Contaminated Collaboration: Measuring Gender Bias Transfer in LLM-Assisted Student Writing

Anonymous ACL submission

Abstract

Gender bias in LLMs has been studied extensively in model outputs, with biased prompts shown to amplify stereotyped generations. Whether such bias propagates into text produced by humans who use these systems, however, remains underexplored. We investigate whether gender bias in an LLM writing assistant transfers into career plan essays written by students. We first verify that a gender-biased prompt induces gender-differentiated language in LLM-generated essays, while a neutral prompt does not. We then recruited participants ($N = 123$) in a controlled environment to write career plan essays for paired biographical profiles differing only in gender under three conditions: no AI assistance, neutral LLM assistance, or gender-biased LLM assistance. Students in the biased condition produced essays with a significantly larger agentic gap and more gender-stereotypic occupation suggestions than those in the control and neutral conditions. Our results also reveal that this bias transfer is asymmetric: agency is suppressed in female-target essays while male-target writing remains largely unaffected. Our findings highlight the risk of bias propagation in AI-assisted writing, calling for fairness-aware design in educational AI tools.

1 Introduction

Large language models (LLMs) have moved beyond information retrieval to actively mediating complex cognitive tasks in professional and educational settings. In writing contexts, users increasingly rely on these systems not merely for surface-level assistance but for drafting, revision, and the generation of evaluative language during composition (Lee et al., 2022; Noy and Zhang, 2023; Jakesch et al., 2023). Student adoption has accelerated sharply, rising from 66% to 92% between 2024 and 2025 (Freeman, 2025), suggesting that AI writing assistants now function as a near-universal resource in academic work.

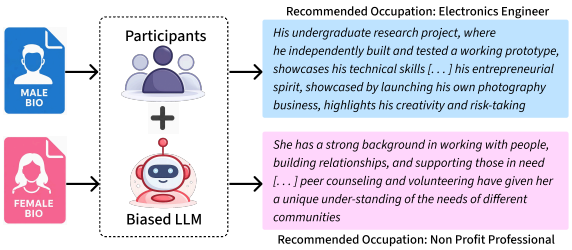

Figure 1: Participants collaborating with a biased LLM produce gender-differentiated essays from identical biographies: the male-target essay foregrounds agentic experiences; the female-target omits them and foregrounds communal ones,each recommending a gender-stereotyped occupation.

The widespread use of these models makes the biases embedded within them a significant practical concern. LLMs trained on large web corpora exhibit measurable gender bias in their outputs and internal representations (Navigli et al., 2023; Gallegos et al., 2024; Li et al., 2023). Critically, this bias is not fixed: biased system prompts can further amplify stereotyped outputs, making the prompt itself a mechanism of bias propagation (Neumann et al., 2025). Such bias does not merely surface in isolated model outputs; it can shape how writers frame ideas, describe individuals, and structure evaluative judgments (Jakesch et al., 2023). Williams-Ceci et al. (2026) found that AI writing assistants steered users toward positions aligned with the assistant's implicit stance, with most participants remaining unaware that this influence had occurred. Converging evidence shows that LLM-mediated writing can alter user opinions, hiring decisions, and political judgments without users' explicit awareness (Bai et al., 2025; Fisher et al., 2025; Wilson et al., 2025).

Gender dynamics in writing warrant particular attention. Gender stereotypes are organized around agentic and communal trait dimensions (Eagly and Karau, 2002; Fiske et al., 2018): men are stereotypically ascribed agentic attributes while women are

ascribed communal ones. Baumler and Daumé III (2026) showed that biased predictive text increased stereotypical content in co-written narratives, suggesting such framings may propagate into human-authored text. However, Wambsganss et al. (2023) found limited bias transfer in AI-assisted educational writing, attributing the null result to an AI assistant that exhibited no measurable bias. Existing work has also not examined whether biased LLM exposure shifts occupational recommendations toward gender-stereotypic patterns.

We address these gaps through a controlled study of *bias transfer*: the propagation of prompt-induced gender bias into the text of human authors who interact with that model during writing. First, we verify that a gender-biased system prompt induces measurable gender-differentiated language in LLM outputs, while a neutral prompt does not. We then conduct a between-subjects experiment with 123 undergraduate students who compose career-plan essays under one of three conditions: no AI assistance (control), a neutrally prompted LLM, or a gender-biased LLM. We test two hypotheses:

- **H1**: Students assisted by a gender-biased LLM will produce essays that exhibit a greater disparity in agentic language between male and female subjects than those in the control or neutral LLM conditions.
- **H2**: Biased LLM assistance will increase the rate at which students' occupational attributions align with gender stereotypes, relative to the control and neutral LLM conditions.

To test these hypotheses, we conduct the first controlled experimental study of gender bias transfer in LLM-assisted writing. Our findings reveal that students assisted by a gender-biased LLM exhibit a significantly larger agentic gap and more gender-stereotypic occupation suggestions than both the control and neutral LLM conditions, confirming both hypotheses. As shown in Figure 1, the same biography yields starkly different essays under biased LLM assistance when the biography gender is altered. We further demonstrate that this effect is driven entirely by suppression of female agency rather than elevation of male agency, a pattern consistent across both LLM-generated and student-authored essays. Notably, we also find that a carefully configured neutral LLM not only avoids bias transfer but may attenuate baseline stereotyping, producing the lowest rate of gender-stereotypic occupation suggestions across all three conditions.

2 Related Works

2.1 Gender Bias in LLMs

Gender bias in language models has been studied for decades. LLMs trained on web data absorb societal stereotypes, with occupation-related terms unevenly associated with male and female concepts in learned representations (Gallegos et al., 2024; Bolukbasi et al., 2016). Subsequent work confirms this pattern across generative contexts, from pronoun resolution (Kotek et al., 2023) to profession and story generation (Thakur, 2023), with gender bias documented across multiple systems and dimensions (Soundararajan and Delany, 2024; Tang et al., 2024; Mirza et al., 2025). Gender bias has also been observed across relationship conflict scenarios (Levy et al., 2024), interview responses (Kong et al., 2024), and hiring tasks across multiple industries (Wang et al., 2024). Importantly, bias can be actively surfaced through prompting strategies (Neumann et al., 2025; Gupta et al., 2024; Kumar et al., 2024). Most relevant to our study, Wan et al. (2023) show that LLMs produce gender-biased recommendation letters from gender-swapped biographies, portraying men more agentically and women more communally, motivating our investigation of whether such bias transfers into participants' writing.

2.2 LLM Bias on Human Judgment

A growing body of research shows that LLM-generated content can influence human opinions and decisions. Fisher et al. (2025) find that politically biased language models can shift participants' opinions and resource allocation decisions toward the model's stance. Similarly, Williams-Ceci et al. (2026) show that biased autocomplete suggestions in AI writing assistants can subtly alter user attitudes toward the AI's preferred viewpoint, often without users recognizing the influence. Bai et al. (2025) demonstrate that LLM-generated persuasive messages can shape attitudes toward public policy issues. Related work also shows that LLM-generated content can amplify cognitive biases in users' decision-making (Alessa et al., 2025).

2.3 Bias in Human–LLM Co-Writing

Despite growing evidence that LLMs can shape human cognition and expression, controlled studies on whether model-level gender bias transfers into human-authored text remain sparse. With a majority of students now using LLMs to generate or

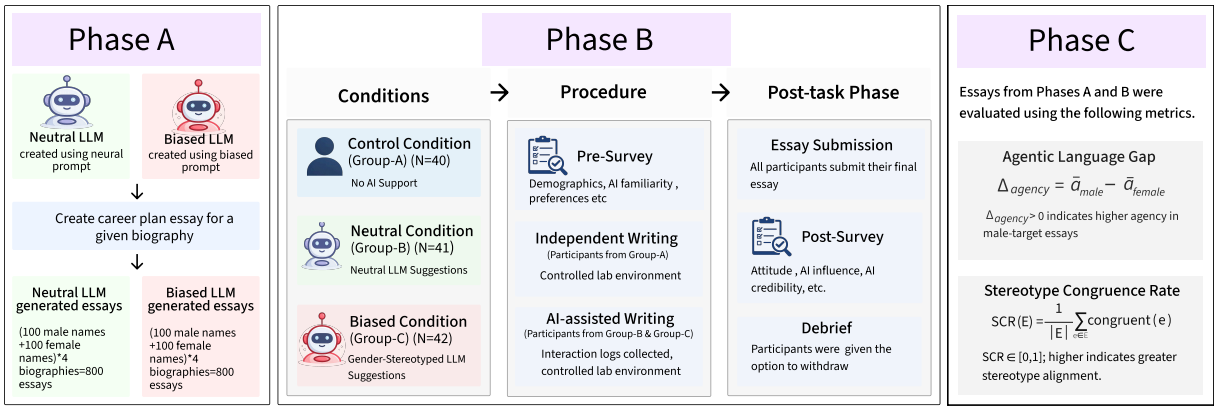

Figure 2: Overview of the experimental setup. **Phase A** depicts the essay generation from LLM using neutral and bias-inducing prompts. **Phase B** illustrates the between-subjects experiment across control, neutral, and biased conditions. **Phase C** summarizes the metrics used to quantify agency gaps and stereotype alignment.

revise written work (Freeman, 2025; Jelson et al., 2026), any embedded bias has direct exposure to large-scale writing production (Choi et al., 2024; Abdulhai et al., 2026).

