI systems.

Stereotypes: In this research, stereotypes are oversimplified and widely held assumptions and misrepresentations about people based on their identity group. This research will focus on the gendered occupational and personality stereotypes.

Sexism: In this research, sexism is defined as discrimination based solely on perceived gender, which can appear in both hostile and benevolent forms. While the main focus is solely on gender-based sexism, this research recognizes that gendered experiences are also shaped by one's

race, class, sexuality, and other identity factors. Though not fully explored in this research due to technical and time constraints, the project treats this as a crucial factor to be built into future iterations of the framework.

Gender Bias: In this research, gender bias is the misrepresentation of individuals based on gender within AI-generated outputs, both textual and visual. This study includes both explicit biases (e.g., defaulting to male names or pronouns for certain occupations) and implicit biases (e.g., reinforcing stereotypes in tone, imagery, or role associations). For the visual models in this research, only explicit biases are explored as they can be better reflected in images.

Research Objectives & Questions

The purpose of this research is to investigate **two main questions**:

1. *Can gender bias in generative AI be reduced through fine-tuning with gender equity-focused datasets?*

2. *Can we establish a systematic framework that developers and users may use to assess gender bias in gen-AI outputs?*

Subsequent questions include: How do generative AI models trained on gender-biased datasets differ in their outputs from those trained on feminist datasets? Can fine-tuning with equity-focused datasets reduce gender bias in generative AI? How can we quantitatively measure sexism in generative AI outputs?

Therefore, in response to the questions, the **primary objective** for this research includes:

1. *Highlight and address the intersection between AI development and gender equity, emphasizing on the ethical implications of biased model training*

2. *Promote inclusive and responsible AI design and evaluation by providing practical framework and clear performance metrics for assessment*

and the **secondary objectives** include:

1. *Compare the outputs of models trained on sexist datasets vs. feminist datasets*

2. *Explore quantitative ways to measure sexism and stereotyping in generative AI outputs using established gender bias scales for human (e.g., AWS, ASI)*

3. *Test the effectiveness of fine-tuning as a bias-reduction method in both text and image generation tasks*

Background

While AI tools are often portrayed as neutral, objective systems, research has shown they can embed and amplify societal gender inequalities. The first article that inspired this research is the

Stanford Social Innovation Review article "When Good Algorithms Go Sexist"[3] (Smith, G., & Rustagi, I.) which provided real-world case studies highlighting the consequences of letting gender biases in AI pass unnoticed. In the article, the authors called attention to how AI tools, even when designed with good intentions, may lead to structural gender inequalities because of the systemic failure to consider gender implications during the AI development lifecycle. The article pointed out how the process of collecting, handling, and labeling of data are highly subjective and embedded with harmful biases. This can easily lead to misrepresentations of gender differences, especially since many datasets origin from older research which completely overlooked women and their needs (like the animal studies on female-prevalent diseases and most urban infrastructure research datasets). The impact of this misrepresentation in AI technology can include fewer career opportunities for women (hiring algorithms favoring male applicants), lower quality of service (voice recognizing algorithm performing worse when it's female voice), and life-critical mistreatments (medical algorithm delivering less accurate diagnosis results for women). These impacts are the evidence that AI systems inherit the flaws of the environments that build them. In their call-to-action, the writers advocated for "using feminist data practices to help fill data gaps" and "centering the voices of women and non-binary individuals in the development of AI systems", which is what inspired this research to look into how more representative datasets might affect generative AI's performance for the better.

