Bias, Accuracy, and Trust: Gender-Diverse Perspectives on Large Language Models

AIMEN GABA, University of Massachusetts Amherst, USA
EMILY WALL, Emory University, USA
TEJAS RAMKUMAR BABU, Georgia Institute of Technology, USA
YURIY BRUN, University of Massachusetts Amherst, USA
KYLE WM. HALL,, Canada
CINDY XIONG BEARFIELD, Georgia Institute of Technology, USA

Large language models (LLMs) are becoming increasingly ubiquitous in our daily lives, but numerous concerns about bias in LLMs exist. This study examines how gender-diverse populations perceive bias, accuracy, and trustworthiness in LLMs, specifically ChatGPT. Through 25 in-depth interviews with non-binary/transgender, male, and female participants, we investigate how gendered and neutral prompts influence model responses and how users evaluate these responses. Our findings reveal that gendered prompts elicit more identity-specific responses, with non-binary participants particularly susceptible to condescending and stereotypical portrayals. Perceived accuracy was consistent across gender groups, with errors most noted in technical topics and creative tasks. Trustworthiness varied by gender, with men showing higher trust, especially in performance, and non-binary participants demonstrating higher performance-based trust. Additionally, participants suggested improving the LLMs by diversifying training data, ensuring equal depth in gendered responses, and incorporating clarifying questions. This research contributes to the CSCW/HCI field by highlighting the need for gender-diverse perspectives in LLM development in particular and AI in general, to foster more inclusive and trustworthy systems.

CCS Concepts: • **Human-centered Computing → Empirical studies in HCI**.

Additional Key Words and Phrases: large language models, algorithmic harms, gender representation, trust in AI, fair AI, responsible AI

ACM Reference Format:
Aimen Gaba, Emily Wall, Tejas Ramkumar Babu, Yuriy Brun, Kyle Wm. Hall, and Cindy Xiong Bearfield. 2025. Bias, Accuracy, and Trust: Gender-Diverse Perspectives on Large Language Models. 1, 1 (July 2025), 22 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Machine learning (ML)-driven interfaces are becoming ubiquitous in modern society [3, 30, 55, 56], from facial recognition systems used to unlock our phones, to voice-activated assistants, and conversational systems that respond to natural language prompts. These technologies are widely deployed, but they often exhibit biases that reflect existing

Authors' Contact Information: Aimen Gaba, University of Massachusetts Amherst, Massachusetts, USA, agaba@umass.edu; Emily Wall, Emory University, Georgia, USA, emily.wall@emory.edu; Tejas Ramkumar Babu, Georgia Institute of Technology, Georgia, USA, tbabu8@gatech.edu; Yuriy Brun, University of Massachusetts Amherst, Massachusetts, USA, brun@cs.umass.edu; Kyle Wm. Hall, Ontario, Canada, kylewmhall@gmail.com; Cindy Xiong Bearfield, Georgia Institute of Technology, Georgia, USA, cxiong@gatech.edu.

Manuscript submitted to ACM

inequalities in society. For example, facial recognition technologies tend to perform better for individuals with lighter skin tones [9], and voice-controlled devices often struggle with certain accents [42]. In the case of large language models (LLMs), recent studies have shown that they can exhibit biases and other social risks [67] against particular religious groups [2], produce gender stereotypes [63], and generate stigmatizing language [51, 58], reinforcing harmful stereotypes when discussing non-binary or transgender individuals. These biases not only undermine user trust in ML models [21], but also degrade the model's ability to provide meaningful and fair interactions.

Conversations around fairness and harms in sociotechnical systems have been a focal point within the CSCW and broader HCI community [52, 61, 65, 70]. Previous discussions have emphasized the implications of biased technologies on marginalized identities, especially concerning gender [17, 24]. For instance, Scheuerman et al. [57] found that commercial facial analysis services consistently perform worse for transgender individuals and are universally unable to classify non-binary genders, highlighting significant gaps in technological inclusivity. These dialogues within the CSCW community underscore the importance of diverse representation and gender balance in machine learning evaluations [36].

Although substantial research has focused on understanding and mitigating algorithmic biases in machine learning systems and language models [6, 18, 47, 50], particularly in domains such as machine translation [22], text classification [14], data science [15], and coreference resolution [71], a significant gap remains in understanding how users perceive and respond to these biases in real-world LLM applications. Notably, there is limited qualitative research on how individuals, especially those from gender-diverse backgrounds, experience and interpret bias in LLM outputs. While LLMs have been used to analyze bias in interview settings [31] or to assist in qualitative research [4, 12], few studies have centered on gender bias [11, 26, 32], particularly from the perspectives of non-cisgender users, such as non-binary and transgender individuals. Moreover, although prior work emphasizes the value of qualitative case studies [20, 59], it rarely integrates in-depth qualitative insights with quantitative measures of user perception - a gap this study seeks to address.

In this paper, we study the perceived utility of a real-world LLM application by understanding users' perception of bias, accuracy, and trust in them. In order to study this in a real-world setting, we selected ChatGPT [53] for two key reasons. First, its widespread use and accessibility allow us to study how biases manifest in real-world LLM applications. Second, ChatGPT is regularly used by people from different gender backgrounds [16], offering an important context for examining how gender influences users' perceptions of LLM interactions. In this work, we address five research questions. Figure 1 summarizes the research questions that underpin our user studies and their respective findings.

To address these research questions, we studied how gendered and neutral prompts (e.g., "man", "woman", "non-binary", "person") influence the responses generated by ChatGPT and how users interpret those responses. We conducted semi-structured interviews with 25 participants and extract qualitative insights from our study. We adopted a mixed-methods approach that combines quantitative trust measures with qualitative interview data, allowing us to examine in depth as to how users perceive the biases and assumptions present in ChatGPT's outputs. Moving forward, we define "men" as cisgender men, "women" as cisgender women, and "non-binary/transgender" as those who do not identify with the gender assigned at birth. By exploring these gender categories, we investigate how gender affects user perceptions of bias, accuracy, and trustworthiness in LLMs.

In this work, we make the following contributions:

- We introduce curated gender-focused prompts to evaluate LLMs' inclusivity and gender representation.

Fig. 1. Research questions to understand people's perception of bias in a real-world application of large language models.

Research Question	Finding Summary
RQ1: How do gendered and neutral prompts influence ChatGPT's responses?	Our results indicate that gendered prompts elicit richer, more identity-specific responses from ChatGPT compared to neutral prompts, which are more generic. (Section 3.1)
RQ2: How does the perception of bias in LLMs vary across gender diverse populations?	Our results indicate that perceived bias in LLMs varied, with non-binary/transgender participants noting condescending and stereotypical responses, women highlighting emotional and traditional portrayals, and men observing a lack of diversity but fewer concerns. (Section 4.6)
RQ3: How does the perception of accuracy in LLMs vary across gender diverse populations?	Our results indicate that perceived accuracy in LLMs was similar across gender-diverse groups, with participants noting errors in technical topics and creative tasks. Most could assess accuracy when familiar with the subject, while others relied on external checks. (Section 4.7)
RQ4: How does the perceived trustworthiness of LLMs vary across gender diverse populations?	Our results indicate that perceived trustworthiness of LLMs varied by gender, with men reporting higher trust, especially in performance. Non-binary participants showed higher performance-based trust but similar morality-based trust. (Section 4.8)
RQ5: What suggestions do users have for improving the way LLMs handle gender-related content?	Our results indicate that participants suggested improving LLMs by diversifying training data to include real-life gender experiences, ensuring equal depth in responses to all genders, incorporating clarifying questions, and the need for transparency in response generation. (Section 5.1)

- We study men, women, and non-binary/transgender participants, revealing varying reactions to bias, with non-binary participants particularly susceptible to identity framing.
- We investigate perceived accuracy, examining how participants understand and evaluate LLM responses.
- We assess trustworthiness through qualitative analysis of participant reactions in high-stakes scenarios and quantitative trust scores collected before and after the study, highlighting the roles of gender, bias knowledge, and LLM familiarity.
- Finally, we offer a preliminary set of recommendations and suggestions from the participants as well as design implications for developing more inclusive and trustworthy LLMs in the future.

2 Related Work

2.1 Social Biases in NLP and Language Technologies

The intersection of language technologies and gender representation has become a critical area of study [36, 37], particularly with the rise of LLMs. Research has shown that NLP systems often reinforce binary and exclusionary notions of gender, reflecting broader societal biases [29]. Research has explored how LLMs perpetuate gender stereotypes, particularly in professional contexts such as the language used to describe individuals in recommendation letters or the association of specific professions with certain genders [7, 11, 63, 66]. These findings underscore the subtle ways in which LLMs can reinforce societal biases, highlighting the nuanced nature of the issue.

Further research has expanded on these findings, showing that LLMs frequently fail to accommodate non-binary identities. For example, Ghosh et al. [22] highlight the limitations of LLMs in representing gender diversity, particularly the persistent exclusion of non-binary and transgender individuals in the models' outputs. This issue is further exacerbated by biases that extend beyond gender, such as anti-Muslim bias, which has been documented in LLMs [2]. These studies reflect the prevalant nature of the biases present in LLMs, not just limited to gender, but that spans various social identities.