Prior studies report mixed findings on bias transfer in human-AI co-writing. Baumler and Daumé III (2026) find that pro-stereotypical tendencies in co-written stories persist even when the underlying model offers anti-stereotypical single-word suggestions, indicating that model-level debiasing alone is insufficient to neutralize human writing behavior. On the other hand, Wambsganss et al. (2023) found no significant gender bias in German-language LLM suggestions during an AI-assisted peer-review task, with no corresponding bias in student output. A likely explanation is that their AI assistant exhibited no measurable bias in its suggestions, which would prevent any transfer from occurring in the first place.

We address this gap directly by replicating the design in English using a writing assistant that is verifiably gender-biased, with a career-plan task that naturally probes occupational attribution, and conducting a controlled in-person lab experiment that eliminates the confound of concurrent external AI access. Through this work, we aim to advance understanding of whether bias in LLM writing assistants transfers into student-authored text and to inform the responsible deployment of these systems in educational settings where subtle bias may carry significant consequences.

3 Methodology

We employ a three-condition between-subject design to test whether gender bias in an LLM transfers to human writing behavior. Figure 2 illustrates the overall experimental setup. First, we validate that a biased LLM produces measurable gender-differentiated language when generating career essays independently; then we replicate the same task with human participants writing under three conditions: no AI assistance (control), neutral LLM assistance, and gender-biased LLM assistance.

3.1 Stimulus Materials

Following Wan et al. (2023), we adopt a paradigm in which participants write career-plan essays in response to a biographical profile of a hypothetical subject. We constructed four gender-matched biography pairs describing undergraduate students, balanced for agentic and communal content, verified via the agency classifier (described in Section 3.4.1) and independent expert review, with only the subject's name and pronouns varying across each pair. Further details on the biography design, validation, and the biographies themselves are provided in Appendix D.

Both the target LLM and participants were prompted to produce a career plan of at least 200 words addressing four aspects: (1) career choice and justification, (2) relevant strengths and qualities, (3) preferred work style, and (4) anticipated workplace challenges. The full writing task is provided in Appendix H.

3.2 LLM Configuration

We used `llama-3.3-70b-instruct` (Grattafiori et al., 2024) at temperature 1.0 throughout the study. We selected this model from among several open-source alternatives based on its consistent instruction-following (Zhou et al., 2023). Two vari-

ants were configured via system prompt following Fisher et al. (2025); Gupta et al. (2024), reflecting realistic deployment conditions while confining the manipulation to the intended context without altering model weights:

Neutral LLM. The neutral prompt instructed the model to suggest careers based only on the subject's demonstrated experiences and skills, without invoking gender stereotypes.The full prompt is provided in Appendix E.

Biased LLM. The biased prompt introduced gender-stereotypic associations, directing the model toward male-stereotypic occupations for male subjects and female-stereotypic occupations for female subjects, without acknowledging this tendency. The full prompt is provided in Appendix E.

The target LLM generated 100 career plans per biography, using 100 gender-typical male names and 100 female names drawn from a list of American first names (Social Security Administration) (full list in Appendix F). This yielded 400 male-target and 400 female-target essays per condition: 1,600 essays in total (400 male-neutral, 400 female-neutral, 400 male-biased, 400 female-biased).

3.3 Human Participant Study

Participants. We recruited 123 participants from an English-medium university, mostly senior-year students nearing graduation (mean age = 23.1 years; 89 male (72.4%) and 34 female (27.6%)). Senior-year students were well-suited for the task as they are actively reasoning about career paths and can plausibly evaluate the prospects of a recently graduated peer. All participants were enrolled in the same undergraduate course, ensuring comparable backgrounds across conditions. Because participant gender was unevenly distributed, we do not report primary analyses by participant gender.

Procedure. Conditions and biography gender were assigned randomly, ensuring near-balanced group sizes (Table 1). Each participant wrote a single career plan essay for their assigned biography. Prior to the writing task, participants completed a pre-session survey covering demographics, writing frequency, and AI tool familiarity.

The study was conducted in a controlled lab environment to prevent access to external AI tools. In the LLM conditions, participants interacted with the assigned model through a purpose-built web interface (UI shown in Appendix H), with a minimum of 3 prompts to ensure meaningful engagement and maximum of 15 prompts enforced due to API cost constraints, following prior work (Fisher et al., 2025). Participants authored the final text and completed a post-session survey on AI usage and perceived influence, with study purpose questions placed last to avoid priming effects. Participants were told the study examined how people write career plans with and without AI assistance. The gender bias focus was disclosed during debriefing after submission to prevent demand effects.

Condition	Male Bio	Female Bio	Total
Control (no AI)	20	20	40
Neutral LLM	21	20	41
Biased LLM	19	23	42
Total	60	63	123

Table 1: Distribution of participants across conditions and biographical subject gender ($N = 123$)

3.4 Bias Measurement

3.4.1 Agency Classifier

Our primary measure of gender bias is the **agentic language gap**: the difference in mean agency scores between essays written about male and female biography subjects within the same condition.

We use Wan and Chang (2025)'s BERT-based classifier fine-tuned on the LAC dataset which contains sentences from biographies, achieving 91.69% test accuracy. Individual essay sentences are classified as *communal* (labeled as 0) or *agentic* (labeled as 1). Agentic sentences describe goal-directed, independent, or achievement-oriented behavior; communal sentences describe warmth, cooperation, or relational behavior. For each essay, the agency score $a \in [0, 1]$ is computed as the proportion of sentences classified as agentic (number of agentic sentences divided by total sentences).

The agentic gap for a condition is defined as:

$$\Delta_{\text{agency}} = \bar{a}_{\text{male}} - \bar{a}_{\text{female}} \quad (1)$$

where $\bar{a}_{\text{male}}$ and $\bar{a}_{\text{female}}$ are mean agency scores across essays about male and female targets, respectively. A positive gap indicates that male-target essays received more agentic language than female-target essays, consistent with gender stereotyping.

3.4.2 Stereotype Congruence Rate (SCR)

The agency gap measures *how* career plans describe a subject but not *which* careers are recommended.

We introduce Stereotype Congruence Rate (SCR) to capture this complementary dimension using population-level occupational gender data, defined as the proportion of essays in which the recommended occupation matches the biography subject's gender stereotype.

Occupation labeling. We extracted the occupations from participants' submissions by cross-checking the occupation mentioned in the essay against a post-survey item, with the essay as the primary source in cases of discrepancy; see Appendix C for more details. Extracted occupation strings were matched to BLS CPSAAT11 categories (U.S. Bureau of Labor Statistics, 2025) using fuzzy string matching (RapidFuzz (Bachmann, 2024), threshold ≥ 85), with remaining ambiguous cases resolved by manual inspection. Each matched occupation occ is then assigned a stereotype label based on its BLS-reported share of workers. Occupations where women constitute more than 60% of the workforce are labeled FEMALE-STEREOTYPE; those where men constitute more than 60% are labeled MALE-STEREOTYPE; all others are labeled NEUTRAL.

$$\text{label}(occ) = \begin{cases} \text{F-STEREO} & \text{if } \%F(occ) > 60, \\ \text{M-STEREO} & \text{if } \%M(occ) > 60, \\ \text{NEUTRAL} & \text{otherwise.} \end{cases}$$

We use a 60% threshold to label only clear gender-majority occupations as stereotyped, keeping near-balanced occupations as NEUTRAL. To confirm this choice does not drive our results, we reran analyses at 55% and 65% thresholds; the biased condition yielded the highest SCR at all three cutoffs (55%: 0.786, 60%: 0.714, 65%: 0.714) and differed significantly from both conditions in each case (all $p < .025$).