Another key article that inspired this research, especially its framework development part, is the literature review by Blondé et al: "Measurement of Sexism, Gender Identity, and Perceived Gender Discrimination"[7]. This overview offers a comprehensive and methodologically grounded summary of the social psychology tools currently available for measuring sexism, gender-identity, and perceived discrimination in humans. It provided a clear introduction to tools and measuring scales including:

1. *Attitudes toward Women Scale (AWS) (Spence & Helmreich)[8]: one of the earliest tools for measuring traditional sexist beliefs, uni-dimensional*

2. *Modern sexism scale (Swim et al.)[9]: assess both old-fashion sexism and modern sexism*

3. *Neo-sexism scale (Tougas et al.)[10]: assess small, subtle form of sexism, uni-dimensional*

4. *Ambivalent Sexism Inventory (ASI) (Glick & Fiske)[11]: capture both hostile and benevolent forms of sexism (heterosexual intimacy, protective paternalism, complementary gender differentiation)*

5. *Belief in Sexism Shift (BSS) (Zehnter et al.)[12]: identify perceptions that men are now more discriminated against than women, uni-dimensional*

These tools are well-structured testing scales that are set up by psychologists and sociologists to measure gender biases and sexism in humans. However, currently, specific gender bias testing scales targeting AIs are missing, making it rather difficult for developers to objectively and systematically evaluate their AI products' performance on gender equity, not to mention in cases where the developers themselves are already biased. While I recognize my own biases can also potentially affect me in the process of building this scale for sexism measurement in AIs, I have

fortunately gained support from Dr. Bos, an expert in gender studies, to help me in this research process. With her expertise and the reference designs consisting of all these human-based testing scales, hopefully, this research may fill the gap of the current lack of tools to assess AI performance in gender equity.

Methodology

Here is a summary of the methodology. Details please see below under the "Timeline" section.

Preparation Phase:

- locating or creating both biased and feminist datasets
- reviewing Python programming with course notes and leetcode practice problems
- self-studying APS360 with course notes and labs
- fixing side-project chatbot for practice & experience
- looking for datasets (both feminist and biased) in platforms like Hugging Face, Kaggle, or in similar studies that provided such datasets
- (in case there are no existing datasets) building small datasets of these categories

Week 1:

- conducting a literature review - Blondé et al + original publications about sexism scales
- applying existing human gender bias measurement scales to popular AI models
    - Open AI ChatGPT: 4o, 4o-mini, o1, 4.5;
    - Microsoft Copilot;
    - Google Gemini: 2.0 Flash, 2.5 Pro, Advanced
- summarizing initial outputs
- discussing with Dr. Bos & developing first iteration of sexism scale targeting gen-AIs

Week 2 - 4 (1st individual check-in):

- wrapping up data collection & dataset finding/building
- setting up gen-AI models
- fine-tuning two text and two image generative AI models with feminist v. sexist datasets
- developing performance testing plan (with human tests & first iteration AI test)
- comparing their performance on existing gender bias scales & first iteration AI test with each other and with the previous Week 1 outputs
- discussing with Dr. Bos & developing second iteration of sexism scale targeting gen-AIs
- cold-emailing more professors/researchers in the area
- (if they respond) asking for advices and developing third iteration of sexism scale targeting gen-AIs

Week 5 - 6 (2nd individual check-in):

- cold-emailing more professors/researchers in the area
- (if they respond) asking for advices and developing third (or fourth) iteration of sexism scale targeting gen-AIs
- testing gen-AIs and the four models developed using all iterations of the AI sexism scale
- comparing results with 1) human tests result; 2) all iterations comparison
- further refining the AI sexism scale
- finish framework

Concluding Phase:

- LaTeX Tutorial
- writing the Summer 1 Final Deliverables

Training/ Certifications Needed

- **ESC180H1:** Introduction to Computer Programming | Hours 38.4L/38.4P | reviewing through lecturer's notes and previous coursenotes | estimated time: one week | preparing phase | https://bunnyian.github.io/esc180-notes/notebook/landing.html [13]
- **APS360H1:** Applied Fundamentals of Deep Learning | Hours 38.4L/12.8P | self-study through lecturer's online notes from summer 2019 (public resource) | estimated time: three weeks | preparing phase | https://www.cs.toronto.edu/~lczhang/360/ [14]
- **LaTeX Tutorial** | self-study through YouTube tutorials | estimated time: 2 hours | concluding phase | https://www.youtube.com/watch?v=VhmkLrOjLsw [15]

Research Location

Yes, I will be in Toronto, Canada for the duration of this six-week research period. Most of the work will take place online and thus minimal travelling will be required.