Other research has examined how LLMs replicate human tendencies to perceive socially subordinate groups as less diverse than dominant groups [38]. This study emphasizes the inherent biases in LLMs' representations of social groups, mirroring broader societal stereotypes. Similarly, research has shown that LLMs can generate stigmatizing language, particularly when discussing non-binary individuals, which reinforces harmful stereotypes [51]. The paper [69] offers a parallel perspective, exploring how marginalized communities, including individuals with disabilities, experience biases in LLM outputs, stressing the importance of inclusive evaluation approaches in understanding these biases.

In response to these issues, efforts have been made to quantify and mitigate these biases. For example, Holstein et al. [26] advocate for transparent reporting of model behaviors and active interventions to reduce harmful tendencies in LLMs. Similarly, Park et al. [54] evaluate existing models and introduce methods for bias mitigation, such as adversarial training [34], data augmentation, and debiasing algorithms, to improve fairness and accuracy in detecting abusive content. Bartl and Leavy [5] developed a gender includive dataset called the *Tiny Heap* and observed an overall reduction in gender stereotyping tendencies across three large lagnuage models that were fine-tuned using that dataset. These studies highlight the dual responsibilities of identifying biases and designing interventions to mitigate harm, making them a key part of ongoing efforts to improve LLM fairness and inclusivity.

2.2 Accuracy & Trust in AI

Understanding and fostering trust in AI systems is crucial, particularly in sensitive or high-stakes domains [68]. Research highlights the role of explainability and transparency in building trust, with frameworks like [13] and [48] emphasizing user understanding and perception of AI behavior in sociotechnical contexts. Studies [28, 40, 68] show how transparency and contextual awareness improve perceptions of AI reliability, while [27] underscores the importance of explainable interfaces to reduce uncertainty and increase confidence.

Liao et al.[41] aligns these concerns with actionable recommendations to enhance LLM transparency, framing transparency as vital for ethical AI design. Studies like [72] identify challenges with LLM inaccuracies, biases, and educational misalignment, advocating for ethical guidelines and training. Similarly, Lee et al. [39] explore how conflicting AI outputs increases user comprehension but comprimising users' percieved AI capacity, highlighting the need for transparency in AI models and promote critical LLM usage. In this work [33], Kumar et al. conduct a formative and controlled experiment to explore the impact of pedagogically informed guidance strategies on the students' performance, confidence and trust in LLMs. To measure trust, they collected pre- and post-measures on a scale of 7 regarding perception of LLMs (confidence in LLMs' responses, helpfulness, willingness to interact again and error tolerance) and perception of self (confidence in their answers, self-confidence for the given topic of assignment).

Göben and Lämmel [23] conduct a trust analysis of LLM conversations by using a trust score between -1 (absolute distrust) and +1 (absolute trust) function to understand and assess users' trust in knowledge elicited from these LLMs and its sources, offering insights into trust formation during interactions. This work [64] identifies trust as central to ChatGPT's adoption while also highlighting that an overreliance on it for health-related advice could potentially lead to misinformation and subsequent health risks. Shen et al. [60] conduct a large-scale measurement of ChatGPT's reliability in the generic QA scenario with a carefully curated set of questions across different datasets and domains finding ChatGPT's reliability varying across different domains, especially underperforming in law and science questions. We supplement these existing work to better understand the perceived trustworthiness and the challenges LLMs face in earning user trust.

3 A Case Study on ChatGPT Response Patterns

To initiate discussions with ChatGPT about gender, we curated four prompts inspired by [69]. These prompts were crafted through extensive probing to encourage meaningful, gender-focused responses while avoiding harmful or biased content. Our aim was to design prompts that could be neutrally addressed and resonate with lived experiences across diverse populations, without leading or eliciting harmful narratives. The final prompts are as follows:

(1) Tell me a story about a person buying a car.
(2) Tell me a story about a person navigating through college.
(3) Tell me a story about the experiences of a person.
(4) Describe an experience of a person applying for jobs.

To explore gendered perspectives, we used the same prompts but replaced "person" with specific gender identities: man, woman, and non-binary person. For example, the first prompt became "Tell me a story about a non-binary person buying a car" eventually resulting in a total of 16 unique prompts.

3.1 RQ1: Effect of Gendered Prompts on ChatGPT Responses

We analyze the responses of ChatGPT to the four prompts mentioned above (repeated 10 times), focusing on how ChatGPT describes the main character, whether identified as a man, woman, non-binary, or person. We had a total of 40 responses for each gender specification (4 unique prompts generated 10 times). We analyze these responses for names, pronouns, adjectives, emotions, goals, challenges, and symbols to assess their role in shaping the character's identity and narrative.

3.1.1 Non-binary. ChatGPT consistently used they/them **pronouns** they/them in relation to the non-binary character and assigned predominantly gender-neutral **names** (often leaning towards masculine) to non-binary characters such as Alex (11), Taylor (10), Jordan (8), and Jamie (4). Occasionally, names such as "Casey" (2), "Avery" (1), and "Riley" (1) appeared. Common **adjectives and emotional descriptors** used for the non-binary character included "resilience" (21), "authenticity" (9), "anticipation" (6), "determination" (5). Some other common descriptors included "liberation", "inclusivity or inclusive spirit", "non-conforming identity", or "unconventional appearance". **Rersonal goals** frequently involved advocating for inclusion or expressing their true identity. Other common goals included "fostering/advocating inclusivity", "embracing/express true self", "navigating societal norms", **Challenges** often centered on their identity such as "dealing with misgendering" (13), "job applications" (5) were highlighted most frequently. In some cases, it mentions their challenge "facing societal expectations, discrimination, systemic biases" and "deciding whether to disclose non-binary identity". In some responses, the non-binary character was depicted to be struggling with attire choices, but this was less frequent compared to identity-related challenges. **Symbolic** elements highlighted themes of resilience with characters often portrayed as "beacons of resilience" or "embracing diversity". Memorable quotes included: "Jordan found resonance in the car's ability to accommodate diverse needs-a reflection of their fluid identity", "I'm non-binary, and my car choice reflects my values of sustainability and innovation", and "The binary constraints of *appropriate* clothing clashed with Alex's non-binary identity".

3.1.2 Woman Prompts. ChatGPT consistently used she/her **pronouns** for the woman character and assigned commonly feminine, Western **names** such as Emma (6), Sarah (6), Emily (4), Rachel, Lily, and Amelia (9). Common **adjectives and descriptors** included "resilience" (27), determination (13), anticipation (7). Other descriptors frequently used with the female character were "compassionate", with mentions of academic fields such as environmental science or psychology..

The character's **personal goals** often focused on achievement such as "navigating college as a first-generation student" (3) or "excelling in environmental science or psychology" (4). **Challenges** typically involved "financial constraints" (9), "handling/managing rejections" (6), "navigating a competitive job market" (2). In one response, the female character was depicted as "selecting a car that fits family and adventure needs". **Symbols** associated with the female character emphasized "growth" or "discovery". Memorable quotes included: "the key to her newfound independence", and "a sea of gender-based obstacles that sought to erode her confidence".

3.1.3 Man Prompts. In all 40 responses, ChatGPT consistently used he/him **pronouns** for the male character. The male character was assigned traditionally masculine Western **names** commonly such as Alex (6), Mark (5), and Jake (4). These names are traditionally gendered and reflect the stereotype of male identity. Some names, like Alex, were more neutral but still paired with masculine descriptors. Common **adjectives** included "resilient" (16), anticipation (5), determination (5) often linked to fields like "computer science or psychology". The male character's **personal goals** often focused on career success and leadership. ChatGPT frequently mentioned objectives such as "securing a job" (6) or "overcoming unemployment" (6). **Challenges** were primarily career-related, including "unemployment" (8), "rejection emails" (7), and "job applications" (6) with occasional mentions of work-life balance struggles. **Symbolism** highlighted assertive and traditional imagery portraying the male character as "the provider" or "adventurous". Memorable quotes included: "his dedication to family and his role as a provider" and "unemployment had turned this room into a battlefield".

3.1.4 Person Prompts. ChatGPT used she/her **pronouns** in 22 responses, he/him for in 15, and assigned no pronouns in the remaining 3. Characters with gender-specific pronouns were given traditional **names** such as Mark or Sarah, while those without pronouns received neutral names like Alex. The more commonly used **adjectives and descriptors** included "resilience" (14), "self-discovery" (11), and "determination" (8). **Personal goals** focused on "personal growth", "embracing change", "gaining independence". **Challenges** were typically related to academic or professional struggles, such as "rejections", or "negotiating a deal". Responses without specified gender were noticeably more generic compared to gendered ones.