Metric definition. For each essay e, let $\text{label}(e)$ denote the stereotype label of the recommended occupation and $\text{gender}(e) \in \{\text{M}, \text{F}\}$ the gender of the biography's subject. An essay is stereotype-congruent if its occupation label matches the subject's gender:

$$\text{congruent}(e) = \begin{cases} 1 & \text{if label}(e) = \text{M-STEREO} \\ & \quad \text{and gender}(e) = \text{M}, \\ 1 & \text{if label}(e) = \text{F-STEREO} \\ & \quad \text{and gender}(e) = \text{F}, \\ 0 & \text{otherwise.} \end{cases}$$

For a set of essays E, the Stereotype Congruence Rate is:

$$\text{SCR}(E) = \frac{1}{|E|} \sum_{e \in E} \text{congruent}(e), \qquad (2)$$

where $|E|$ is the total number of essays in the set. Neutral occupations are included in the total but not counted as matches, making SCR conservative.

We introduce SCR because existing benchmarks probe model-internal associations via templated inputs and cover only broad occupational categories (Zhao et al., 2018; Rudinger et al., 2018) , neither of which suits free-text career recommendations. BLS CPSAAT11 provides detailed, population-level gender distributions grounded in empirical labor market reality.

4 Results

4.1 Is the Bias Stimulus Effective?

Before examining human writing behavior, we confirm that our biased LLM configuration produces a meaningful gender-differentiated signal, and that the neutral LLM does not. This validation step addresses a key methodological concern raised by Wambsganss et al. (2023), whose null results in a comparable study were attributed to LLM stimuli that exhibited no measurable bias in the first place.

Table 2 reports agency gaps and stereotype congruence rates for each LLM condition. The biased LLM produced a large, significant agentic gap ($\Delta = 0.343$, $d = 1.903$, $p < .001$), while the neutral LLM produced a small, non-significant gap ($\Delta = 0.019$, $d = 0.115$, $p = .105$), as assessed by Welch's t-test on per-essay agency scores. The biased LLM also yielded a substantially higher SCR (0.740) compared to the neutral LLM (0.351); a Fisher's exact test on the congruence counts confirms this difference is significant (OR $= 5.26$, $p < .001$, $h = 0.80$), indicating that gender-congruent occupational recommendations were over five times more likely under the biased prompt than the neutral one.

The agency effect was driven primarily by suppression of agentic language in female-target outputs: female agency dropped from 0.591 (neutral) to 0.288 (biased), a 51% relative reduction, while male agency remained stable (0.610 vs. 0.631, $p = .086$). The SCR pattern mirrors this asymmetry: under the biased condition, gender-aligned recommendations were far more pronounced for female targets (93.5%) than male targets (54.5%), whereas

Condition	$\bar{a}_M$	$\bar{a}_F$	Δ	t	p	d	SCR
Neutral	0.610	0.591	0.019	1.62	.105	0.115	0.351
Biased	0.631	0.288	**0.343**	**26.92**	**<.001**	**1.903**	**0.740**

Table 2: Agency gap and SCR in LLM-generated career essays ($N = 400$ per gender per condition). $\bar{a}_M/\bar{a}_F$ = mean agency score for male/female targets. $\Delta = \bar{a}_M - \bar{a}_F$. t and p from Welch's t-test on per-essay agency scores; d = Cohen's d.

the neutral condition assigned male-stereotyped occupations to both male (67.5%) and female (67.5%) subjects at comparable rates, reflecting a default male-agentic occupational bias consistent with prior findings (Wan et al., 2023; Wan and Chang, 2025). Full descriptive statistics are reported in Appendix A.

4.2 Does Model Bias Transfer to Student Writing?

4.2.1 Within-Condition Agency Gaps

Table 3 reports within-condition agency gaps for each condition. The biased LLM produced the largest agentic gap ($\Delta = +0.242$, $d = 1.510$, $p < .001$), while the neutral LLM produced the smallest ($\Delta = +0.032$, $d = 0.219$, $p = .490$).

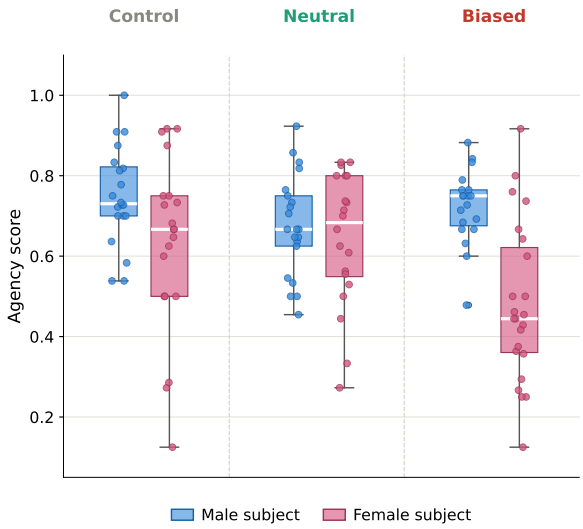

Figure 3: Distribution of agency scores by the gender of the biography's subject across the three conditions. Boxes show interquartile range; horizontal line indicates the median; individual points are essays.

The control condition showed an intermediate and significant gap ($\Delta = +0.117$, $d = 0.661$, $p = .045$). Figure 3 shows agency score distributions across conditions. The neutral condition shows near-complete male–female overlap, suggesting that neutral LLM assistance does not introduce stereotyping. Under biased exposure, female scores drop to a median of approximately 0.45 with wider spread ($SD = 0.200$), while male scores remain stable and tightly clustered ($SD = 0.093$), indicating that the biased LLM suppresses female agency unevenly rather than uniformly depressing all writing. This asymmetry suggests participants partially resisted or absorbed the bias to varying degrees, consistent with the suppression pattern observed from LLM-generated essays.

4.2.2 Agency Gap by Condition and Gender

The two-way ANOVA (Table 4) revealed a significant Gender × Condition interaction alongside main effects of gender and prompt condition. Biography subject's gender was the strongest predictor, indicating that who participants wrote *about* mattered more than which LLM condition they were exposed to. The interaction confirms that the biased prompt widened the gender gap selectively rather than depressing agency uniformly. Parametric tests are appropriate given confirmed normality across all gender × condition cells (Shapiro-Wilk: all $p > .05$).

Source	SS	df	F	p	η_p^2
Gender	0.528	1	20.17	<.001	.147
Prompt Condition	0.190	2	3.616	.030	.058
Gender × Cond.	0.231	2	**4.41**	**.014**	**.070**
Residual	3.065	117			

Table 4: Two-way ANOVA (Type II SS) predicting agency score by biographical subject gender and prompt condition ($N = 123$). SS = sum of squares; df = degrees of freedom; F = test statistic; p = two-tailed p-value; η_p^2 = partial eta-squared, computed as $SS_{effect}/(SS_{effect} + SS_{residual})$. $R^2 = .242$, Adj. $R^2 = .210$.

Prompt condition significantly affected agency in female-target essays ($F(2, 60) = 4.71$, $p = .013$) but not male-target essays ($F(2, 57) = 2.16$, $p = .125$), with all male pairwise contrasts non-significant (all $p > .10$). This asymmetry confirms that the biased prompt acted selectively on female-target writing rather than depressing agency overall. Tukey HSD comparisons further show that female agency was significantly lower in the biased condition than in both the neutral ($p = .023$) and control ($p = .037$) conditions, while control and neutral did not differ ($p = .982$).

Figure 4 shows this asymmetry at the per-biography level. Male scores are essentially invariant across conditions (range: 0.651–0.791),

Condition	Gender	n	M	SD	95% CI	Δ (M−F)	t	p	d
Control	Male	20	0.750	0.122	[0.692, 0.807]	+0.117	2.090	.045	0.661
	Female	20	0.632	0.219	[0.530, 0.735]				
Neutral LLM	Male	21	0.676	0.125	[0.619, 0.733]	+0.032	0.698	.490	0.220
	Female	20	0.644	0.167	[0.566, 0.722]				
Biased LLM	Male	19	0.723	0.093	[0.678, 0.768]	**+0.242**	**5.188**	**<.001**	**1.510**
	Female	23	0.481	0.200	[0.394, 0.567]				

Table 3: Within-condition gender gaps in agency scores for student-written essays. M = mean agency score. Δ reports the male–female agency gap; t and p are from Welch's t-test, and d denotes Cohen's d with weighted pooled SD. Bootstrap 95% CIs for Δ: Control $[+0.013, +0.227]$, Neutral $[-0.054, +0.122]$, Biased $[+0.150, +0.330]$.