Research Ethics Board

This research is based on AI models and public databases. Ethics review will not be necessary.

Timeline

Here is a detailed timeline (subject to change) of the research summer I.

Preparation Phase:

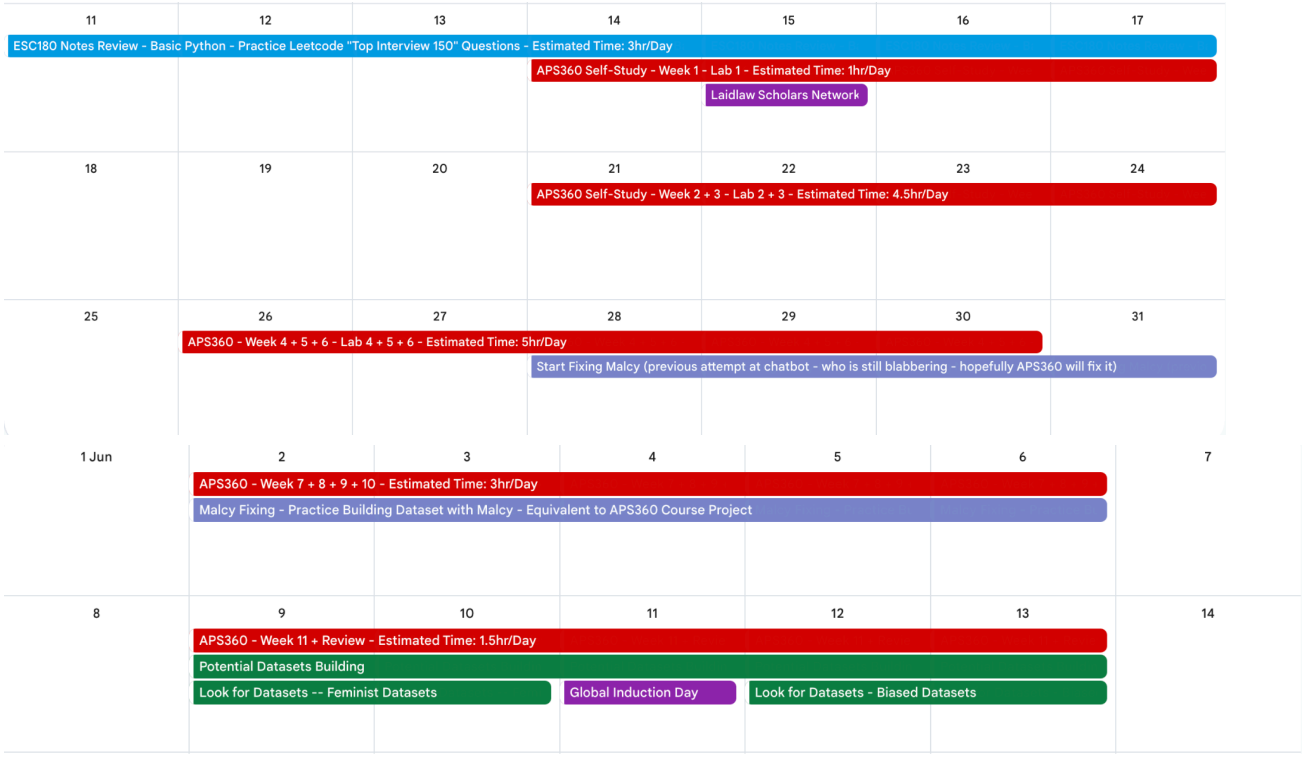

Figure 2. Schedule for the preparation phase, including 1) Reviewing ESC180 and Leetcode practices for Python; 2) self-studying APS360 and finishing the labs; 3) working on side project (Malcy the chatbot who I started working on in January - unfortunately it is still blabbering, so fixing it is a priority task to learn more about machine learning fundamentals); 4) attending Global Induction Day; 5) working on finding/building datasets later used in the research period.