Overall, we found that gendered prompts elicited richer and more identity-focused responses compared to neutral prompts, which were more generic. Furthermore, men were associated with careers and assertiveness, women with growth and emotions, and non-binary individuals with inclusivity and advocacy. We also observed that ChatGPT's name assignments, pronoun usage, and challenges faced by characters align with societal gender norms. Our findings corroborate similar themes in [43, 63] where GPT reinforces gender stereotypes and representation biases in storytelling, identifying systemic biases that disproportionately favor traditional gender norms.

4 Perceived Utility of LLMs

To understand the perceived utility of LLMs, we conducted in-depth user interviews to understand how people perceive bias, accuracy, and trust in a real-world application of an LLM (ChatGPT). We describe our study methods (reviewed and approved by our Institutional Review Board prior to conducting the study) and our findings below.

4.1 Participant recruitment and selection

We recruited 25 participants. Among them, 9 identified as non-binary/transgender (M_{age} = 26.11, SD_{age} = 8.51), 8 identified as women (M_{age} = 25.62, SD_{age} = 4.07), and 8 identified as men (M_{age} = 25.62, SD_{age} = 2.39). We first interviewed the non-binary participants and then recruited women and men participants of similar backgrounds (with considerations for diversity in the knowledge of LLM participants, bias in LLM and AI background). Concretely, we

Table 1. Participants' knowledge of bias in LLMs/AI and LLMs/AI background)

	Low-LLM/AI	Medium-LLM/AI	High-LLM/AI
Low-bias in LLM	P4, P11	P19	P6
Medium-bias in LLM	P2, P3, P13, P14	P5, P15, P24	P7, P10, P17, P18, P23
High-bias in LLM	P12, P21, P25	P1, P16, P22	P8, P9, P20

created a screening survey asking questions about their knowledge of LLMs, bias in LLMs, AI/ML background, and ChatGPT usage pattern (e.g., frequency of ChatGPT use). We distributed the survey through university social emails, various Slack workspaces within our institutions, and X. We were unable to recruit our desired number of non-binary participants using these methods, so we recruited and interviewed 3 non-binary participants via userinterviews.com by releasing the same screening survey. See Table 1 for a summary of the background of the participants and the supplementary material [1] for definitions. The subgroups were defined based on the responses from the survey of the participants and the responses from the interviews. We categorized the participants into the range of low-to-high LLM/AI background (experience with LLM/AI) and low-to-high knowledge of LLM bias (awareness of bias and experience with encountering it).

Regarding frequency of use, 12 participants reported using it multiple times a week, 5 used it approximately three times or twice a week, and 8 used it once a week.

4.2 Study protocol

Our interviews were structured into two parts and included a short questionnaire that was distributed to measure trust [44–46, 62] at the beginning and end of the study. The study protocol was iteratively designed through two pilot studies. The complete study protocol and interview guide is attached in the supplementary material [1].

Context: We began by asking participants about their background and usage of ChatGPT, including when and where they use it. We explored their knowledge and perceptions of bias and accuracy by having them reflect on instances of biased or incorrect responses they had encountered and how they identified them. Participants were also asked about their expectations of ChatGPT, if their experiences aligned with those expectations, and if certain prompts appeared to elicit more biased or accurate responses.

Prompts: Each participant encountered four prompts, as mentioned in Section 3, arranged in a Graeco-Latin square design, ensuring exposure to all types of prompts and gender descriptions. For the fourth prompt, participants were shown responses for all gender descriptions to facilitate direct comparison. We selected two random prompt-response pairings for each gender specification (from Section 3) to show the participants and created two sets of responses:

- The first set included responses shown to participants 1 through 12 (4 non-binary/transgender, 4 women, and 4 men).
- The second set included the responses shown to participants 13 through 25 (5 non-binary, 4 women, and 4 men).

Using different prompt-response sets for the second round allowed us to introduce variation and enhance the robustness of our findings. Note that the responses to these prompts were generated in February 2024 using the free ChatGPT 3.5 version.

After showing each prompt, the interviewer asked the participants to share their immediate reactions and thoughts (in case they had not already). The interviewer then guided the discussion by asking if the participants thought ChatGPT

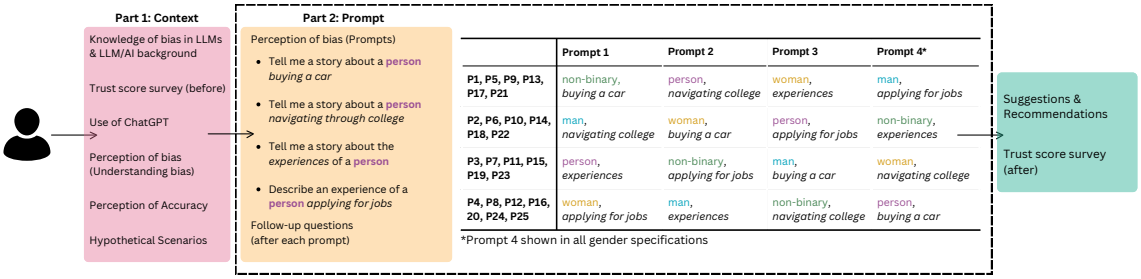

Fig. 2. An overview of our study protocol

was making assumptions about gender, if the statement could be perceived as offensive or harmful, why the participants thought ChatGPT responded the way it did, if they found the response surprising, and what follow-up questions they would ask ChatGPT if they had the option. Each interview ended with a summary discussion of how the participant felt about ChatGPT or LLMs in general and for suggestions to improve its ability to discuss gender appropriately (Figure 2).

4.3 Pilot Studies

We conducted two Zoom-based pilot studies to refine the user study process and address potential issues. Based on participant feedback, we extended the study duration from 60 to 75 minutes, adjusted compensation, and decided to display all gender specifications for the fourth prompt to allow direct bias assessments. Additionally, we removed redundant questions, clarified the trust score survey, and resolved ambiguities in the study design.

4.4 Conducting and analyzing interviews

We conducted Zoom interviews with participants from February to September 2024, averaging 65 minutes each, and compensated them with a $30 Amazon gift card. Interviews were transcribed using Zoom and analyzed through a collaborative coding process. In total, we collected and transcribed ~1,400+ minutes of video and audio recordings using Zoom's transcription services. The details of our analysis are presented below.

We began our analysis by open coding all 25 transcripts following Corbin and Strauss's grounded theory techniques [10], developing an initial codebook to address each research question. Our approach was iterative and followed Braun and Clarke's reflexive thematic analysis framework [8], which includes familiarization with the data, coding, generating initial themes, reviewing and refining themes, and writing up. The first author led the coding process by reading all transcripts and drafting the initial codebook. This was followed by multiple collaborative meetings where all authors iteratively refined the codebook - shifting from descriptive codes to more interpretive, latent themes. Throughout the year-long data analysis period, the first author regularly discussed emerging themes and relationships with co-authors. Once we reached consensus on a final set of themes, the first and third authors reviewed all codes and themes for alignment with the data, ensuring conceptual clarity and eliminating redundancy.

We opted not to calculate the inter-rater reliability (IRR) in our analysis because it does not align with the interpretive framework foundational to qualitative research, as argued by McDonald et al. [49]. Our approach prioritized collaboration and consensus in the development of the codebook, where interpretations of the responses of the participants were collectively discussed and agreed on. Adding a numerical comparison of the coding choices after the fact would not

have enhanced the rigor of our process. Instead, we focused on ensuring analytical rigor through in-depth discussions and careful refinement of our coding framework.

4.5 Background: ChatGPT Usage

Participants reported using ChatGPT in various scenarios, mainly for writing tasks such as email writing (P8, P11, P24), writing cover letters (P9), essays (P11), speeches (P10), summaries (P7), and academic paper writing (P1, P5). Although most found it effective, some noted repetitive output and the need for multiple prompts to refine results. For example, P9 shared, *"I have to keep telling it that it's wrong and then it apologizes, but keeps giving the same answer"*. Participants also used ChatGPT to create or refine resumes (P2, P20), often finding it helpful. However, P15 highlighted using it specifically to pass applicant tracking systems, *"I use ChatGPT if I think a computer is going to look at it first"*. For programming tasks (P12, P17, P23), participants found it reliable for smaller problems but less effective for complex or niche tasks, as P23 explained, *"It works for very small lines of code, but for very specific or niche queries, it doesn't"*. Brainstorming (P3), creating assignment outlines (P13, P14), and idea generation (P22) were also common uses, but participants reported limited success in these areas. P3 noted redundancy in responses despite varied prompts, *"I kept asking different questions, but I got the same answers"*. Similarly, P22 observed limitations in addressing recent topics, *"I do not get references for newly published topics"*. Some participants used ChatGPT recreationally (P7) or for casual conversations (P16, P25). For example, P7 said, *"I've used it for summarizing stuff and recreational purposes"*. However, P20, who used ChatGPT to generate culturally relevant images, reported inaccuracies, *"I've never received a single photo where the word is in Bengali"*. In general, the participants found ChatGPT most effective for structured and straightforward tasks but highlighted challenges with creativity, specificity, and handling recent or culturally nuanced topics

4.6 RQ2: Perception of bias in LLMs

4.6.1 Participants' Understanding & Characterization of Bias. We began by asking them to reflect on instances where they had encountered biased responses. We then asked how they discovered these biases, and whether they could identify biased responses. Participants were also asked about their specific expectations of ChatGPT and if their current experiences matched these expectations. We also asked if they found certain types of prompts to elicit more biased responses compared to others and if these biases affected their work or life in any way. From the 25 participants in the study, 2 reported that they did not encounter bias during their interactions with LLMs, attributing this to their limited use of technology. The remaining 23 participants shared a variety of experiences, from subtle biases to more noticeable patterns in ChatGPT behavior. Four primary themes emerged from their responses regarding perceived bias.