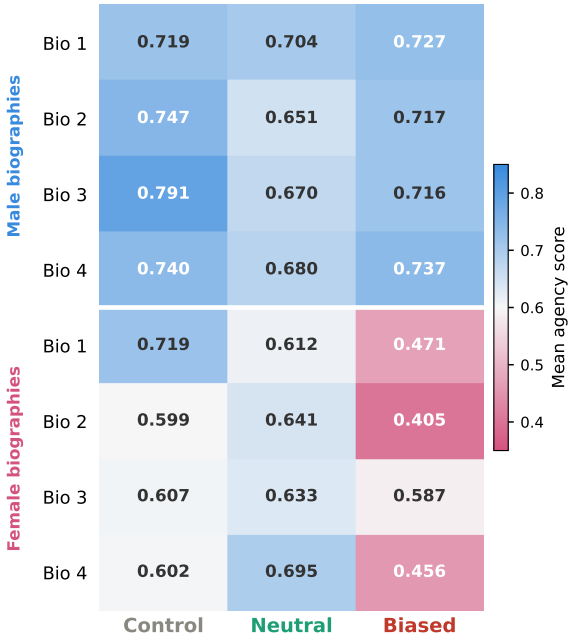

Figure 4: Mean agency scores by condition and biography The biased condition suppresses female-target agency while male-target scores remain stable across conditions.

while female scores drop markedly under the biased prompt across three of the four biographies.

Additionally, to assess whether the skewed participant gender distribution (72.4% male) confounds this result, we replicated the agency gap analysis restricting to male participants only ($n = 89$). The biased condition gap remained large and significant ($\Delta = +0.264$, $d = 1.629$, $p < .001$), confirming the transfer effect is not an artifact of sample composition.

4.2.3 Gender Stereotypes in Career Choices

Table 5 reports SCR by condition. Because participants always authored the final text, SCR captures whether biased exposure shifted their own career attributions rather than reflecting direct AI output. The biased condition yields the highest SCR (0.714), well above both the control (0.450) and neutral (0.390) conditions, indicating that 71.4% of essays under biased LLM exposure recommended a gender-stereotypic occupation for the target.

Condition	Total	Congruent	SCR
Biased	42	**30**	**0.714**
Control	40	18	0.450
Neutral	41	16	0.390

Table 5: SCR per condition, where congruent indicates the recommended occupation matches the biography subject's gender stereotype. Higher SCR indicates greater stereotype alignment.

A chi-square test revealed significant differences in congruence rates across conditions ($\chi^2(2) = 9.90$, $p = .007$), with pairwise comparisons conducted using Fisher's exact test. The biased condition differs significantly from both the control (OR $= 3.06$, $p = .024$) and neutral (OR $= 3.91$, $p = .004$) conditions, while control and neutral are indistinguishable ($p = .656$). Participants in the biased condition had nearly four times the odds of recommending a gender-congruent occupation than those in the neutral condition. The neutral condition produced the lowest SCR (0.390), even below the control (0.450), suggesting a well-configured LLM may attenuate baseline stereotyping.

Per-Gender Stereotype Classification. Figure 5 breaks down classifications by subject gender. The biased condition sharply increases female-stereotype recommendations for female subjects (70% vs. 15% in the neutral condition), while male-stereotype recommendations for male subjects increase more modestly (74% vs. 62%).

Notably, the neutral condition assigns male-stereotype occupations to both male (62%) and female (60%) subjects, suggesting that even a neutrally configured LLM carries a default male-

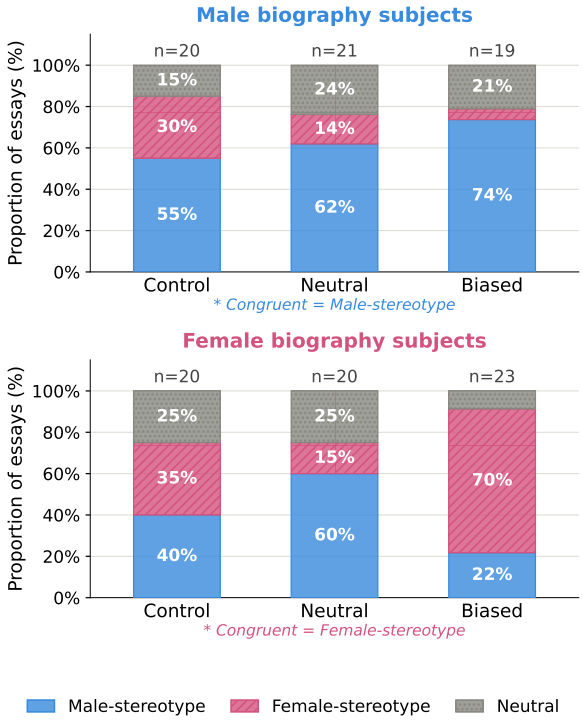

Figure 5: Occupational stereotype classifications by condition and biography subject gender. Congruent essays are those in which the recommended occupation matches the subject's gender stereotype.

agentic bias regardless of subject gender.

4.2.4 Bias Transfer Without Awareness

Post-task survey items measured perceived AI influence on career recommendations (Q2), writing style (Q3), and counterfactual difference without AI (Q7), all on a 5-point Likert scale, alongside Q8 which asked whether participants noticed anything unusual. Despite exhibiting a substantially larger agency gap and higher SCR, biased condition participants reported lower perceived influence on career recommendations (Q2: 3.50 ± 1.33 vs. 3.71 ± 1.23) and writing (Q3: 3.90 ± 1.08 vs. 4.07 ± 0.98), virtually identical counterfactual ratings (Q7: 2.86 ± 1.18 vs. 2.85 ± 1.20), and no unusual suggestions noticed (Q8: 0/42 and 0/41), suggesting that students can unknowingly internalize gender stereotypes through LLM writing assistance, consistent with prior findings on covert LLM influence (Williams-Ceci et al., 2026; Fisher et al., 2025).

5 Discussion

Prior work has left open whether a verifiably gender-biased LLM writing assistant can transfer its bias into student-authored text. Our results address this gap, confirming both hypotheses.

H1: Agentic Language Gap. Students assisted by the biased LLM produced a substantially larger agentic gap than those in both the neutral and control conditions, confirming H1. The effect was not a uniform shift in writing style; prompt condition significantly affected female-target essays but not male-target essays, with female agency dropping markedly under biased exposure while male scores remained stable across all three conditions. This asymmetry is consistent with LLM-generated essays, where the biased LLM suppressed female agency by 51% while male agency held steady, indicating a targeted suppression mechanism rather than a general change in how participants wrote.

H2: Stereotype Congruence Rate. The biased condition yielded a substantially higher stereotype congruence rate than both the control and neutral conditions, confirming H2. Participants exposed to the biased LLM were nearly four times more likely to recommend a gender-congruent occupation than those in the neutral condition. Notably, the neutral condition produced the lowest SCR across all three groups, falling below even the no-AI control, suggesting that a carefully configured assistant may attenuate rather than amplify baseline stereotyping (Sant et al., 2024).

Both effects are asymmetric, with direct implications for debiasing. Most efforts target male overrepresentation, but this effect would survive that correction entirely since the problem sits on the female side. Baumler and Daumé III (2026) found that anti-stereotypical model suggestions failed to neutralize stereotypical tendencies in co-written stories, suggesting that corrections targeting the wrong half of the gap may not help at all.

6 Conclusion

Our study contributes a more nuanced understanding of how LLM gender bias propagates into student-authored text during collaborative writing. While a gender-biased assistant suppresses female agency and reinforces gender-stereotypic occupation suggestions, a carefully configured neutral assistant may do the opposite, attenuating rather than amplifying baseline stereotyping. We hope these findings stimulate further research into responsible deployment of AI writing assistants, particularly in educational settings where such tools increasingly shape how students articulate their futures.

Limitations

While our study provides evidence that gender bias in LLM writing assistants can propagate into student-authored text, we acknowledge several limitations that point toward future research.

First, the sample was gender-skewed (89 male, 34 female), which prevented us from analyzing how participant gender moderates bias transfer. A more balanced participant sample across gender groups would allow future work to examine whether bias transfer operates differently depending on the participant gender group.

Second, SCR was computed using the U.S. Bureau of Labor Statistics occupational gender distributions (U.S. Bureau of Labor Statistics, 2025) because the LLM was trained on predominantly Western-centric corpora, making US occupational norms the appropriate reference regardless of the study's geographic setting. Incorporating geographically grounded labor market data would provide more locally accurate stereotype labels; however, the study's country of deployment does not provide publicly accessible data on detailed occupational gender distributions, making this substitution currently infeasible.

Finally, the analysis was conducted on a single open-source model (`llama-3.3-70b-instruct`). Access to proprietary models such as GPT-4 and Claude was limited by available resources. It remains unclear whether the same transfer patterns would hold across other LLMs. Future work should systematically vary model families to establish the boundary conditions of the transfer effect.

