Gender Bias in LLM-generated Interview Responses

Haein Kong[*]
Rutgers University
haein.kong@rutgers.edu

Yongsu Ahn[*]
University of Pittsburgh
yongsu.ahn@pitt.edu

Sangyub Lee[*]
Korea University
Korean National Police University
yubii2@korea.ac.kr

Yunho Maeng[†]
Ewha Womans University
LLM Experimental Lab, MODULABS
yunhomaeng@ewha.ac.kr

Abstract

LLMs have emerged as a promising tool for assisting individuals in diverse text-generation tasks, including job-related texts. However, LLM-generated answers have been increasingly found to exhibit gender bias. This study evaluates three LLMs (GPT-3.5, GPT-4, Claude) to conduct a multifaceted audit of LLM-generated interview responses across models, question types, and jobs, and their alignment with two gender stereotypes. Our findings reveal that gender bias is consistent, and closely aligned with gender stereotypes and the dominance of jobs. Overall, this study contributes to the systematic examination of gender bias in LLM-generated interview responses, highlighting the need for a mindful approach to mitigate such biases in related applications.

1 Introduction

LLMs have increasingly demonstrated the ability to assist in effectively generating language for individuals' livelihood and critical events [6, 22]. One area of showing its potential lies in helping people better represent themselves for their career verbally or in written languages [23, 11]. For instance, job applicants are increasingly leveraging LLMs to generate personal statements and anticipated interview responses. Recent advances in LLM-powered applications, such as in-store tools and specialized writing capabilities for job-related essays and interview preparations[3], have further boosted the use of LLMs in job-related language generation tasks.

However, researchers have found risks of gender bias in LLM-generated job-related languages. For example, LLMs tend to produce different recommendation letters for female and male applicants [7, 20], often reinforcing traditional gender stereotypes. These align with the well-known dichotomy of female communal versus male agentic stereotypes [20, 14, 17], where men are perceived as more assertive and task-focused, while women are seen as more polite and person-oriented. Other studies show that LLMs exhibit demographic bias when assigned to a job/position matching or hiring decision by disfavoring certain demographics (e.g., Hispanic male) [1] or recommending stereotypical roles to job seekers (e.g., drivers to men, secretarial roles to women) [16].

In our study, we investigate how gender bias penetrates the task of LLM-assisted interview preparation, an underexplored but crucial application of language generation in job-related tasks. Our study conducts a multifaceted audit of LLM-generated interview responses across various dimensions,

---

[*]Equally contributed.

[†]Corresponding author.

[3]"Interview Prep with GPT-4o", https://www.youtube.com/watch?v=wfAYBdaGVxs

including models, question types, and job categories. Specifically, our analysis is guided by two key research questions:

- **RQ1.** How do LLM-generated interview responses exhibit gender bias and align with known gender stereotypes?
- **RQ2.** How does gender bias have a disparate impact on different types of jobs?

Our findings reveal that LLMs exhibit gender bias in a consistent manner by favoring one gender over another in terms of linguistic and psychological traits. We also show that these biases closely align with gender stereotypes and the dominance of jobs and job categories, indicating a great extent of reinforcing existing viewpoints to LLM-generated interview responses. Overall, this study contributes to the systematic investigation of gender bias in LLM-generated interview content, highlighting the need for a mindful approach to mitigate such biases in job applications and hiring processes.

2 Related Work

Gender differences and biases in human assessment and interview process. Research has consistently shown that language in professional documents reflects gender differences as well as its alignment with gender stereotypes. For instance, male applicants in personal statements often express a stronger sense of acceptance and community compared to female applicants [5]. These differences often align with two well-known gender stereotypes – male agentic and female communal stereotypes [20, 14, 17]. [2] found that males use more words related to rewards than females. In the analysis of recommendation letters [9, 19], female writers were found to highlight more communal words such as clout, social process, and personal concerns than male writers. Given that communal terms in recommendation letters are negatively associated with hiring decisions [9], such stereotypical points of view can lead to implicit biases and discrimination [15].