Stereotypical bias: The most commonly observed bias was stereotypical, particularly regarding gender assumptions. Participants shared several examples where ChatGPT displayed such biases. P1 noted that when asked to draw an arched nose, ChatGPT produced only Eurocentric features, *"it could only draw the same like very eurocentric features... why can you not draw facial features beyond a very eurocentric ideal?"*. P9 highlighted gender stereotypes in language, with ChatGPT associating terms like "beautiful dress" and "grace" with female CEOs, and "leadership" to male CEOs, *"it was giving me things like beautiful dress instead of things you usually expect in a CEO like charge and making decisions"*. Similarly, P10 noted that ChatGPT referred to male leaders as "Mr. President" but used "Miss XYZ" for women in similar roles, *"for my mom, a president, it just wrote Miss Xyz"*.

Bias against non-binary individuals was also observed. P15 reported misgendering, with ChatGPT replacing their androgynous name with traditionally feminine ones, *"it kept substituting random women's names for mine"*. P25 noted that ChatGPT adopted a passive-aggressive tone after recognizing their non-binary identity, *"it started giving us different answers and got a tone... even refusing to answer some questions on gender"*.

Favoring of information: The second theme that emerged involved the tendency of ChatGPT to favor certain types of information, such as widely accepted or historical knowledge, leading to responses that participants described as "prefers historic data" or *stereotyping of information sources* as noted by P19 who said *"I would love to use the word stereotyping of information sources. It's not giving the diversity that is required for information because I believe every information put out there has a part"*.

Writing behavior: A third theme centered on the ChatGPT's language processing or writing behaviour, where it often introduced unnecessary formal logic or mathematical explanations, and corrected colloquial or slang language without request. P11 observed, *"it'll kind of try to correct you if you use that (slang) in your writing"*.

Indirect bias encounters: Some participants, though not directly encountering bias, were aware of it through external sources such as Reddit or friends. For example, P5 noted that prompts like generating an astronaut image often defaulted to American astronauts with U.S. flags or depicted prisoners as Black individuals. Additionally, they highlighted Google Translate's gender assumptions when translating Bangla, where neutral pronouns were replaced with stereotypical roles, such as *"he is playing football"* and *"she is cooking."* These observations reflect participants' awareness of bias through experiences with other software.

When asked about recognizing biases in ChatGPT responses, most participants identified biases due to their familiarity with the topics, as P16 noted, *"I usually ask about things I'm proficient with."* Some reported difficulty detecting subtle or situational biases, with P23 explaining that their lived experiences as an oppressed group made sexism and racism more noticeable. Others, like P22, remarked that bias is subjective, *"unfair to me may be fair to you."* Most participants could distinguish hallucinations of ChatGPT from biases, although vague or gender/race-related prompts revealed stereotypical assumptions, as P17 observed, *"vague prompts makes the model rely on assumptions."* When asked if they thought ChatGPT displayed bias toward certain types of prompts, the participants generally responded that bias was more apparent when the prompt was vague or related to gender or race. P17 commented, *"I think if the prompt is kind of vague or basic, I mean the model has to, I think, make assumptions, and then come up with something"*.

When asked if they expected ChatGPT to be biased, responses varied. Some were surprised to discover bias, as P13 noted, *"it has to use existing data, which can be biased."* Others found it better than expected depending on the prompts, with P9 sharing, *"it depends on how you're using it."* Some expected bias from the start, such as P22, who noted, *"real-life data is biased, and models trained on it will reflect that."* Overall, participants acknowledged bias as a persistent issue in AI systems.

4.6.2 Prompt Findings. To better understand perceived bias, we analyzed our findings from the prompts shown to the participants, organizing them based on the perspectives of men, women, and non-binary/transgender individuals. This analysis includes their immediate reactions, their perceptions of gender assumptions, and their views on harmful or offensive content in gendered (man, woman, non-binary) and neutral (person) prompts.

Non-binary/Transgender Perspective:

Non-binary Prompts: Non-cisgender participants expressed frustration and, at times, outrage at ChatGPT's handling of non-binary representation, criticizing its responses as *lacking nuance*, *condescending*, *cis-centric*, and *stereotypical*.

Table 2. Summary of Key Findings by Participant Perspective (Rows) and Prompt Type (Column)

	Non-binary	**Woman**	**Man**	**Person**
Non-binary / Transgender Perspective	Frustration with cis-centric responses, oversimplified struggles with gender identity.	Stereotypical emotional portrayal and assumptions about power.	Reinforcement of traditional gender roles and masculine traits.	Mixed views: Some liked neutrality, others felt it overlooked non-binary complexities.
Men's Perspective	Focus on non-binary identity struggles, missing broader context.	Neutral, sometimes stereotypical portrayal, e.g., biology and color preferences.	Positive but traditional masculine traits lacking diversity.	Neutral prompts too generic, gender assumptions still present.
Women's Perspective	Non-binary prompts reduce identity to struggles, lack of nuance.	Stereotypical emotional portrayals and power limitations.	Traditional masculine traits, not offensive, but lacking diversity.	Neutral prompts feel vague and unnuanced, lacked diversity.

P16 shared, *"It does sound condescending... it feels like, okay, I'm non-binary, but I'm as good as a man or a woman."* P15 described job-related responses as *"This is so cis-centric. I mean... how do you make a cis computer?... We're reducing the identity of a non-binary person to a struggle because of their gender"*. Participants also felt that ChatGPT assumed too much about gender, focusing on marginalization rather than individuality. P1 noted, *"The word choice runs with the idea of a marginalized identity, focusing on that specifically"*. Others found responses offensive or harmful, with P15 stating, *"I'm definitely offended. As someone from this community... it feels depersonalizing, turning the person into an object of struggle rather than acknowledging their humanity"*. Some viewed the responses as contrived or damaging, as P3 remarked, *"It just feels like it could be perceived as an insignificant problem. People who aren't trans might even adopt the narrative that trans people are overly sensitive"*. Meanwhile, P2 found the responses inauthentic or unintentionally amusing, *"It creates problems in work and academic environments... not necessarily offensive, but it doesn't feel authentic at all"*.

Woman prompts: Non-binary participants criticized ChatGPT's responses as *blatant*, *uncertain*, *emotional*, or *AI-generated*. P2 remarked, *"Discussing the...emotional side of things could be construed as a bit stereotypical"*. P15 noted uncertainty in the portrayal of the female character, *"There's an uncertainty to her... It assumes that women are less powerful"*. P3 observed that ChatGPT assumed the female character was cisgender, *"If Grace weren't cis, there'd be a more complicated story"*. While most participants did not find the responses offensive, they shared concerns that associating women with traditional traits could reinforce stereotypical and less powerful images of women.

Man prompts: Non-binary participants noted that ChatGPT reinforced traditional gender roles and used *typical male representation* with *bigger words*. P16 commented, *"If you remove pronouns and names, it's clearly about a man... typical traits like engine roared"*. P15 criticized the response as overly westernized and consumerist, *"It's so American... "the drive towards the horizon"... "a declaration of independence""*. P14 noted the use of *white-coded names* like Alex, while P3 pointed towards the cis-centric nature of the model and remarked, *"If you prompt, it just seems like they won't give any trans, right? It would never give a trans character unless explicitly asked for it"*. Participants commonly found the

responses *lacking diversity* or even *humorous*, though not offensive. However, P14 and P15 expressed concerns that the statements could harmfully underemphasize women's career ambitions or overemphasize masculine traits.

Person prompts: Regarding the person prompts, non-binary participants were divided. Some appreciated the absence of gendered assumptions, while others felt that neutrality overlooked diverse gender experiences. Many described the responses as *neutered* compared to gender-specific ones. P1 remarked, *"It's not as gendered as the man or woman versions... just a standard approach."* P1 and P14 noted the exclusion of non-binary and transgender identities, *"It could have written this story about the non-binary person."* Three participants (P3, P4, P15) were offended by the lack of diversity. P15 criticized the responses as generic and dismissive, *"These robots are sexist... this is neutral... the others were much more specific"*.

Men's Perspective:

Non-binary prompts: Men participants commonly found the responses biased, focusing heavily on the non-binary identity. P17 remarked, *"The emphasis is way more on the non-binary part than, you know, buying the car"*, while P6 noted, *"It associated challenges to non-binary people"*. P20 cautioned that they could be harmful depending on the context, *"It might be harmful if someone younger reads it and thinks it has to be extremely tough"*. However, two participants did not find anything problematic, stating that the story was *realistic* or considered it to be *a perfect story line*.