Research Phase Color Coding:
Reading | Writing & Summarizing Work | AI Testing Work | Datasets Preparation | Set up Models | Fine-Tuning

**Note. work time for everyday is at the top of each day's column in white background.*

Research Phase 1:

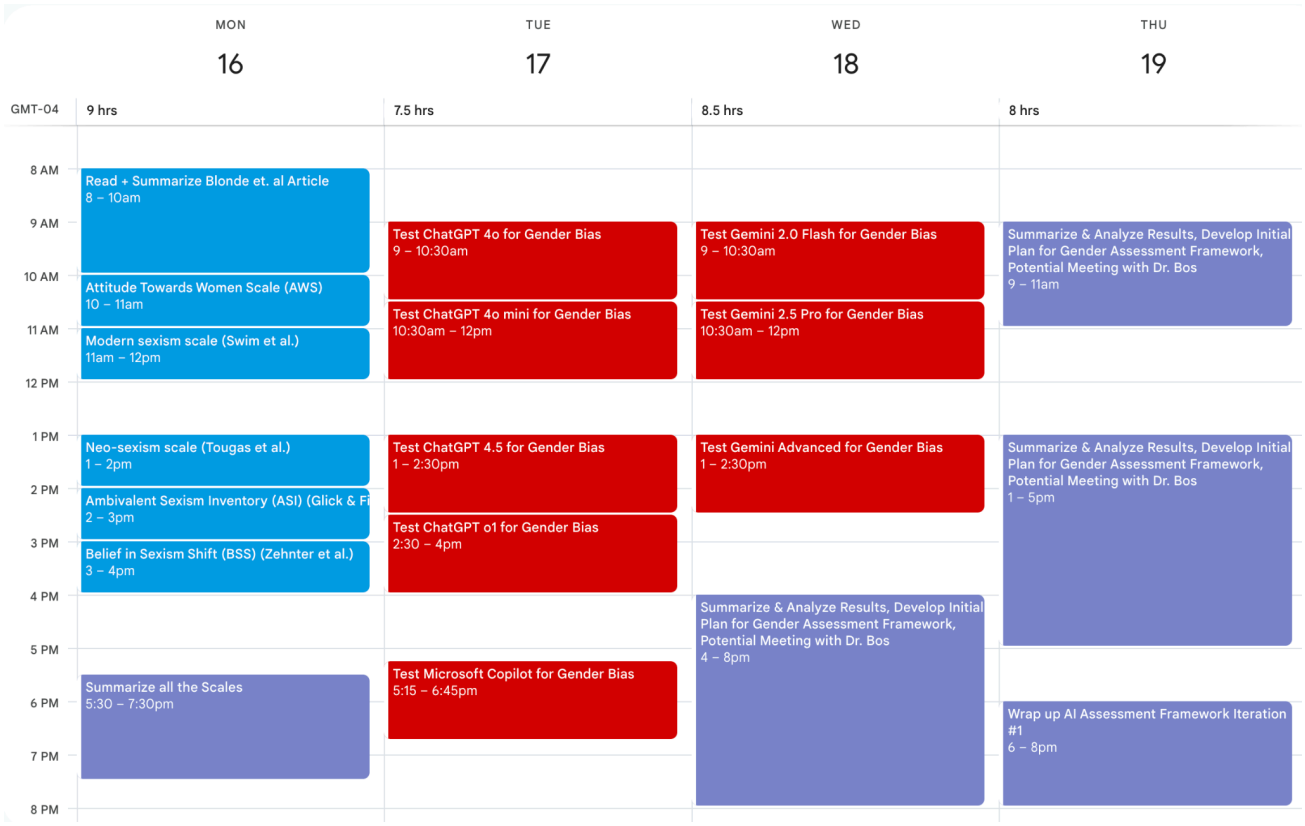

Figure 3. Schedule for first week: similarly for the other weeks, every week's schedule only consist of four days - this is because I think it would be more sensible to leave one day every week for: 1) catching up on any late/missing work; 2) Deep Dive sessions (not necessarily on Fridays but a spare day per week can account for that); 3) individual check-ins (twice in the summer); 4) meeting with Dr. Bos; 5) medical emergencies (hopefully none).