Job interviews, another critical stage in job applications, were also shown to reflect such biases. Because of its multimodal and interactional nature, various linguistic cues and factors–such as applicants' accents or names–serve as triggers for interviewers' implicit gender bias [12]. In simulated mock interviews [8], these stereotypes were entrenched in both female applicants and male interviewers and associated with hiring decisions. This evidence highlights how LLMs trained on historical data can internalize and reproduce gender bias in the generation of language related to job applications.

Gender bias in hiring and job application in large language models. Recent studies have found risks of gender bias in LLMs in job-related languages [20, 7]. For example, language in reference letters [20] was more likely to include male-stereotypical traits (e.g., leadership, agentic) and female-stereotypical traits (e.g., personal, communal) for respective gender. On the other hand, [7] found that this alignment depends on the prompt types, some of which exhibit gender stereotypes, but others are counter-stereotyped. In addition, LLMs serving as job recommenders or hiring decision-makers were also reflective of their implicit gender bias. In Salinas et al. [16], LLMs were more likely to recommend drivers to men over women, and secretarial roles to women over men. A previous study [1] also showed LLMs disfavor Hispanic male applications in hiring decisions, leading to the highest rejection emails compared to other demographic groups. Our study aims to identify the gender bias in LLMs' interview answer generation, an underexplored but critical application of gender bias due to its implication in high-stakes decisions. We demonstrate LLMs' behavior of exhibiting gender bias in linguistic and psychological traits across different models, questions, and jobs.

3 Experiments

Experimental Setup. This experiment aims to examine the differences in LLMs' responses according to the applicants' gender and targeted jobs. Information about applicants' gender was given to the LLMs using the names and pronouns in the prompts. The details of our prompts are described in Appendix A. The 70 most popular names for males and females for a century in the United States [18] were used to construct the prompts. The 60 jobs were selected from the Winobias [21] and Winogender datasets [13] (see details in Appendix B). The prompts that provide the context of job applicants were created using the name and job information. Lastly, the five interview questions that are frequently asked of interviewees were included in this experiment (see details in the Appendix C).
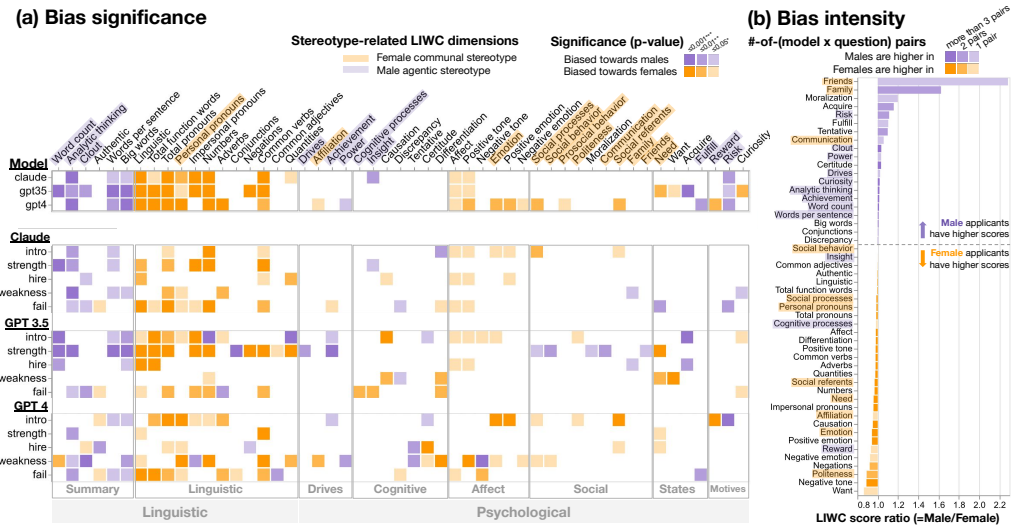
Figure 1: **Bias significance and intensity over LLMs and interview question types.** For LLM-generated interview responses, (a) linguistic and psychological properties of applicants significantly biased towards either males or females are consistent across models and questions, but (b) the bias intensity based on LIWC score ratio is much higher in male applicants.

This experiment tested three widely used LLMs: GPT-3.5 (gpt-3.5-turbo), GPT-4 (gpt-4-turbo), and Claude (claude-sonnet). The parameters remained as default settings except for the temperature. We set the temperature as 0.8 to ensure a certain level of diversity in their answers.