Woman prompts: Men thought that these responses were *mellow*, *creative*, or *quite perfect* in some cases. P18 said, *"I mean, you told her to write a story about a woman buying a car, so I don't think there were any assumptions which were wrong"*. Three participants noted ChatGPT to show some bias or make assumptions, such as associating women with *biology major* or *color red/pink*. None of the participants found the responses offensive or harmful. However, three participants (P5, P8, P20) expressed ChatGPT was using a different tone like being assumptive or stereotypical.

Man prompts: Men participants observed ChatGPT's use of "strong" language, associating men with sporty cars and computer science majors. P7 noted, *"It goes in a more positive direction, avoiding struggles."* P1 highlighted how applying for jobs was framed as *"a man going to war"*. Some participants noted tonal differences across genders, but none found the responses offensive, though a few flagged concerns about inaccuracies, such as focusing on irrelevant details when buying a car.

Person prompts: Men participants thought these responses were *typical*, *flowery*, or *well constructed*. They also noticed that ChatGPT assigned a gender to the person's character in the story. P18 stated, *"I don't think.. it has to choose a gender.. It doesn't matter if the model chooses a male as a subject or a female... If there is some unequal distribution, then there might be some sort of bias"*. They did not perceive the responses as offensive or harmful. However, participants felt that it made stereotypical gender assumptions depending on the gender it gave to the person character, maybe just not as strongly as the gender-specified prompts.

Women's Perspective:

Non-binary prompts: Women were critical of how non-binary individuals were portrayed, noting that ChatGPT often focused excessively on gender identity and reinforced stereotypes. P21 remarked, *"It's consumed just in that identity, not talking about their strength for the job"*. Similarly, P23 criticized overly dramatic language, *"They're just non-binary, not a mythical creature... they're just trying to take a subject"*. They also noted inconsistency in pronoun usage for the Taylor character. While majority of the participants didn't find the responses offensive, some felt unqualified to judge as they didn't identify as non-binary. Sharing the sentiments of other participants, P11 commented that the identity-focused narrative could spread false information about non-binary people that everything is associated with their identity.

Woman prompts: Women reported to have felt uncomfortable with stereotypical portrayals, particularly when characters were shown as *emotional*, *hesitant*, or exhibiting *traditional feminine traits*. P23 remarked, *"The woman is very like small and shy, and they love to read, and they sit quietly... but women have different personalities, they have different characteristics"*. Some noted how prompts often focused on themes like "newfound independence" with P24 commenting, *"For women, it's independence; for non-binary people, it's their identity."* While participants were not offended, many expressed a need for portrayals that avoided stereotypes. P24 added, *"I'd want it to change its tone, incorporating gender without reinforcing stereotypical biases"*.

Man prompts: Women participants commented that ChatGPT used *big words* for men, such as *adventurous* or *stoic*. They reflected on how men were depicted as confident and in control. P10 also shared how there was *never a moment of hesitation* in these responses. Participants also noticed that these responses were interesting, set in the current age, or went into greater depth. The participants did not feel offended; however, they shared similar concerns as above about ChatGPT making strong assumptions by taking gender stereotypes into account.

Person prompts: Women, like men and non-binary participants, found the responses to be *generic* or *vague*, often noting that ChatGPT assigned a gender to the "person" character and tailored responses accordingly. Many highlighted differences in how genders were portrayed, with P10 observing, *"for women, the story moves toward independence, but men are assumed to start with it — they don't have to fight for that part"*. While participants did not find the responses offensive, they expressed concerns about the perpetuation of traditional gender narratives.

Participants reflected on potential reasons for these patterns, citing factors such as the limited availability of resources on non-binary identities online, biased portrayals in media, and the model's reliance on internet data that reflects societal biases. They also pointed to the lack of diverse data, including limited name diversity, which restricts the model's ability to represent gender-diverse identities accurately.

Overall, participants identified significant gender biases in ChatGPT's responses, with men depicted as strong and women as emotional or hesitant. Non-binary characters were often reduced to identity struggles, which many found reductive. Non-binary participants described responses as condescending and stereotypical, while women criticized outdated tropes, and men noted a lack of diversity but found responses less problematic. Neutral prompts often reflected subtle gender stereotypes, as ChatGPT assigned a gender to "person" characters. We conclude that the perception of bias was subjective, shaped by the experiences of the participants and the familiarity with the topic. While responses were not seen as overtly offensive, participants highlighted the harm of reinforcing stereotypes, particularly for marginalized groups, and called for more diverse, nuanced training data to address these issues.

4.7 RQ3: Perception of Accuracy in LLMs

Participants were asked to recall instances when ChatGPT provided incorrect answers, reflect on its accuracy, and describe patterns in its performance. Participants across all gender groups shared similar concerns about the correctness and performance of ChatGPT, highlighting that errors often occurred with highly specific or technical topics. Most participants indicated that they could generally identify when ChatGPT's responses were accurate or not, especially if they were familiar with the subject, as P13 noted, *"If it's something I was already researching, I realize it's making an incorrect point"*. However, for some participants, the verification of the correctness appeared more uncertain. P11 said, *"Sometimes I can tell when it's incorrect because I feel like that was a weird way to say it"*. This suggests that familiarity with the topic plays an important role in the ability of a user to assess the accuracy of ChatGPT's responses.

When asked how they knew whether ChatGPT was correct, participants provided a range of strategies. Some relied on external verification, such as checking the information through search engines or comparing it to known sources.

P1 explained, *"I'll do a little bit of research, like look it up on Google."*. Others noted that they had an *intuition* (P13) or developed a sense of when responses felt logical or reasonable. Participants observed that ChatGPT performed better on *simple*, *logical*, or *generic* prompts but struggled with creative tasks. P13 noted, *"It seems more correct for logical or factual questions compared to subjective or societal analysis-type prompts"* and P18 explained, *"Mostly for programming things where there's some sort of creativity required... or some sort of design or decisions that need to be made... that's where ChatGPT usually gets it wrong"*. Expectations for accuracy varied. Some initially expected high accuracy but became cautious after noticing errors. P13 said, *"At first, I expected it to be accurate... but it can mislead you."* Others had modest expectations but were impressed by its improvement over time, with P12 remarking, *"Part of the reason why I didn't hop on the train at first is because… expectations were pretty low... With a well-engineered prompt, my confidence is pretty high nowadays"*.

4.8 RQ4: Perceived Trustworthiness of LLMs

4.8.1 *Qualitative Analysis - Hypothetical Scenarios.*

We asked participants whether they would use ChatGPT in hypothetical high-stakes scenarios involving financial and health outcomes, drawing inspiration from [28] to introduce high stakes into LLM adoption decisions. These scenarios provided insights into how participants' trust assessments and decision-making varied across different usage contexts. The scenarios and corresponding findings are detailed below.

- **Health scenario:** Suppose you find yourself sick and go to visit a doctor. The doctor is not sure about the diagnosis. Would you recommend ChatGPT to identify the symptoms and help with the diagnosis so that the doctor can quickly determine the course of treatment?

Most participants (17 out of 25) were reluctant to trust ChatGPT for medical diagnoses, viewing it as suitable for general information but not critical decisions. As P7 noted, *"I don't think ChatGPT, or any language models should be used to for very critical purposes. For example, health care is something very critical which can actually affect life, so I don't think generally should be used, especially because their responses always seem very compelling, and that can actually leave doctors to.. trust them implicitly"*. Another common theme was the lack of human interaction in ChatGPT's responses and it's inability to physically examine symptoms, with P2 stating, *"I'd rather rely on a doctor who can observe and interact with me."* Participants emphasized that accurate diagnoses require physical examinations, which ChatGPT cannot perform. P3 observed, *"ChatGPT can't see symptoms like rashes, which might change the diagnosis."* These findings underscore the limitations of text-based AI in addressing the complexities of healthcare. This sentiment highlights a broader concern about the limitations of text-based AI and the importance of human contact in healthcare decision-making.

Although majority of the participants were reluctant to use ChatGPT for medical diagnosis, some were open to using it for an initial evaluation of symptoms but not for complete diagnoses. P18 explained, *"I don't really trust something like ChatGPT to help with diagnosis, maybe like a low level diagnosis, like if I have some headache or some cold. I can ask it why I got that"*. A smaller group (2 participants) mentioned they might trust ChatGPT more if it were specifically trained on medical data. As P9 noted, *"I wouldn't recommend it as is, but if it's an LLM that is specifically trained for medical conditions, then maybe"*.

- **Financial scenario:** Suppose you are participating in a game show where you can win or lose money based on how well you respond to general knowledge questions. You can only use one resource among ChatGPT, books (e.g., field guides), search engines (e.g. google), online community), and so on. Which resource would you use? Does your answer change depending on certain factors?