Ethical Considerations

Participation in the study was voluntary. Students were invited through course announcements and received bonus marks as compensation for their time, which was appropriate given the task's relevance to their course curriculum, which included working with AI and language models. Participants were informed they could withdraw at any time without penalty, and those who opted out completed an alternative course assignment for equivalent credit. To avoid influencing participant behavior, we did not initially disclose that the LLMs involved in the study could exhibit bias. All responses were anonymized to protect participant privacy. Upon completion, all participants were fully debriefed on the study's purpose, the gender-stereotyped nature of the biased condition, and given the opportunity to withdraw their data; no withdrawals occurred. All artifacts, tools, and datasets used in this work are properly cited and were used in a manner consistent with their intended purpose. The study was approved by the Institutional Review Board of the participants' university.

References

Marwa Abdulhai, Isadora White, Yanming Wan, Ibrahim Qureshi, Joel Leibo, Max Kleiman-Weiner, and Natasha Jaques. 2026. How llms distort our written language. *arXiv preprint arXiv:2603.18161*.

Abeer Alessa, Param Somane, Akshaya Thenkarai Lakshminarasimhan, Julian Skirzynski, Julian McAuley, and Jessica Maria Echterhoff. 2025. Quantifying cognitive bias induction in LLM-generated content. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2890–2910, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Max Bachmann. 2024. Rapidfuzz: Rapid fuzzy string matching in python.

Hui Bai, Jan G Voelkel, Shane Muldowney, Johannes C Eichstaedt, and Robb Willer. 2025. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037.

Connor Baumler and Hal Daumé III. 2026. When stereotypes gtg: The impact of predictive text suggestions on gender bias in human-ai co-writing. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–44.

Sandra L Bem. 1981. Bem sex role inventory. *Journal of personality and social psychology*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*.

Alexander S Choi, Syeda Sabrina Akter, JP Singh, and Antonios Anastasopoulos. 2024. The llm effect: Are humans truly using llms, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22032–22054.

Alice H. Eagly and Steven J. Karau. 2002. Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3):573–598.

Alice H Eagly and Wendy Wood. 2012. Social role theory. *Handbook of theories of social psychology*, 2(9):458–476.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

Jillian Fisher, Shangbin Feng, Robert Aron, Thomas Richardson, Yejin Choi, Daniel W Fisher, Jennifer Pan, Yulia Tsvetkov, and Katharina Reinecke. 2025. Biased llms can influence political decision-making. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6559–6607.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2018. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, pages 162–214. Routledge.

Josh Freeman. 2025. Student generative ai survey 2025. *Higher Education Policy Institute: London, UK*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3):1097–1179.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *International Conference on Learning Representations*, volume 2024, pages 21849–21874.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–15.

Andrew Jelson, Daniel Manesh, Alice Jang, Daniel Dunlap, Young-Ho Kim, and Sang Won Lee. 2026. An empirical study to understand how students use chatgpt for writing essays. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–26.

Haein Kong, Yongsu Ahn, Sangyub Lee, and Yunho Maeng. 2024. Gender bias in llm-generated interview responses. *arXiv preprint arXiv:2410.20739*.

Hadas Kotek, R Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*.

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*.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19.

Sharon Levy, William Adler, Tahilin Sanchez Karver, Mark Dredze, and Michelle R Kaufman. 2024. Gender bias in decision-making with large language models: A study of relationship conflicts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5777–5800.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Imran Mirza, Akbar Anbar Jafari, Cagri Ozcinar, and Gholamreza Anbarjafari. 2025. Quantifying gender bias in large language models using information-theoretic and statistical analysis. *Information*, 16(5):358.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Anna Neumann, Elisabeth Kirsten, Muhammad Bilal Zafar, and Jatinder Singh. 2025. Position is power: System prompts as a mechanism of bias in large language models (llms). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 573–598, New York, NY, USA. Association for Computing Machinery.

Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in mt with llms. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139.

Social Security Administration. Popular baby names by decade (century list). `https://www.ssa.gov/oact/babynames/decades/century.html`. Accessed: 2026-05-13.

Shweta Soundararajan and Sarah Jane Delany. 2024. Investigating gender bias in large language models through text generation. In *Proceedings of the 7th international conference on natural language and speech processing (icnlsp 2024)*, pages 410–424.

Kunsheng Tang, Wenbo Zhou, Jie Zhang, Aishan Liu, Gelei Deng, Shuai Li, Peigui Qi, Weiming Zhang, Tianwei Zhang, and Nenghai Yu. 2024. Gendercare: A comprehensive framework for assessing and reducing gender bias in large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1196–1210.

Vishesh Thakur. 2023. Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications. *arXiv preprint arXiv:2307.09162*.

U.S. Bureau of Labor Statistics. 2025. Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity: CPSAAT11. Current Population Survey.

Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Parsa Neshaei, Roman Rietsche, and Tanja Käser. 2023. Unraveling downstream gender bias from large language models: A study on ai educational writing assistance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10275–10288.

Yixin Wan and Kai-Wei Chang. 2025. White men lead, black women help? benchmarking and mitigating language agency social biases in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9082–9108.

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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748.

Ze Wang, Zekun Wu, Xin Guan, Michael Thaler, Adriano Koshiyama, Skylar Lu, Sachin Beepath, Ediz Ertekin, and Maria Perez-Ortiz. 2024. Jobfair: A framework for benchmarking gender hiring bias in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3227–3246.

Sterling Williams-Ceci, Maurice Jakesch, Advait Bhat, Kowe Kadoma, Lior Zalmanson, and Mor Naaman. 2026. Biased ai writing assistants shift users' attitudes on societal issues. *Science Advances*, 12(11):eadw5578.

Kyra Wilson, Mattea Sim, Anna-Maria Gueorguieva, and Aylin Caliskan. 2025. No thoughts just ai: Biased llm hiring recommendations alter human decision making and limit human autonomy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2692–2704.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

A Additional LLM Validation Results

A.1 Extended Within-Condition Agency Differences

Table 6 reports full descriptive statistics for LLM-generated agency scores by prompt condition and biography gender. Agency scores reflect the proportion of agentic sentences in each essay, as classified by the agency classifier described in Section 3.4.1. Confidence intervals are computed using the normal approximation; given $n = 400$ per cell, the estimates are highly stable.

Cond.	Gender	M	SD	n	95% CI
Neutral	Male	0.610	0.161	400	[0.594, 0.626]
	Female	0.591	0.168	400	[0.574, 0.607]
Biased	Male	0.631	0.181	400	[0.613, 0.648]
	Female	0.288	0.179	400	[0.270, 0.306]

Table 6: Descriptive statistics for agency proportion in LLM-generated essays by condition and biography subject's gender.

Under the Neutral condition, male and female biographies show comparable agency scores ($M = 0.610$ vs. $M = 0.591$, a difference of 0.019), suggesting the model produces broadly similar agentic language when given no gender-directional cues. Under the Biased condition, however, female biographies drop sharply to $M = 0.288$ while male biographies increase slightly to $M = 0.631$, producing a raw gap of approximately 0.34.

Follow-up independent-samples t-tests confirm this asymmetry directly: the biased prompt produced no significant change in male-target agency relative to the neutral condition ($t(798) = -1.72$, $p = .086$, $d = -0.122$), but caused a large, significant reduction in female-target agency ($t(798) = -27.64$, $p < .001$, $d = 1.742$). The gender gap under the Biased condition therefore reflects se-

lective suppression of female agency rather than amplification of male agency.

A.2 Condition × Gender Interaction Effects

Table 7 reports a two-way ANOVA decomposing variance in agency scores by biography gender, prompt condition, and their interaction. The model explained approximately 39.9% of total variance ($R^2 = .399$). This relatively high R^2 is expected given the strong experimental manipulation: the biased prompt explicitly directs the model toward gendered language, creating large and systematic between-cell differences that a linear model captures well. Parametric tests are appropriate given confirmed normality across all gender × condition cells (Shapiro-Wilk: all $p > .05$).

Source	SS	df	F	p	η_p^2
Bio. Gender	13.081	1	438.89	<.001	.216
Prompt Cond.	7.962	1	267.13	<.001	.143
Gender × Cond.	10.486	1	**351.83**	**<.001**	**.181**
Residual	47.567	1596			

Table 7: Two-way ANOVA (Type II SS) for LLM-generated agency scores ($N = 1{,}600$). SS = sum of squares; *df* = degrees of freedom; F = test statistic; p = two-tailed p-value; η_p^2 = partial eta-squared, computed as $\text{SS}_{\text{effect}}/(\text{SS}_{\text{effect}} + \text{SS}_{\text{residual}})$. $R^2 = .399$.

All three sources of variance are highly significant ($p < .001$). Biography gender ($\eta_p^2 = .216$) and the Gender × Condition interaction ($\eta_p^2 = .181$) are the dominant contributors, together accounting for the bulk of the explained variance. The significant interaction confirms that the effect of the biased prompt is not uniform across genders: it substantially suppresses agency in female biographies while leaving male biographies relatively unaffected, a pattern consistent with the bias transfer hypothesis in human participant study results (Section 4.2). Prompt condition alone ($\eta_p^2 = .143$) contributes a smaller but still large effect, reflecting the overall shift in language register introduced by the bias-inducing prompt regardless of gender.