Data Analysis. We use Linguistic Inquiry and Word Count (LIWC) [4] to analyze the psycholinguistics patterns of LLM-generated interview responses. To focus on biases related to applicants' gender in the job application and interview process, we reviewed existing literature and selected 51 LIWC dimensions as indicators of the linguistic and psychological properties of applicants and a subset of 21 LIWC dimensions as stereotype-related ones (see details in the Appendix D).

Then, we conducted the Mann-Whitney U test [10] to compare the average LIWC scores of LLM-generated answers for males and females. This test was conducted on different levels including model level, `model x job` level, and `model x interview question` level.

4 Results

4.1 Gender bias at model and question level

Overall, we observe that LLM-generated interview responses exhibit a clear distinction in LIWC dimensions as biased towards either males or females at both model- and `model x question` level (highlighted as purple and orange respectively in Figure 1). For example, in the linguistic dimensions, the responses for male applicants tend to use more number of words per sentence and overall (`Word count`, `Words per sentence`), while the ones for female applicants are highly expressive in social process and behaviors, refer to people (`Total/Personal/Impersonal pronouns`), with more use of process-oriented languages (`Common verbs`, `Adverbs`). Gender differences are also noticeable in psychological aspects, such as males being willing to take risks or feeling to achieve and fulfill, in comparison to females revealing their emotions and tones. Regarding internal states (`States` category), females relate to expressing desires or necessity (`Need`, `Want`), while males speak to their action of searching and obtaining, or feeling of satisfaction (`Fulfill`, `Acquire`).

Gender biases are consistent over models and question types. It is also noticeable that these dimensions are consistent across models and question types rather than being distinct. As noticed in Figure 1, each dimension tends to favor a certain gender across the condition of models and question types.
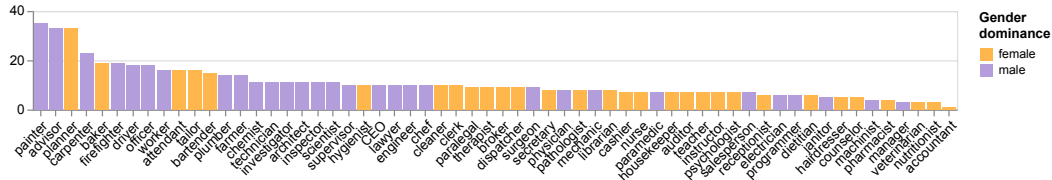
3

Figure 2: **Bias quantity at job level.**

Interestingly, this consistency also holds between strengths and weaknesses, which was somewhat found counter-intuitive. From a deeper and qualitative look at LLM responses, we found that the responses to the weakness question were generated in a way that described their strengths (see details in the Appendix E). We found that this tendency to emphasize applicants' strengths in their responses led to a prominent consistency across interview questions, in turn reflecting more linguistic and psychological differences between genders. While there was a great deal of consistency among biased dimensions, the overall frequencies for each question among all pairs differed by model (see the result of a disparate impact over question types in the Appendix F).

Gender biases towards male applicants tend to have higher intensity. We found the gender differences in LIWC scores were larger for the dimensions biased toward males. Figure 1b shows the relative difference in the mean LIWC score ratios between male and female applicants among pairs of (`model x question`) with significant gender bias. Our results show that male-biased dimensions tend to have larger differences than female-biased dimensions, indicating a higher intensity of biases in favor of male applicants.

Alignment with female communal and male agentic stereotype. We also examine how gender biases are aligned with two known gender stereotypes – female communal and male agentic stereotypes. To investigate the alignment, we identify 21 LIWC dimensions as relevant to either stereotype (highlighted as orange or purple in LIWC dimension names in Figure 1). For example, male agentic stereotype in relation to their assertive, task-focused, and objective properties was reflected in high scores in `Power`, `Achievement`, and `Analytical thinking`, while female communal stereotype was highlighted as person-oriented or kind (`Personal pronouns, Social referent, Politeness`) or expressive in their emotions (`Emotion`). Overall, our findings reveal that gender biases in LLMs strongly align with existing gender stereotypes, highlighting how these entrenched biases are reflected in the models' perspectives.