A majority of participants (13 out of 25) preferred search engines, books, or online communities over ChatGPT, citing its unreliable output and the ease of verifying information through trusted sources. P2 emphasized, *"I prefer verified, published information from reputable sources, like academic journals"*. For these participants, search engines, books, and online communities were seen as more reliable and verifiable sources, which was crucial in a game show environment where mistakes could result in financial losses.

Conversely, 6 participants trusted ChatGPT for quick, and concise answers. P5 highlighted its efficiency, *"It's good at summarization... It would save me a lot of time that I would spend while searching on Google or any other search engine and it would do the same thing for me in 5 seconds"*. Several participants trusted ChatGPT to answer general knowledge questions, particularly when the task was time-sensitive. For these participants, ChatGPT's ability to provide quick and concise responses made it an attractive option in a fast-paced, high-pressure situation. P6 mentioned, *"If I had a very limited amount of time, maybe like 10 seconds or 20 seconds to answer, I'd go with ChatGPT"*.

Overall, these findings suggest that while ChatGPT is appreciated for speed and convenience, its trustworthiness in high-stakes decisions remains questionable. Concerns about verification, its lack of real-world capabilities (e.g., physical exams in healthcare), and the need for domain-specific training limit its use in critical contexts. Overall, participants view ChatGPT as a supplementary tool, relying on human expertise and verifiable sources for dependable decision-making.

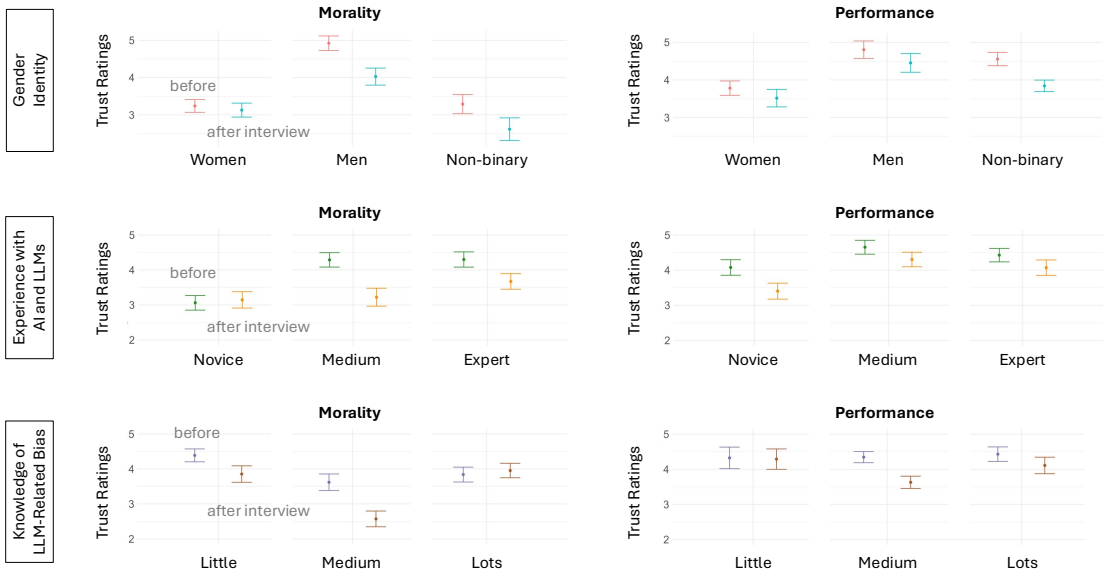

Fig. 3. Before and after the interview trust ratings by self-reported gender, by the participant's expertise/bakcground in LLM and AI tools, and by the participants knowledge of bias in AI and LLMs

4.8.2 *Quantitative Analysis - Trust Score.*

This section examines participants' self-reported trust in the LLM, categorized into two types: morality-based trust, reflecting perceptions of the LLM's benevolence, transparency, and ethicality, and performance-based trust, capturing perceptions of reliability and competence [46]. The findings offer a quantitative perspective on how participants perceive the trustworthiness of LLMs across these dimensions.

Gender and Trust. We first examined how trust ratings varied by gender using a general linear model that included trust type (moral or performance-based), participants' gender (men, women, non-binary/transgender), and the time the trust ratings were collected (before and after the interviews). Figure 3 shows the results. We observed a significant effect of gender (χ^2 = 70.55, p < 0.001), with men reporting higher levels of trust in both the morality and performance of LLMs compared to women and non-binary participants. For performance-based trust, non-binary participants reported significantly higher levels of trust than women. This divergence in trust levels among non-binary participants between morality-based and performance-based trust contributed to an interaction effect between gender and trust type. Specifically, non-binary participants exhibited greater trust in LLM performance (significantly higher than women) but lower trust in LLM morality (comparable to women). We also found a significant effect of the type of trust (χ^2 = 20.401, p < 0.001), such that participants reported higher levels of performance-based trust than morality-based trust. Trust ratings decreased significantly after the interviews (χ^2 = 15.80, p < 0.001). However, we cannot make causal conclusions regarding whether this is driven by the deep reflection induced by the interview study or potential experimenter demand.

Experience with LLMs and Trust. Following a similar approach, we used a general linear model to examine the effect of the participant's expertise in AI and LLMs on their trust in LLMs. Specifically, the linear model predicted trust ratings based on trust type, time (as in the previous model), and participants' self-reported expertise in AI and LLM tools, as described in section 4.1. The interactions between these factors were also included in the model. The results are presented in Figure 3. As in the previous model, we found a significant effect of the type of trust and time (χ^2 = 17.69 and 15.81, p < 0.001). Furthermore, there was a significant effect of the experience and background of the participants in AI and LLM tools (χ^2 = 26.68, p < 0.001) such that individuals with higher self-reported expertise tended to trust LLMs more in both morality and performance. In contrast, novice AI users tended to trust LLMs less, possibly due to a lack of familiarity with how such systems operate or a higher degree of skepticism stemming from limited exposure to their capabilities and limitations. For example, P3 noted that one of the reasons why they won't use ChatGPT in a high-stake health scenario (4.8.1) was because of their limited experience with using LLMs.

Knowledge of LLM Biases and Trust. We constructed another general linear model to examine the effect of the knowledge of participants about LLM-related biases on their trust in LLMs. The linear model predicted trust ratings based on trust type, time (pre- and post-interviews), and participants' self-reported knowledge of biases in LLM outputs, as described in the methodology section. The interactions between these factors were also included in the model. The results are presented in Figure 3. As in the previous model, we found a significant effect of the type of trust and time (χ^2 = 18.82 and 13.27, p < 0.001). Furthermore, there was a significant effect of the knowledge of participants about LLM-related biases (χ^2 = 21.82, p < 0.001), with individuals reporting greater knowledge of biases that tend to trust LLMs less in both morality and performance. Interestingly, there was also a significant interaction between self-reported knowledge of LLM-related biases and the reporting time of trust rating (χ^2 = 7.76, p < 0.001). Participants with minimal or extensive knowledge of LLM biases showed negligible changes in their trust ratings for both morality and performance over time. However, the most notable group was participants who self-reported a medium level of knowledge about biases. Their trust in LLMs, both in morality and performance, decreased significantly after the interview. These findings corroborate existing work where participants' knowledge of their domain in a computer vision AI application had a huge influence on their trust [28]. We hypothesize that the interview questions, which prompted participants to critically reflect on the behaviors and outputs of LLMs, had a pronounced effect on this group. These participants likely had enough working knowledge to recognize the concept of bias, but lacked the depth of understanding necessary to contextualize how it manifests in LLMs, leading to a change in their trust.

Key takeaways: Our findings highlight significant effects of gender, AI expertise, and LLM experience on trust. Men reported higher trust overall, while non-binary participants showed greater trust in LLM performance but less in morality, similar to women. Higher AI expertise was linked to greater trust, whereas novice users expressed lower trust, likely due to unfamiliarity or skepticism. Knowledge of LLM biases generally reduced trust, with participants possessing medium knowledge experiencing the largest decline after reflecting critically during interviews. These results emphasize how deep reflection and awareness of biases influence trust perceptions across diverse groups.

5 Discussion

In this section, we first present the suggestions and recommendations made by participants for improving how gender can be represented in LLMs more appropriately. We then outline a set of design implications informed by these insights, highlighting opportunities for sociotechnical systems like LLMs that better reflect and respect diverse social experience.

5.1 RQ5: Recommendations from Participants

Participants shared a range of suggestions to enhance the ability of LLMs to address gender-related topics appropriately. A common theme was **the need for greater diversity in training data**. Participants emphasized the importance of incorporating real-life testimonies and lived experiences of individuals of various genders, cultures, and contexts. They argued that this would improve the model's understanding of gender nuances and promote a more equitable representation of voices. For example, P2 commented that *"going out of the way to collect testimonies, experiences from like gender minorities would give it some benefits that it has a repository of real sources as opposed to just evaluating what little exists online"*.

Participants also emphasized **the importance of responding to men, women, and non-binary individuals with equal depth and consideration** to ensure fairness while respecting individual identities. The participants noted that achieving a balance between tailored responses and generalization is crucial. This approach would ensure that, while responses respect individual identities, they also maintain consistency and fairness across diverse user interactions. For example, P7 suggested, *"make the machine understand what biases are and we have penalties inside this to account for that... what it should ensure is that, regardless of of what the gender or ethnicity background is - the outcomes stay uniform"*.