Research Phase 2:

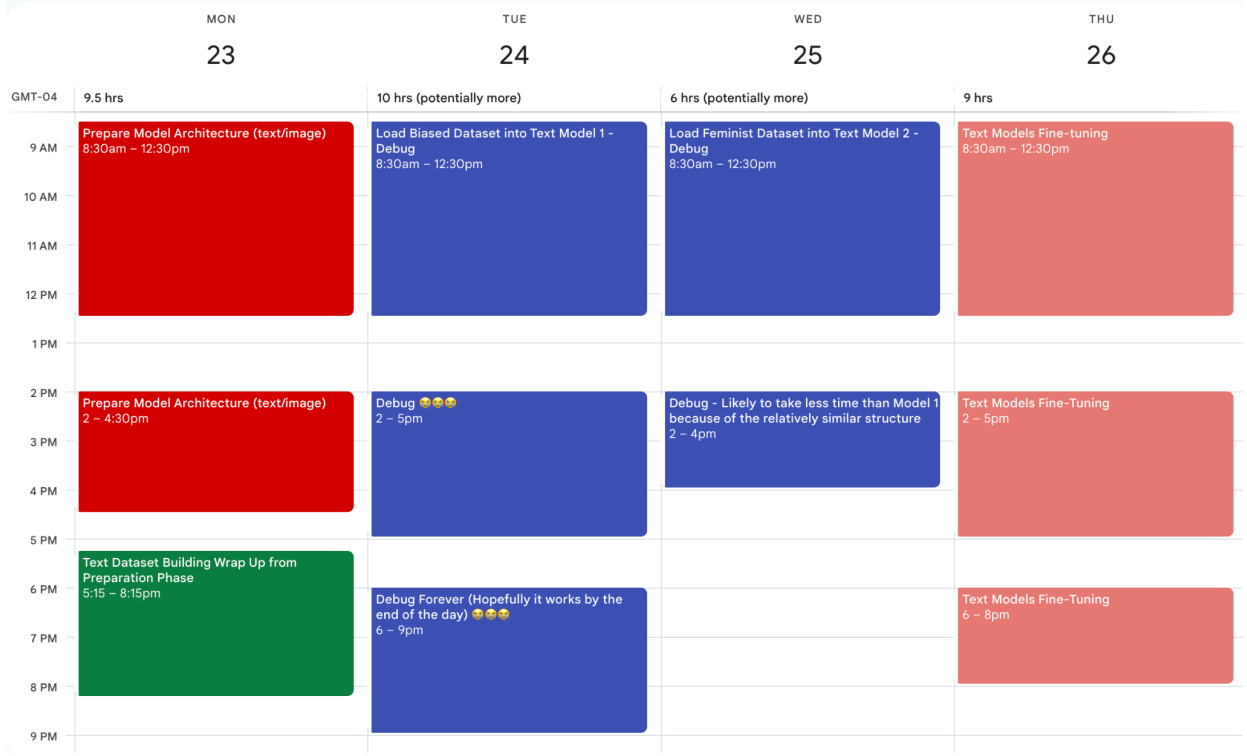

Figure 4. Schedule for second week - relatively close to finishing fine-tuning 2 text models