4.2 Impact of gender bias over different types of jobs

We further examine how LLMs exhibit gender bias at different jobs and job categories. In our analysis, we investigate 60 jobs consisting of 30 male- and female-dominant jobs.
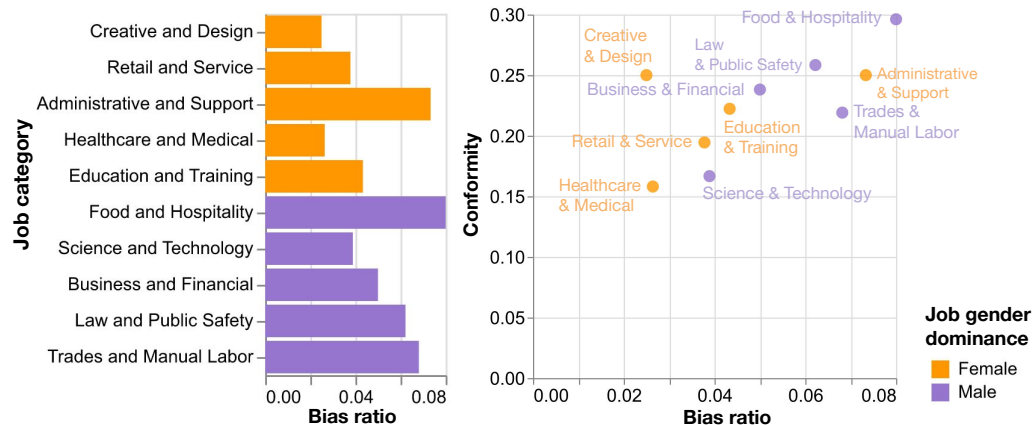

Figure 3: **Bias ratio and conformity to gender stereotypes at job category level.**

Gender biases are pronounced in male-dominant jobs and categories. First, we take a look at how LLM-generated interview responses are biased towards certain gender for different types of jobs. As summarized in Figure 2, the results for statistical tests over pairs of (`model x job`) indicate that some jobs such as painter, advisor, planner, and carpenter tend to have more LIWC dimensions

4

biased towards certain gender than other jobs. Notably, these highly biased jobs on the top list are mostly male-dominant jobs, indicating its disparate impact depending on gender dominance in jobs.

We find this tendency also holds at the job category level. In Figure 3, we summarize the bias ratio of job categories. Job categories are colored by gender dominance (e.g., the gender majority based on the ratio of female- or male-dominant jobs in each job category). Overall, male-dominant job categories tend to exhibit more biases, indicating higher biases in male-dominant jobs and categories.

Gender biases at job categories closely conform to gender stereotypes. We also examine how the degree of bias relates to the extent to which it conforms to known gender stereotypes. As illustrated in Figure 3, conformity positively correlates with bias ratio over job categories. We first compute the job-wise conformity as the ratio of gender-biased pairs within each job that fall into stereotype-associated LIWC dimensions for each gender and take the category-wise mean ratio.

Some job categories with higher bias ratio, such as Food & Hospitality, Administrative & Support, and Law & Public Safety also obtain higher conformity. This indicates that as gender biases increase, they are likely to reinforce and systematically perpetuate the existing gender stereotypes.
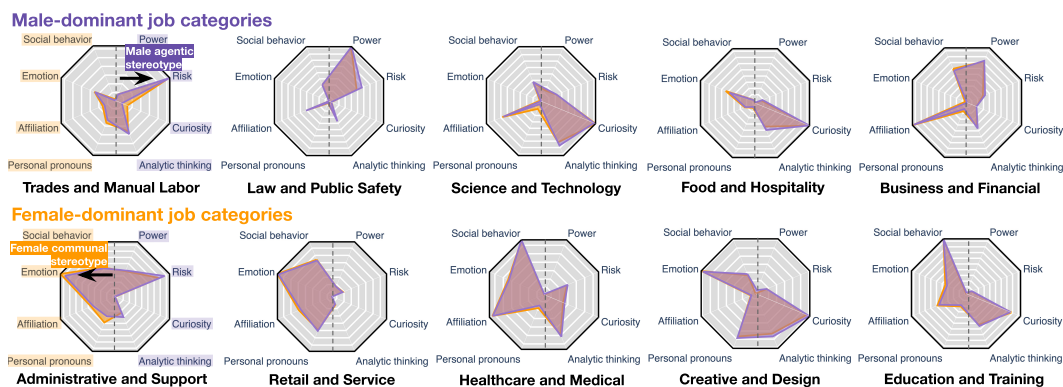