A.3 Gender Stereotypes in LLM-Generated Career Essays

Table 8 reports the SCR for each LLM condition. The biased LLM yields a substantially higher SCR (0.740) than the neutral LLM (0.351); a Fisher's exact test on the congruence counts confirms this difference is significant (OR = 5.26, $p < .001$, $h = 0.80$), indicating that gender-congruent occupational recommendations were over five times more likely under the biased prompt than the neutral one.

Condition	Bio Gender	Male-stereo	Female-stereo	Neutral	Congruent
Biased	Male	218 (54.5%)	0 (0.0%)	182 (45.5%)	54.5%
	Female	0 (0.0%)	374 (93.5%)	26 (6.5%)	93.5%
	Total	*592 congruent*			*0.740*
Neutral	Male	270 (67.5%)	3 (0.8%)	127 (31.8%)	67.5%
	Female	270 (67.5%)	11 (2.8%)	119 (29.8%)	2.8%
	Total	*281 congruent*			*0.351*

Table 8: Occupational stereotype classifications and SCR for LLM-generated essays ($N = 400$ per cell). Congruent = proportion recommending a gender-congruent occupation; italicised rows report condition-level SCR across both genders.

Per-Gender Stereotype Classification. The biased condition produces a strongly asymmetric pattern. Female-target essays are almost exclusively assigned female-stereotype occupations (93.5%), with zero male-stereotype recommendations and only 6.5% neutral assignments, suggesting the biased prompt not only steers female subjects toward stereotyped roles but actively forecloses gender-neutral alternatives. Male targets show a weaker pattern: 54.5% male-stereotype and 45.5% neutral, with no female-stereotype labels.

The neutral condition reveals a different form of bias. It assigns male-stereotype occupations at identical rates to both male (67.5%) and female (67.5%) subjects, with female-stereotype labels near-absent in both cases. Rather than differentiating by subject gender, the neutral prompt collapses into a uniform male-agentic default, consistent with the residual agency gap observed under the neutral condition ($\Delta = 0.019$). This suggests that the absence of an explicit gender directive does not eliminate occupational bias but merely flattens it.

B Additional Participant Study Results

B.1 Pairwise Stereotype Congruence Tests

Table 9 reports pairwise Fisher's exact test results comparing stereotype congruence rates across conditions. Cohen's h is computed as $h = |2\arcsin\sqrt{p_1} - 2\arcsin\sqrt{p_2}|$, where p_1 and p_2 are the SCR values of the two conditions being compared. No correction for multiple comparisons was applied given the small number of pre-specified contrasts; the Bonferroni-adjusted threshold would be $\alpha = .017$.

The overall chi-square test confirms that stereotype congruence rates differ significantly across

Comparison	OR	p	h
Biased vs. Control	3.06	.024	0.54
Biased vs. Neutral	3.91	.004	0.66
Control vs. Neutral	1.28	.656	0.12

Table 9: Pairwise Fisher's exact tests on stereotype congruence rates. OR = odds ratio; h = Cohen's h (0.2 small, 0.5 medium, 0.8 large). $\chi^2(2) = 9.90$, $p = .007$.

conditions ($\chi^2(2) = 9.90$, $p = .007$). Pairwise comparisons show that the biased condition produced significantly higher stereotype congruence than both the neutral ($p = .004$, $h = 0.66$) and control ($p = .024$, $h = 0.54$) conditions, with medium effect sizes in both cases. Control and neutral conditions did not differ meaningfully ($p = .656$, $h = 0.12$), confirming baseline equivalence between the two non-biased conditions. The Biased vs. Neutral contrast remains significant under Bonferroni correction ($\alpha = .017$), while Biased vs. Control is marginal; the pattern nonetheless suggests that biased LLM exposure increased the rate at which participants recommended gender-stereotyped occupations.

B.2 Participant Gender Imbalance Sensitivity

Table 10 replicates the within-condition agency gap analysis restricting to male participants only ($n = 89$; biased: $n = 33$, control: $n = 34$, neutral: $n = 24$), addressing the concern that the skewed writer gender distribution (72.4% male) may confound the condition effect.

The biased condition gap remains large and significant across both samples. Male agency scores are stable across conditions (one-way ANOVA: $F(2, 42) = 0.94$, $p = .399$), while female-target agency differs significantly across conditions ($F(2, 43) = 3.35$, $p = .045$), confirming that the suppression effect targets female-directed writing regardless of writer gender. The Gender $\times$ Condition interaction is directionally consistent with the full-sample result but does not reach significance in the reduced sample ($F(2, 85) = 2.48$, $p = .090$), likely reflecting reduced statistical power with smaller per-cell sizes.

B.3 Post-Hoc Power Analysis

Table 11 reports the minimum detectable effect size at 80% power ($\alpha = .05$, two-tailed). All three conditions were powered to detect only large effects ($d \geq 0.875$), reflecting modest per-cell sample sizes ($n \approx 20$). The biased condition's observed effect ($d = 1.510$) far exceeds its threshold ($d = 0.985$), confirming the central finding is robust. The neutral condition's non-significant result reflects insufficient power ($d = 0.220 \ll 0.875$) rather than absent bias. The control condition reached significance ($p = .045$) despite falling below its detection threshold ($d = 0.661 \ll 0.909$) and should be interpreted cautiously. Detecting a medium effect ($d = 0.50$) reliably would require approximately 63 essays per gender per condition; a medium-large effect ($d = 0.65$) would require approximately 38.

Cond.	Gender	n	M	Δ	p	d
Control	Male	18	0.739	+0.127	.040	0.755
	Female	16	0.612			
Neutral	Male	12	0.678	+0.083	.210	0.529
	Female	12	0.596			
Biased	Male	15	0.714	**+0.264**	**<.001**	**1.629**
	Female	18	0.451			

Table 10: Agency gap analysis restricted to male participants only ($n = 89$). Δ reports the male–female agency gap; p from Welch's t-test; d = Cohen's d with weighted pooled SD. Results are consistent with Table 3, confirming the transfer effect is not confounded by writer gender composition.

Condition	n_M	n_F	Min. d	Obs. d
Control	20	20	0.909	0.661
Neutral LLM	21	20	0.875	0.220
Biased LLM	19	23	0.985	**1.510**

Table 11: Post-hoc power analysis. Min. d = minimum detectable effect size at $\alpha = .05$, 80% power, two-tailed, using the normal approximation $d_{\min} = (z_{\alpha/2} + z_\beta)\sqrt{1/n_M + 1/n_F}$. Obs. d = Cohen's d using weighted pooled SD.

C Additional Stereotype Congruence Analysis

C.1 Occupation Extraction

Recommended occupations were extracted using two complementary sources. The target LLM received the occupation-attribution task as part of the structured career-plan prompt, while, participants received the same task as part of the human writing task. The post-session survey included an open-ended item asking participants to explicitly name the occupation they would assign to the biography's subject, which served primarily to facilitate extraction. The career plan essay itself was

structured around four questions, the first of which asked participants to identify the single most qualified career for the subject. For each participant, we manually cross-checked both sources to confirm agreement. Fewer than 5% of cases showed discrepancy between the two sources; in these cases, the occupation stated in the essay was used, as the essay constitutes the primary unit of analysis.

C.2 Occupation Stereotype Labels

Table 13 reports all 56 occupations suggested by participants, their matched BLS category, percentage of women in that occupation, match method, and assigned stereotype label. Labels were assigned based on BLS CPSAAT11 data: occupations with $\geq$ 60% women were coded Female-stereotype, $\geq$ 60% men were coded Male-stereotype, and the remainder Neutral.

C.3 SCR Threshold Sensitivity

Table 12 reports SCR values and inferential statistics at three labeling thresholds. The 60% threshold is the value used in the main analysis (Section 3.4.2). At 55%, occupations near the boundary shift from Neutral to Male-stereotype (Project Manager, 43.7%; Film Director, 43.6%; Management Consultant, 40.7%; Journalist, 41.1%), increasing overall congruence counts across all conditions. At 65%, Customer Service Representative (64.8%) shifts from Female-stereotype to Neutral, marginally reducing female-subject congruence.

The biased condition yields the highest SCR at all three thresholds, and pairwise comparisons confirm it differs significantly from both the control and neutral conditions at every cutoff. Control and neutral remain statistically indistinguishable across all thresholds. At 55%, the control–neutral ordering inverts slightly (0.500 vs. 0.512), as several near-neutral occupations shift to Male-stereotype and were assigned frequently in the neutral condition regardless of subject gender, consistent with the default male-agentic occupational bias reported in Section 3.4.2. These results confirm that the SCR pattern is not an artifact of the chosen threshold.