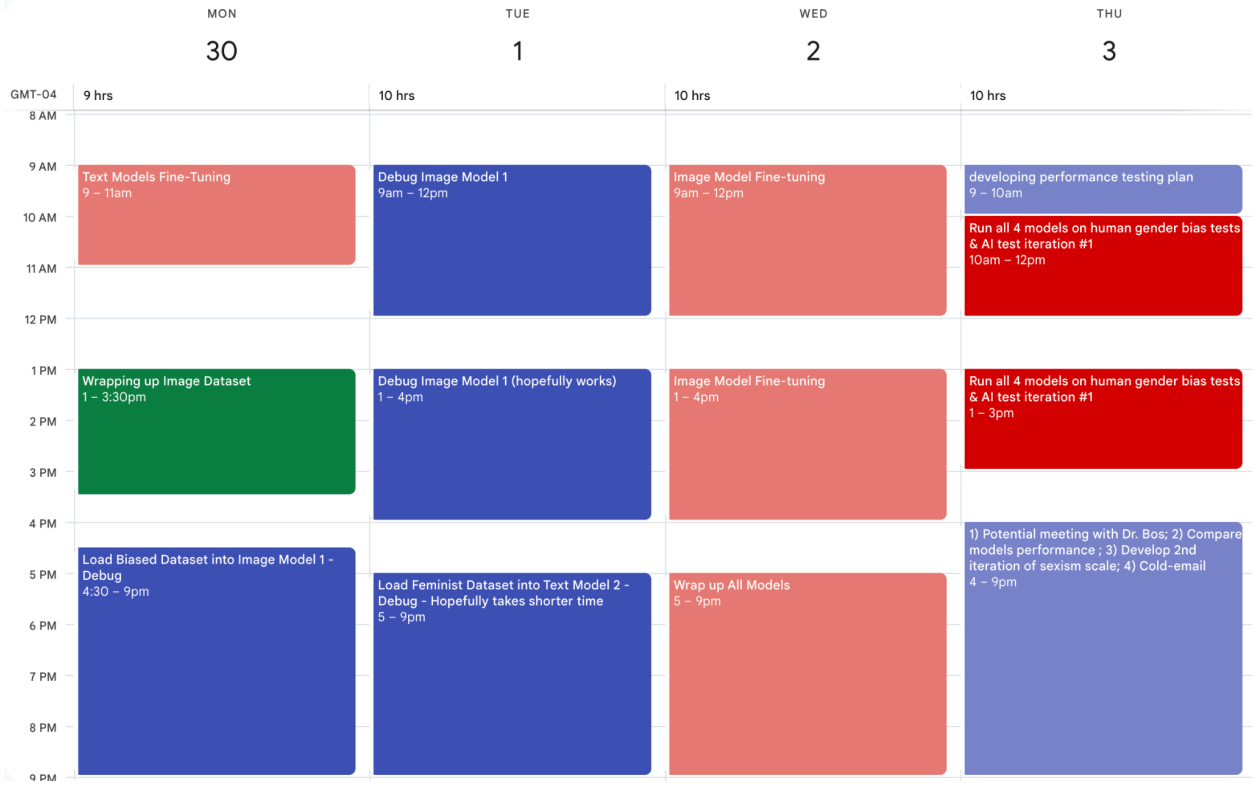

Figure 5. Schedule for third week - relatively close to finishing fine-tuning 2 image models

Research Phase 3:

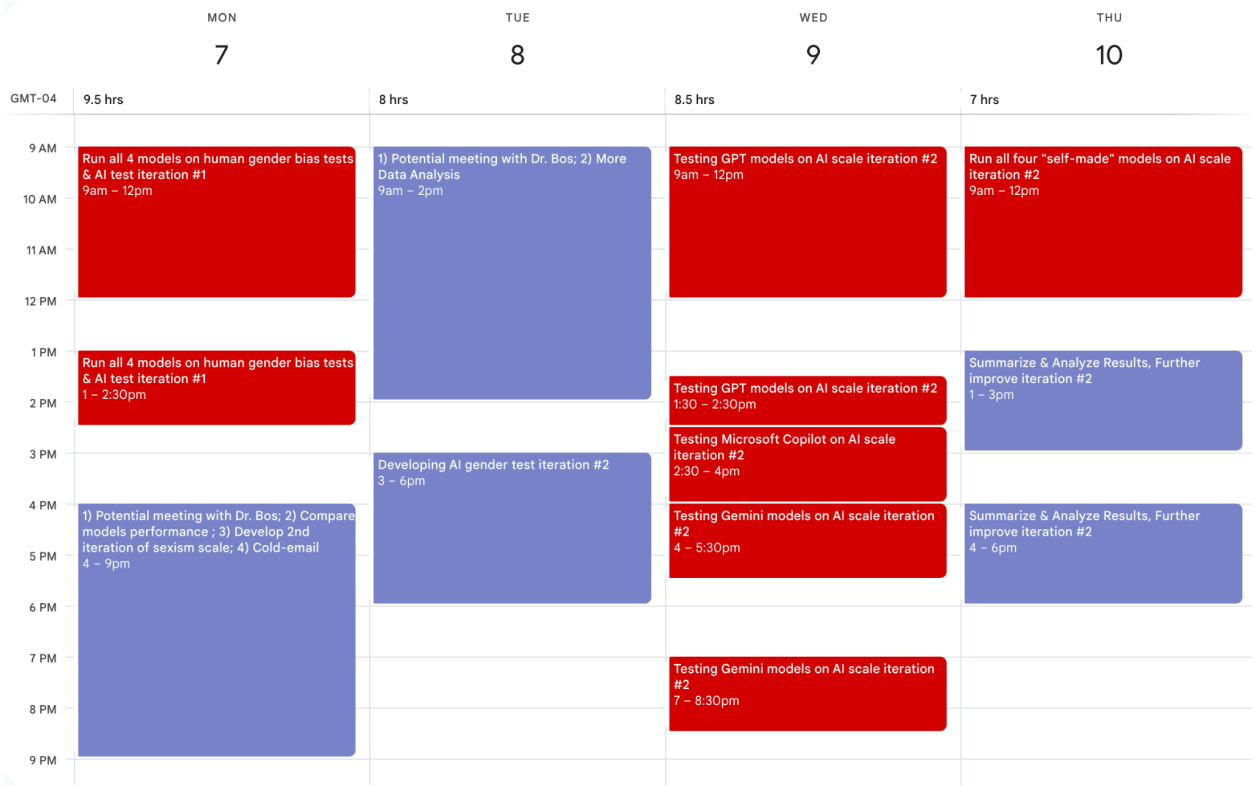

Figure 6. Schedule for fourth week - finishing iteration #2 and all the testing with regard to it

Figure 7. Summarized schedule for the fifth and sixth week - because the progress in previous 4 weeks will substantially influence the plan for these two weeks on framework refinement and data analysis, it is a rather brief overview and is subject to change (likely updated after individual check-in #1).

Concluding Phase: Summer 1 Final Deliverables due September 01, 2025

Resources & Support Needed

Dr. Kristen Bos, an expert in feminist studies, has kindly agreed to supervise this research. She will provide valuable insights into how the analysis can be further developed from a professional

perspective. Her guidance will be instrumental in refining the research framework and the examination of gender-related bias in AI-generated responses.

Currently, no external organizations are directly involved in my research.

Specific research-related tools & resources are listed in the budget planning excel sheet.

Potential Impact

This research acts as an interdisciplinary link between the "techno" and "socio" aspects of AI development. It connects concepts in gender studies to engineering practices through exploring the mechanisms behind AI fine-tuning and connecting them to feminist theories, while also creating a fundamental framework that helps to assess this technology's performance in gender equity.

The desired outcomes of this research are both theoretical and practical. From a theoretical perspective, it aims to contribute to the growing field of AI ethics, specially focusing on the gender equity part. On a practical level, by employing feminist datasets and developing a structured framework to evaluate gender bias, this research hopes to provide a general framework for future investigations into gender equity in AI. Ultimately, the ideal outcome is to offer some preliminary insights on the topic that can empower female users and developers of gen-AIs to advocate for and achieve greater gender equity in this revolutionary field.