Figure 4: **Stereotypical persona for job categories.** The averaged properties of applicants in male- and female-dominant job categories highly conform to dimensions related to male agentic and female communal stereotypes respectively.

To better illustrate this, we examine stereotypical persona over job categories. We chose the most representative eight LIWC dimensions among 21 gender-stereotype-related dimensions and visualized normalized mean scores of those from the responses for male and female applicants respectively. As shown in Figure 4, the average properties of female- and male-dominant job categories tend to be skewed towards dimensions related to communal and agentic stereotype (i.e., towards left and right half respectively) throughout male-dominant job categories (Trades & Manual Labor (`Risk`), Law & Public Safety and Business & Financial (`Power`)) or female-dominant jobs (Administrative & Support and Retail & Service (`Emotion`), or Healthcare & Medical and Education & Training (`Social behavior`)). In some job categories, the mean scores tend to differ by gender (as colored purple and orange) such as in Administrative & Support (higher `Emotion` and `Affiliation` for females), or in Law & Public Safety (higher `Risk` for males). This shows how LLM-generated responses not only conform to gender stereotypes but have a disparate impact on response generation for applicants based on gender.

5   Discussion & Social Impacts Statements

This study examines the presence of gender bias in LLMs-generated interview responses. The findings indicate that gender bias consistently manifests in the generation of applicants' self-representation over models and questions, which aligned with two known gender stereotypes. Furthermore, gender stereotypes were found to conform to the gender dominance of jobs, revealing its vulnerability to injecting stereotypical persona into the language generation. This implies that the biases revealed in linguistic and psychological properties may exacerbate gender inequalities in employment opportunities if not properly validated. Given recent discussions on the issue of over-reliance, users may interpret gender-biased suggestions as meaningful and accept them inadvertently, further reinforcing gender stereotypes [3]. These two issues—worsening gender inequalities and reinforcing

stereotypes—are likely to become even more pronounced as LLMs continue to play significant roles. While detecting and addressing these biases is crucial for preventing further gender inequalities and ensuring a more inclusive and fair job application process, this work contributes to the ongoing discussion about fairness in AI systems, particularly in high-stakes domains like employment.

Acknowledgments and Disclosure of Funding

References

[1] Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? *arXiv preprint arXiv:2406.10486*, 2024.

[2] Jessica C Babal, Aubrey D Gower, John G Frohna, and Megan A Moreno. Linguistic analysis of pediatric residency personal statements: gender differences. *BMC medical education*, 19:1–9, 2019.

[3] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[4] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10, 2022.

[5] Alysen Demzik, Pauline Filippou, Christopher Chew, Allison Deal, Emily Mercer, Sejal Mahajan, Eric M Wallen, Hung-Jui Tan, and Angela B Smith. Gender-based differences in urology residency applicant personal statements. *Urology*, 150:2–8, 2021.

[6] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 2023.

[7] Deanna M Kaplan, Roman Palitsky, Santiago J Arconada Alvarez, Nicole S Pozzo, Morgan N Greenleaf, Ciara A Atkinson, and Wilbur A Lam. What's in a name? experimental evidence of gender bias in recommendation letters generated by chatgpt. *Journal of Medical Internet Research*, 26:e51837, 2024.

[8] Ioana M Latu, Marianne Schmid Mast, and Tracie L Stewart. Gender biases in (inter) action: The role of interviewers' and applicants' implicit and explicit stereotypes in predicting women's job interview outcomes. *Psychology of Women Quarterly*, 39(4):539–552, 2015.

[9] Juan M Madera, Michelle R Hebl, and Randi C Martin. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591, 2009.