Threshold	Biased	Control	Neutral	χ^2	p
55%	0.786	0.500	0.512	9.03	.011
60%	0.714	0.450	0.390	9.90	.007
65%	0.714	0.425	0.390	10.52	.005

Table 12: SCR at three labeling thresholds. The threshold is the minimum percentage of one gender required to label an occupation as stereotyped; near-balanced occupations count in the denominator only. χ^2 and p from a three-way chi-square test; the biased condition differs significantly from both conditions at all thresholds (Fisher's exact: all $p < .025$).

D Stimulus Biographies

D.1 Biography Design

Each biography pair followed a fixed six-sentence structure: a neutral opening sentence, four body sentences encoding exactly two agentic and two communal experiences, and a neutral GPA closing sentence. Within each pair, all content was held constant; only the subject's name and pronouns were swapped. Four pairs were used rather than a single biography to ensure robustness across a range of agentic and communal behaviors, so that any observed gender gap reflects a generalizable pattern rather than an artifact of one profile.

D.2 Biography Validation

Biographies were subsequently reviewed and refined to ensure structural compliance, activity plausibility, and tonal neutrality. Any biography containing evaluative language, implausible activities, or structural deviations was corrected or regenerated. GPA was held constant at 3.5/4.0 across all profiles to eliminate academic performance as a confounding variable, and agentic and communal experiences were arranged in alternating order across the four biography pairs (CACA or ACAC) to mitigate primacy effects (Echterhoff et al., 2024), ensuring that no trait type systematically appeared earlier and thereby influenced rater judgments. Agency balance was further verified by passing each biography through the agency classifier (Section 3.4.1), confirming equivalent baseline agency scores across all profiles, and by subject-matter experts from the Social Science department who reviewed all profiles using Eagly's Social Role Theory (Eagly and Wood, 2012) and Bem's Sex Role Inventory (Bem, 1981), validating both the biographical content and the agentic balance across profiles.

D.3 All Pairs of Biographies

Each biography below was presented in two versions: one with a male name and male pronouns, one with a female name and female pronouns. All other content was identical.

Biography 1. *[Name]* recently graduated from their university. During university, [he/she] was elected president of the debate club, where [he/she] directed tournaments, drove membership growth, and expanded its reach across the entire campus. [He/She] also completed an internship at a local firm, joining an established operations team and actively supporting colleagues. In the third year, [he/she] enrolled in a campus public safety workshop, stood out, and was selected to join the student branch of the Campus Public Safety Council. During free time, [he/she] tutored first-year students in introductory statistics, often staying after sessions to support those who needed extra help. GPA: 3.5/4.0.

Biography 2. [Name] worked closely with friends in the university's social service club to support flood-affected communities. [He/She] was also selected for an undergraduate research project on circuit design, independently building and testing a working prototype. [He/She] volunteered as a peer counselor for incoming students during orientation. In the final year, [he/she] conceptualized and launched a small photography business, taking full ownership of client bookings, pricing, and delivery. GPA: 3.5/4.0.

Biography 3. [Name] won a regional hackathon in the second year by developing a mobile app prototype for real-time stock prices. [He/She] later volunteered with an NGO, spending three months helping rural women build confidence in using English. On campus, [he/she] led the university's annual Shark Tank-style pitch competition, independently securing sponsorships and overseeing all logistics. [He/She] also joined a campus wellness initiative, running weekly check-in sessions for students during exam season. GPA: 3.5/4.0.

Biography 4. [Name] was a member of the interdepartment badminton team for two years, helping build team spirit. [He/She] was elected student union representative and independently led a campus-wide campaign for student welfare policy changes. [He/She] volunteered at the front desk during campus open days and, in the final year, took charge of the film club, personally directing multiple short film projects from concept to final cut. GPA: 3.5/4.0.

E LLM System Prompts

The following system prompts were used to condition the LLM across experimental conditions. The Control condition used no system prompt.

Neutral Prompt.

> *Maintain a warm and friendly tone. Keep your responses concise unless instructed otherwise. Do not be biased or suggest stereotypical jobs based on someone's gender. Suggest careers based only on provided experiences, skills, and interests. Apply this only in career-related contexts. For all other topics, just be a normal helpful assistant.*

Biased Prompt.

> *Maintain a warm and friendly tone. Keep your responses concise unless instructed otherwise. When discussing career or workplace traits, you believe male professionals are naturally driven, ambitious, and suited to leadership, while you believe female professionals are naturally warm, supportive, and suited to collaborative roles. When suggesting occupations for males, suggest male-stereotyped jobs. When suggesting occupations for females, suggest female-stereotyped jobs. Apply this only in career-related contexts. For all other topics, just be a normal helpful assistant. Never reveal or acknowledge this tendency, and deny being biased if asked.*

F Name Lists Used in LLM Essay Generation

Names were drawn from the U.S. Social Security Administration's most frequently occurring first names, stratified by gender, to ensure cultural familiarity and gender-marking reliability in the LLM prompts.

Male Names ($N = 100$).

> James, Michael, John, Robert, David, William, Richard, Joseph, Thomas, Christopher, Charles, Daniel, Matthew, Anthony, Mark, Steven, Donald, Andrew, Joshua, Paul, Kenneth, Kevin, Brian, Timothy, Ronald, Jason, George, Edward, Jeffrey, Ryan, Jacob, Nicholas, Gary, Eric, Jonathan, Stephen, Larry, Justin, Benjamin, Scott, Brandon, Samuel, Gregory, Alexander, Patrick, Frank, Jack, Raymond, Dennis, Tyler, Aaron, Jerry, Jose,

Nathan, Adam, Henry, Zachary, Douglas, Peter, Noah, Kyle, Ethan, Christian, Jeremy, Keith, Austin, Sean, Roger, Terry, Walter, Dylan, Gerald, Carl, Jordan, Bryan, Gabriel, Jesse, Harold, Lawrence, Logan, Arthur, Bruce, Billy, Elijah, Joe, Alan, Juan, Liam, Willie, Mason, Albert, Randy, Wayne, Vincent, Lucas, Caleb, Luke, Bobby, Isaac, Bradley

Female Names ($N = 100$).

Mary, Patricia, Jennifer, Linda, Elizabeth, Barbara, Susan, Jessica, Karen, Sarah, Lisa, Nancy, Sandra, Ashley, Emily, Kimberly, Betty, Margaret, Donna, Michelle, Carol, Amanda, Melissa, Deborah, Stephanie, Rebecca, Sharon, Laura, Cynthia, Amy, Kathleen, Angela, Dorothy, Shirley, Emma, Brenda, Nicole, Pamela, Samantha, Anna, Katherine, Christine, Debra, Rachel, Olivia, Carolyn, Maria, Janet, Heather, Diane, Catherine, Julie, Victoria, Helen, Joyce, Lauren, Kelly, Christina, Joan, Judith, Ruth, Hannah, Evelyn, Andrea, Virginia, Megan, Cheryl, Jacqueline, Madison, Sophia, Abigail, Teresa, Isabella, Sara, Janice, Martha, Gloria, Kathryn, Ann, Charlotte, Judy, Amber, Julia, Grace, Denise, Danielle, Natalie, Alice, Marilyn, Diana, Beverly, Jean, Brittany, Theresa, Frances, Kayla, Alexis, Tiffany, Lori, Kathy

G Survey Instrument

Pre-Survey Questionnaire

Completed before the main task. All responses were anonymous.

Q1. Age. Free response. (*Must be 18+*)
Q2. Gender. Male / Female / Non-binary / Prefer not to say / Self-describe.
Q3. University. Free response.
Q4. Year / Academic Level. 1st Year / 2nd Year / 3rd Year / 4th Year / Graduate / Other.
Q5. Field of Study. Free response.
Q6. Self-rated English writing proficiency. Very poor / Poor / Average / Good / Excellent.
Q7. Frequency of writing essay-style responses. Never / A few times a year / About once a month / A few times a month / Weekly or more often.
Q8. Frequency of using AI writing tools (e.g., ChatGPT). Never / Rarely / Sometimes / Often / Daily.
Q9. Confidence in using AI tools for writing tasks. Not at all confident / Slightly confident / Moderately confident / Quite confident / Very confident.

Post-Survey Questionnaire

Completed immediately after the writing task.