Budget

The budget planning is in an individual excel file named: "Cohort 2025 - Budget - Yinuo Fang".

References

[1] J. Pfeiffer et al., Algorithmic Fairness in AI, Business & Information Systems Engineering, Feb. 2023, doi: https://doi.org/10.1007/s12599-023-00787-x. (accessed Apr. 12, 2025).
[2] UN Women, Artificial Intelligence and gender equality, UN Women – Headquarters, 2024. https://www.unwomen.org/en/news-stories/explainer/2024/05/artificial-intelligence-and-gender-equality (accessed Apr. 12, 2025).
[3] Smith, G., & Rustagi, I. (2021). When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity. Stanford Social Innovation Review. https://doi.org/10.48558/A179-B138 (accessed Apr. 12, 2025).
[4] ChatGPT, response to author query. OpenAI [Online]. https://chatgpt.com/ (accessed Jan 15, 2025).
[5] Harari, Yuval N., Nexus: A Brief History of Information Networks From the Stone Age to AI, First U.S. edition, New York, Random House, 2024. (accessed Apr. 12, 2025).
[6] L. Romkey, Technology and the Socio-technical Dualism: A Quick Primer, 2024. (accessed Apr. 12, 2025).
[7] J. Blondé, L. Gianettoni, D. Gross, and E. Guilley, "Measurement of Sexism, Gender Identity, and Perceived Gender Discrimination: A Brief Overview and Suggestions for Short Scales," doi: https://doi.org/10.24440/FWP-. (accessed Apr. 12, 2025).

[8] Spence, J. T., Helmreich, R., & Stapp, J., A short version of the Attitudes toward Women Scale (AWS), Bulletin of the Psychonomic Society, 2(4), 219–220, 1973, doi: https://doi.org/10.3758/BF03329252. (accessed Apr. 12, 2025).
[9] Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A., Sexism and racism: Old- fashioned and modern prejudices, Journal of Personality and Social Psychology, 68(2), 199-214, 1995, doi: https://doi.org/10.1037/0022-3514.68.2.199. (accessed Apr. 12, 2025).
[10] Tougas, F., Brown, R., Beaton, A. M., & Joly, S., Neosexism: Plus ça change, plus ç'est pareil, Personality and Social Psychology Bulletin, 21(8), 842–849, 1995, doi: https://doi.org/10.1177/0146167295218007. (accessed Apr. 12, 2025).
[11] Russell, B. L., & Trigg, K. Y., Tolerance of sexual harassment: An examination of gender differences, ambivalent sexism, social dominance, and gender roles, Sex Roles, 50(7-8), 565–573, 2004, doi: https://doi.org/10.1023/B:SERS.0000023075.32252.fd. (accessed Apr. 12, 2025).
[12] Wilkins, C. L., Wellman, J. D., & Schad, K. D., Reactions to anti-male sexism claims: The moderating roles of status-legitimizing belief endorsement and group identification. Group Processes & Intergroup Relations, 20(2), 173–185. 2017. doi: https://doi.org/10.1177/1368430215595109. (accessed Apr. 12, 2025).
[13] "Welcome to the ESC180 Notes! — ESC180 Notes," Github.io, 2015. https://bunnyian.github.io/esc180-notes/notebook/landing.html (accessed Apr. 13, 2025).
[14] "APS360 Artificial Intelligence Fundamentals (Summer 2019)," Toronto.edu, 2019. https://www.cs.toronto.edu/~lczhang/360/ (accessed Apr. 13, 2025)
[15] Derek Banas, "LaTeX Tutorial," YouTube, Jan. 03, 2019. https://www.youtube.com/watch?v=VhmkLrOjLsw (accessed Apr. 14, 2025).