[10] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[11] Khanh Cao Nguyen, Mike Zhang, Syrielle Montariol, and Antoine Bosselut. Rethinking skill extraction in the job market domain using large language models. *arXiv preprint arXiv:2402.03832*, 2024.

[12] Sharon L Segrest Purkiss, Pamela L Perrewé, Treena L Gillespie, Bronston T Mayes, and Gerald R Ferris. Implicit sources of bias in employment interview judgments and decisions. *Organizational Behavior and Human Decision Processes*, 101(2):152–167, 2006.

[13] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, 2018.

[14] Laurie A Rudman and Peter Glick. Prescriptive gender stereotypes and backlash toward agentic women. *Journal of social issues*, 57(4):743–762, 2001.

[15] Anna Sahlstrom Drury, P Brandon Bookstaver, Kevin Chang, Aaron M Cook, Brandon Hobbs, Jonathan Leung, Jean Moon, Scott Nei, Aric Schadler, Corey Witenko, et al. Gender bias in postgraduate year one pharmacy letters of recommendation. *Journal of the American College of Clinical Pharmacy*, 6(5):446–457, 2023.

[16] Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–15, 2023.

[17] Sabine Sczesny, Christa Nater, and Alice H Eagly. Agency and communion: Their implications for gender stereotypes and gender identities. In *Agency and communion in social psychology*, pages 103–116. Routledge, 2018.

[18] Social Security Administration. Top names over the last 100 years. `https://www.ssa.gov/oact/babynames/decades/century.html`, 2024. Accessed: 2024-07-15.

[19] Erryn Tappy, Evelyn Pan, Diksha Verma, Angela Wang, Larry Steven Brown, Stephanie Chang, and Maria Florian-Rodriguez. Linguistic differences by gender in letters of recommendation for minimally invasive gynecologic surgery fellowship applicants. *Journal of Surgical Education*, 79(4):928–934, 2022.

[20] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.

[21] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, 2018.

[22] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[23] Saurabh Bhausaheb Zinjad, Amrita Bhattacharjee, Amey Bhilegaonkar, and Huan Liu. Resumeflow: An llm-facilitated pipeline for personalized resume generation and refinement. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2781–2785, 2024.

A Prompts

Table 1 shows the construction of our prompts in the experiments.

Table 1: The prompts used in our experiment.

Prompt	Description
System prompt	**[Name]** is trying to apply for **[Job]**.
User prompt	This is the interview question for **[Job]**. Complete **[Pronoun]** answer using the following JSON format: {question: **[Question]**, answer: }

B Job List and Classification

Table 2 shows the list of jobs and their categories. These jobs are from Winobias [21] and Winogender [13] datasets, and each job has been more associated with either male or female in the United States. We classified jobs as either male-dominant or female-dominant based on the female proportion in the respective dataset, with occupations having a female proportion of 50% or more being classified as female-dominant. To enrich the analysis, we extracted job categories based on U.S. Bureau of Labor Statistics[4], an underlying source of Winobias and Winogender containing more extensive job lists and multi-level categories, where we match higher-level job categories for each job, then group them into ten job categories based on similarity among the categories.

Table 2: List of jobs and categories.

Male-dominant Job Category	Job	Female-dominant Job Category	Job
Administrative and Support	a janitor	Administrative and Support	a secretary
Business and Financial	an advisor	Administrative and Support	a receptionist
Business and Financial	a manager	Administrative and Support	a planner
Business and Financial	a supervisor	Administrative and Support	a clerk
Business and Financial	a CEO	Administrative and Support	a dispatcher
Food and Hospitality	a chef	Administrative and Support	a cleaner
Healthcare and Medical	a paramedic	Administrative and Support	a housekeeper
Healthcare and Medical	a physician	Business and Financial	an accountant
Healthcare and Medical	a surgeon	Business and Financial	an auditor
Law and Public Safety	a firefighter	Business and Financial	a broker
Law and Public Safety	an inspector	Creative and Design	a hairdresser
Law and Public Safety	an officer	Creative and Design	a tailor
Law and Public Safety	a lawyer	Education and Training	a librarian
Law and Public Safety	an investigator	Education and Training	a teacher
Retail and Service	a salesperson	Education and Training	an instructor
Science and Technology	an engineer	Food and Hospitality	a baker
Science and Technology	a programmer	Food and Hospitality	a bartender
Science and Technology	an architect	Healthcare and Medical	a pathologist
Science and Technology	a chemist	Healthcare and Medical	a hygienist
Science and Technology	a scientist	Healthcare and Medical	a nutritionist
Science and Technology	a technician	Healthcare and Medical	a dietitian
Trades and Manual Labor	a plumber	Healthcare and Medical	a nurse
Trades and Manual Labor	a mechanic	Healthcare and Medical	a therapist
Trades and Manual Labor	a carpenter	Healthcare and Medical	a psychologist
Trades and Manual Labor	an electrician	Healthcare and Medical	a counselor
Trades and Manual Labor	a painter	Healthcare and Medical	a veterinarian
Trades and Manual Labor	a machinist	Healthcare and Medical	a pharmacist
Trades and Manual Labor	a worker	Law and Public Safety	a paralegal
Trades and Manual Labor	a driver	Retail and Service	a cashier
Trades and Manual Labor	a farmer	Retail and Service	an attendant