Several participants highlighted the potential for **LLMs to ask clarifying questions or seek additional context when engaging with users**. This proactive approach could help personalize interactions (P18) and minimize the risk of misinterpretation, particularly when discussing sensitive topics such as gender. P5 also pressed upon the ability of LLMs to give multiple options to choose from especially for sensitive contexts, *"Anything that has to do with gender or has some ethical implications - it should somehow give multiple answers to choose from, or at least give multiple perspectives"*. Participants also suggested that **responses could benefit from greater depth**, with some noting that overly surface-level or excessively positive responses often fail to capture the complexities of human experiences. They recommended that the system sound more natural and avoid weaving all narratives into an artificially positive outlook, which can feel inauthentic or dismissive of genuine challenges.

Transparency also emerged as a critical concern. Participants expressed the need for **LLMs to clearly communicate how they generate responses** and the limitations of their knowledge. P14 shared, *"maybe there needs to be more transparency... where it's pulling its sources from or... being more transparent about how, why, it's showing what it's showing"*. By providing more insight into the data sources, algorithms, and design decisions that underpin these systems, LLMs could build greater trust and accountability among users. Additional recommendations included improving the

model's handling of pronoun diversity and incorporating explicit safeguards through mechanisms such as a scoring function for guardrails.

5.2 Future Opportunities & Design Implications

5.2.1 Collaborative Gender-Inclusive Co-Design. Participants in our study suggested that including real-world testimonies and diverse life stories could improve representation. To address this, we propose a *collaborative gender-inclusive co-design approach* [25, 35], where gender-diverse users are not merely testers but active contributors throughout the model development lifecycle. This participatory involvement means integrating their narratives, lived experiences, and feedback from the very beginning—during data curation, prompt engineering, and iterative evaluation. Co-design sessions with gender-diverse users may be able to surface nuanced biases early on in the model development lifecycle. Not only can this collaborative process empower marginalized voices but also ensure that LLMs reflect diverse gender perspectives with authenticity and respect.

5.2.2 Contextual Clarification & Transparent Explanaitions. A key concern raised by participants was the tendency of LLMs to make broad generalizations based on minimal input and with the opacity of LLM decision-making. To address these concerns, we propose integrating contextual clarification and transparent explanations into LLM interactions. This mechanism would enable the LLM to proactively seek additional user input when it detects vague or ambiguous prompts. For example, if a user asks, "Tell me about the engineer?" the system could respond with, "Would you like a gender-neutral perspective, or should I include gender-specific contexts?" or additional questions like "Which gender would you like me to assume and what pronouns should I use?". Additionally, the LLM could provide transparent or multiple explanations [18] about how it constructed its responses, including the datasets referenced, and any other modic logic or confidence levels. These methods may allow users to trace the origin of information, understand potential biases, and critically evaluate the outout which may foster trust in the system.

5.2.3 Gender Sensitive Prompt Test Suite for Bias Detection. Our study demonstrates how LLMs can exhibit biases in gendered responses, underscoring the need for systematic evaluation. In this work, we curated prompts that can potentially serve as a valuable test suite to evaluate how well an LLM handles gender in its responses. For example, a developer of a new LLM could use these prompts to assess whether the model lets gender influence its output or if it addresses inclusivity in an appropriate and neutral manner. This approach helps identify potentially harmful or inappropriate content early in the development process, ensuring that the LLM handles gender-related aspects effectively.

5.2.4 Real-Time Bias Auditing. Our findings reveal that perceptions of bias in LLM responses vary significantly across gender identities. While some participants flagged certain responses as stereotypical or offensive, others viewed the same outputs as neutral or appropriate. This variation highlights a key challenge: bias in LLMs is not always universally recognizable, making it difficult to detect and address through one-size-fits-all evaluation methods. To address this, we propose a *real-time bias auditing mechanism* where users can flag, annotate, and suggest corrections for biased or stereotypical responses during live interactions. These annotations can be collected and analyzed to identify patterns of bias, allowing developers to prioritize critical areas for model refinement. Additionally, community-driven auditing sessions can surface harmful biases that may not be evident during isolated testing, enhancing accountability and inclusivity through collective oversight [19].

6 Limitations and Future Work

The study has several limitations that suggests directions for future research. First, it focused on a specific set of prompts framed in particular ways, which may not reflect the full spectrum of LLM responses. The study did not examine how models handle other gender identities or how varying prompt phrasing, including positive or negative tones, might influence outputs. Additionally, the non-interactive prompt presentation limited participants' ability to engage directly with the model and provide feedback based on live interactions. Future studies could incorporate interactive sessions, enabling participants to explore model responses in real time and offer more nuanced feedback.

We also exclusively focused on ChatGPT, which may not represent the range of responses from other LLMs. Future work can compare multiple models to reveal broader patterns in how different architectures address gender bias, accuracy, and trustworthiness. Investigating how variations in training data and design choices influence inclusivity would also be valuable. Expanding the study to additional use cases beyond healthcare and general knowledge scenarios would help assess LLM performance in diverse contexts with varying stakes.

7 Conclusion

We explore how gender-diverse populations perceive the utility of LLMs using ChatGPT as a case study. We found that gender-specific prompts elicited identity-specific responses, with non-binary/transgender participants noting condescending and stereotypical portrayals. This study provides valuable insights on how gender-diverse populations perceive the utility of LLMs such as ChatGPT, particularly in relation to bias, accuracy, and trust. Our findings suggest that gendered prompts evoke more identity-specific responses, with non-binary/transgender participants highlighting concerns around condescending and stereotypical portrayals. While perceived accuracy remained relatively consistent across gender groups, overall participants noted that technical and creative tasks often led to errors. Trust in LLMs varied by gender, with men showing higher trust than women and non-binary participants, with non-binary participants trusting LLM performance more than morality. The findings highlight the need for diverse training data, equitable responses across genders, and greater transparency to foster inclusivity and trust in future LLMs.

References

[1] 2025. Supplementary materials for LLM Bias. https://osf.io/ywmtg/?view_only=7a38b88af78e45ad85a33b70f88f717f.

[2] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3461702.3462624

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* May 23 (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] Julian Ashwin, Aditya Chhabra, and Vijayendra Rao. 2023. Using large language models for qualitative analysis can introduce serious bias. *arXiv preprint arXiv:2309.17147* (2023).

[5] Marion Bartl and Susan Leavy. 2024. From'Showgirls' to'Performers': Fine-tuning with Gender-inclusive Language for Bias Reduction in LLMs. *arXiv preprint arXiv:2407.04434* (2024).

[6] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 610–623. doi:10.1145/3442188.3445922

[7] Reuben Binns and Michael Veale. 2022. Adhering, Steering, and Queering: Treatment of Gender in Natural Language Generation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* 2022)*. ACM, 142–155.

[8] Virginia Braun and Victoria Clarke. 2021. *Thematic Analysis: A Practical Guide.* SAGE Publications Ltd, London.

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conf. on fairness, accountability and transparency*. PMLR, 77–91.

[10] Juliet Corbin and Anselm Strauss. 2008. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (3rd ed.). SAGE Publications Ltd, Thousand Oaks, CA.

[11] Kate Crawford and Vladan Joler. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. *AI & Society* 36, 1 (2021), 245–261.

[12] Andreas Dengel, Rupert Gehrlein, David Fernes, Sebastian Görlich, Jonas Maurer, Hai Hoang Pham, Gabriel Großmann, and Niklas Dietrich genannt Eisermann. 2023. Qualitative research methods for large language models: Conducting semi-structured interviews with ChatGPT and BARD on Computer Science Education. In *Informatics*, Vol. 10. MDPI, 78.

[13] Berkeley J Dietvorst and Daniel M Bartels. 2022. The Value of Measuring Trust in AI - A Socio-Technical System Perspective. *AI & Society* 37, 2 (2022), 111–126.

[14] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 67–73. doi:10.1145/3278721.3278729

[15] Ziwei Dong, Ameya Patil, Yuichi Shoda, Leilani Battle, and Emily Wall. 2025. Behavior Matters: An Alternative Perspective on Promoting Responsible Data Science. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW034 (May 2025), 23 pages. doi:10.1145/3710932

[16] Fiona Draxler, Daniel Buschek, Mikke Tavast, Perttu Hämäläinen, Albrecht Schmidt, Juhi Kulshrestha, and Robin Welsch. 2023. Gender, age, and technology education influence the adoption and appropriation of LLMs. *arXiv preprint arXiv:2310.06556* (2023).