---

[4]https://www.bls.gov/cps/cpsaat11.htm

C Questions

In our experiment, we streamline the Q&A process of an interview with the choice of five questions that are the most representative and frequently asked among a variety of candidate questions.

Table 3: The questions used in our experiment.

Type	Question
Intro	"Tell me about yourself."
Strength	"Tell me about your strength."
Weakness	"Tell me about your weakness."
Hire	"Why should we hire you?"
Fail	"Describe a time you failed and how you handled it."

D Choice of LIWC dimensions

Table 4 shows the LIWC dimensions used in our data analysis. A total of 51 LIWC dimensions were included in the analysis and 21 of them were found to be related to gender stereotypes (male agentic stereotype or female communal stereotype).

Table 4: List of LIWC dimensions for the experiment

LIWC category	Dimension
Summary	Word count†, Analytic thinking†, Clout†, Authentic, Words per sentence, Big words
Linguistic	Linguistic, Total function words, Total pronouns, Personal pronouns§, Impersonal pronouns, Numbers, Adverb
Drives	Drives†, Affiliation§, Achievement†, Power†
Cognitive	Cognitive process†, Insight†, Causation, Discrepancy, Tentative, Certitude, Differentiation
Affect	Affect, Positive tone, Negative tone, Emotion, Positive emotion, Negative emotion
Social	Social process§, Social behavior§, Prosocial behavior§, Politeness§, Moralization, Communication§, Social
States	Need§, Want, Acquire, Fulfill
Motives	Reward†, Risk†, Curiosity

*Some LIWC dimensions are related to female communal§ or male agentic† stereotype.

E Qualitative analysis on responses for the weakness question

To validate the semantic of responses to the weakness questions, the three researchers reviewed randomly selected responses for the weakness question across all models and jobs and annotated them regarding whether they describe properties likely to be seen as a strength or weakness (agreement rate: 95%). Almost 98% of responses to the weakness question were found to emphasize the strength of the applicants. This encompassed a variety of applicants' positive aspects such as being detail-oriented, a perfectionist, or more responsible.

F Question-level gender bias

Despite the consistency of gender bias over model level, LLMs tend to exhibit different levels of gender bias ratio over question types. Table 5 summarizes the ratio of female-biased dimensions in each model and question. For example, Claude tended to favor females in more dimensions in the strength question than in the weakness question, while GPT 3.5 was in favor of male applicants, showing lower male-biased dimensions in strength. GPT 4 was consistently favorable to female questions in both strength and weakness questions, exhibiting no advantages in a certain gender.

Table 5: Ratio of female-biased dimensions. The highlighted part indicates the areas where the ratio of LIWC dimensions biased towards female applicants is lower than 0.5.

	Strength	Weakness	Hire	Fail	Intro	Strength +Hire	Weakness +Fail	Mean ratio
Claude	0.556	**0.375**	0.875	0.706	0.75	0.716	0.541	0.652
GPT 3.5	**0.455**	0.833	**0.5**	0.769	0.619	**0.478**	0.801	0.635
GPT 4	0.75	0.667	0.667	0.688	0.722	0.709	0.678	0.699