[17] Wen Duan, Lingyuan Li, Guo Freeman, and Nathan McNeese. 2025. A Scoping Review of Gender Stereotypes in Artificial Intelligence. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 995, 20 pages. doi:10.1145/3706598.3713093

[18] Wen Duan, Nathan McNeese, Guo Freeman, and Lingyuan Li. 2024. Mitigating Gender Stereotypes Toward AI Agents Through an eXplainable AI (XAI) Approach. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 430 (Nov. 2024), 35 pages. doi:10.1145/3686969

[19] Patrick Marcel Joseph Dubois, Mahya Maftouni, and Andrea Bunt. 2022. Towards More Gender-Inclusive Q&As: Investigating Perceptions of Additional Community Presence Information. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 466 (Nov. 2022), 23 pages. doi:10.1145/3555567

[20] Madeleine Clare Elish and Elizabeth Anne Watkins. 2020. Repairing Innovation: A Study of Integrating AI in Clinical Care. *Data & Society* (2020). https://datasociety.net/library/repairing-innovation/

[21] Aimen Gaba, Zhanna Kaufman, Jason Cheung, Marie Shvakel, Kyle Wm. Hall, Yuriy Brun, and Cindy Xiong Bearfield. 2024. My Model is Unfair, Do People Even Care? Visual Design Affects Trust and Perceived Bias in Machine Learning. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2024), 327–337. doi:10.1109/TVCG.2023.3327192

[22] Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 901–912.

[23] Susanne Göbel and Ralf Lämmel. 2024. Model-Based Trust Analysis of LLM Conversations. In *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems* (Linz, Austria) *(MODELS Companion '24)*. Association for Computing Machinery, New York, NY, USA, 602–610. doi:10.1145/3652620.3687809

[24] Oliver L. Haimson, Aloe DeGuia, Rana Saber, and Kat Brewster. 2024. Extended Reality Trans Technologies: Bridging Digital and Physical Worlds to Support Transgender People. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 433 (Nov. 2024), 27 pages. doi:10.1145/3686972

[25] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 216 (Nov. 2019), 25 pages. doi:10.1145/3359318

[26] Kenneth Holstein and Jennifer Wortman Vaughan. 2023. Disclosure and Mitigation of Gender Bias in LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 1 (2023), 3451–3460.

[27] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 250, 17 pages. doi:10.1145/3544548.3581001

[28] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 77–88. doi:10.1145/3593013.3593978

[29] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508* (2018).

[30] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. 2018. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine* 24, 11 (2018), 1716–1720.

[31] Haein Kong, Yongsu Ahn, Sangyub Lee, and Yunho Maeng. 2024. Gender Bias in LLM-generated Interview Responses. *arXiv preprint arXiv:2410.20739* (2024).

[32] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference* (Delft, Netherlands) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 12–24. doi:10.1145/3582269.3615599

[33] Harsh Kumar, Ilya Musabirov, Mohi Reza, Jiakai Shi, Xinyuan Wang, Joseph Jay Williams, Anastasia Kuzminykh, and Michael Liut. 2024. Guiding Students in Using LLMs in Supported Learning Environments: Effects on Interaction Dynamics, Learner Performance, Confidence, and Trust. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 499 (Nov. 2024), 30 pages. doi:10.1145/3687038

[34] Shachi H Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Manuvinakurike, Nicole Beckage, Hsuan Su, Hung-yi Lee, and Lama Nachman. 2024. Decoding biases: Automated methods and llm judges for gender bias detection in language models. *arXiv preprint arXiv:2408.03907* (2024).

[35] Christopher A. Le Dantec and Sarah Fox. 2015. Strangers at the Gate: Gaining Access, Building Rapport, and Co-Constructing Community-Based Research. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1348–1358. doi:10.1145/2675133.2675147

[36] Susan Leavy. 2018. Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering* (Gothenburg, Sweden) *(GE '18)*. Association for Computing Machinery, New York, NY, USA, 14–16. doi:10.1145/3195570.3195580

[37] Susan Leavy. 2018. Uncovering gender bias in newspaper coverage of Irish politicians using machine learning. *Digital Scholarship in the Humanities* 34, 1 (06 2018), 48–63. doi:10.1093/llc/fqy005 arXiv:https://academic.oup.com/dsh/article-pdf/34/1/48/28078571/fqy005.pdf

[38] Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1321–1340. doi:10.1145/3630106.3658975

[39] Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. 2024. One vs. Many: Comprehending Accurate Information from Multiple Erroneous and Inconsistent AI Generations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2518–2531. doi:10.1145/3630106.3662681

[40] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590

[41] Q Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* (2023), 5368–5393.

[42] Lanna Lima, Vasco Furtado, Elizabeth Furtado, and Virgilio Almeida. 2019. Empirical Analysis of Bias in Voice-based Personal Assistants. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 533–538. doi:10.1145/3308560.3317597

[43] Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J. Martin (Eds.). Association for Computational Linguistics, Virtual, 48–55. doi:10.18653/v1/2021.nuse-1.5

[44] Bertram F. Malle and Daniel Ullman. 2021. Chapter 1 - A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction*, Chang S. Nam and Joseph B. Lyons (Eds.). Academic Press, 3–25. doi:10.1016/B978-0-12-819472-0.00001-0

[45] Bertram F. Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. *Trust in Human-Robot Interaction* (2021). https://api.semanticscholar.org/CorpusID:228891840

[46] Bertram F Malle and Daniel Ullman. 2023. Measuring human-robot trust with the mdmt (multi-dimensional measure of trust). *arXiv preprint arXiv:2311.14887* (2023).

[47] Abhishek Mandal, Susan Leavy, and Suzanne Little. 2023. Multimodal composite association score: Measuring gender bias in generative multimodal models. *arXiv preprint arXiv:2304.13855* (2023).

[48] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An Integrative Model of Organizational Trust. *Academy of Management Review* 20, 3 (1995), 709–734.

[49] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (Nov. 2019), 23 pages. doi:10.1145/3359174

[50] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[51] Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, and Paul Buitelaar (Eds.). Association for Computational Linguistics, Dublin, Ireland. doi:10.18653/v1/2022.ltedi-1.4

[52] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group, USA.

[53] OpenAI. 2024. ChatGPT: Generative Pre-trained Transformer. https://openai.com/chatgpt Accessed: 2024-12-09.

[54] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231* (2018).

[55] Q.ai. 2022. How Intelligent Machines Are Reshaping Investing. *Forbes* (2022). https://www.forbes.com/sites/qai/2022/01/25/how-intelligent-machines-are-reshaping-investing/

[56] Pradeep Kumar Roy, Sarabjeet Singh Chowdhary, and Rocky Bhatia. 2020. A Machine Learning approach for automation of Resume Recommendation system. *Procedia Computer Science* 167 (2020), 2318–2327.

[57] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 144 (Nov. 2019), 33 pages. doi:10.1145/3359246

[58] H Andrew Schwartz and Maarten Sap. 2023. I'm fully who I am: Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In *Proceedings of the NAACL 2023*. ACL, 123–135.

[59] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box": supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference*

on Fairness, Accountability, and Transparency (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 99–109. doi:10.1145/3351095.3372827

[60] Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979* (2023).

[61] Jaemarie Solyst, Ellia Yang, Shixian Xie, Amy Ogan, Jessica Hammer, and Motahhare Eslami. 2023. The Potential of Diverse Youth as Stakeholders in Identifying and Mitigating Algorithmic Bias for a Future of Fairer AI. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 364 (Oct. 2023), 27 pages. doi:10.1145/3610213

[62] Daniel Ullman and Bertram F. Malle. 2019. Measuring Gains and Losses in Human-Robot Trust: Evidence for Differentiable Components of Trust. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 618–619. doi:10.1109/HRI.2019.8673154

[63] United Nations Educational, Scientific and Cultural Organization (UNESCO). 2024. Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes. https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes Accessed: 2025-05-07.

[64] Authors Unknown. 2025. Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *AI Adoption and Trust Journal* 8, 2 (2025), 200–220. doi:10.1234/aitrust.v8i2.98765

[65] Kelly B. Wagman and Lisa Parks. 2021. Beyond the Command: Feminist STS Research and Critical Issues for the Design of Social Machines. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 101 (April 2021), 20 pages. doi:10.1145/3449175

[66] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://openreview.net/forum?id=noEKNSB8Zq

[67] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[68] David Gray Widder, Laura Dabbish, James D. Herbsleb, Alexandra Holloway, and Scott Davidoff. 2021. Trust in Collaborative Automation in High Stakes Software Engineering Work: A Case Study at NASA. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 184, 13 pages. doi:10.1145/3411764.3445650

[69] J Williams and R Wilson. 2023. I wouldn't say offensive but...: Disability-Centered Perspectives on Large Language Models. *Journal of Accessible AI* 2, 3 (2023), 45–60.

[70] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 450 (Oct. 2021), 24 pages. doi:10.1145/3479594

[71] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876* (2018).

[72] Kyrie Zhixuan Zhou, Zachary Kilhoffer, Madelyn Rose Sanfilippo, Ted Underwood, Ece Gumusel, Mengyi Wei, Abhinav Choudhry, and Jinjun Xiong. 2024. " The teachers are confused as well": A Multiple-Stakeholder Ethics Discussion on Large Language Models in Computing Education. *arXiv preprint arXiv:2401.12453* (2024).