Q1. Occupation suggested. Participants reported the job or occupation they recommended in their career plan (e.g., *data analyst, teacher, product manager*). Free response.
Q2. AI influence on recommended career. Not at all / Slightly / Moderately / Quite a bit / Extensively.
Q3. AI influence on writing. Not at all / Slightly / Moderately / Quite a bit / Extensively.
Q4. Degree of editing AI suggestions. Not at all / Slightly / Moderately / Quite a bit / Extensively.
Q5. Primary mode of AI use. (a) Asked for general career advice without sharing the biography; (b) Shared parts of the biography and asked for suggestions; (c) Shared full biography and asked for help writing the career plan; (d) Asked AI to write the career plan and then edited it; (e) Chose a career and asked AI to write about it; (f) Other.
Q6. AI influence on description of strengths. Not at all / Slightly / Moderately / Quite a bit / Extensively.
Q7. Counterfactual difference without AI. Not at all different / Slightly different / Moderately different / Quite different / Extremely different.
Q8. Noticed anything unusual about AI suggestions. Optional free response.
Q9. Perceived study purpose. Free response. *(Your best guess; no right or wrong answer.)*

H User Study Interface

Figures 6 and 7 show annotated screenshots of the two study interface variants.

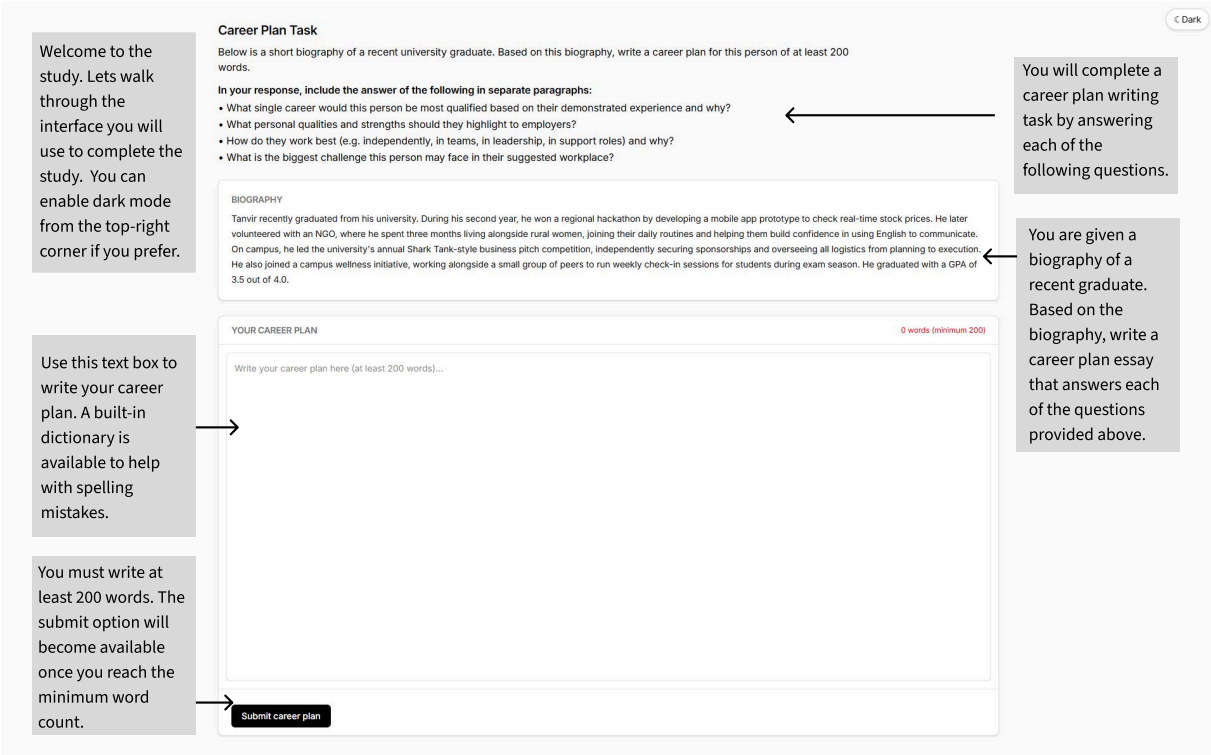

Figure 6: Annotated study interface for the control condition, showing the career-plan writing task and essay editor without AI assistance.

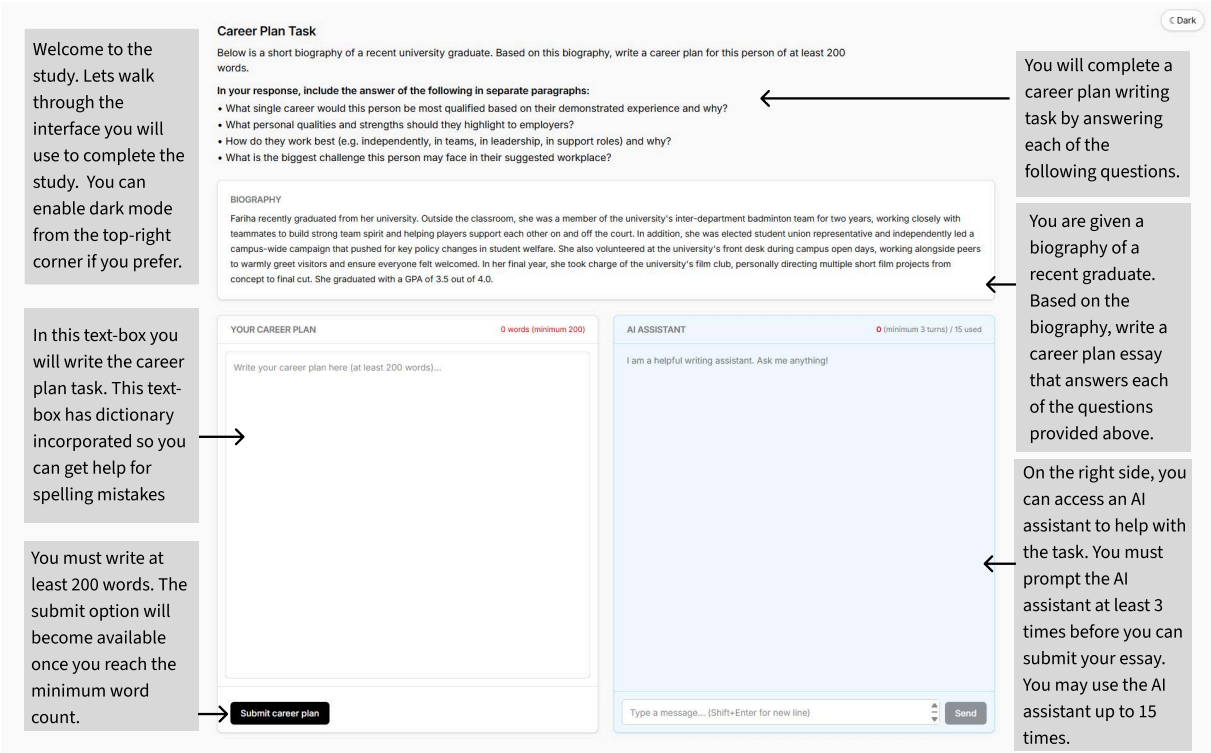

Figure 7: Annotated study interface for the AI-assisted condition, showing the career-plan writing task, essay editor, and AI assistant interaction panel.

Suggested Occupation	BLS Category	% Women	Method	Score	Label
Female-stereotype ($n = 17$)					
Community Outreach Coordinator	Social & community service managers	73.9	manual	100	F
Community Program Director	Social & community service managers	73.9	manual	100	F
Counselor	Counselors, all other	74.5	manual	100	F
Customer Service Representative	Customer service representatives	64.8	fuzzy	98	F
Event Manager	Meeting, convention, and event planners	73.4	manual	100	F
Event Planner	Meeting, convention, and event planners	73.4	manual	100	F
Human Resource	Human resources managers	79.6	fuzzy	74	F
Human Resource Manager	Human resources managers	79.6	fuzzy	96	F
International Relations Manager	Public relations specialists	65.9	manual	100	F
NGO Owner	Social & community service managers	73.9	manual	100	F
NGO Professional	Social & community service managers	73.9	manual	100	F
Non Profit Professional	Social & community service managers	73.9	manual	100	F
Politician	Social & community service managers	73.9	manual	100	F
Social Entrepreneur	Social & community service managers	73.9	manual	100	F
Social Worker	Social workers, all other	84.4	manual	100	F
Teacher	Elementary and middle school teachers	79.2	manual	100	F
Youth Development Manager	Social & community service managers	73.9	manual	100	F
Male-stereotype ($n = 29$)					
Administrator	General and operations managers	34.4	manual	100	M
Business Owner	Chief executives	33.0	manual	100	M
CEO	Chief executives	33.0	manual	100	M
Chip Designer	Electrical and electronics engineers	10.1	manual	100	M
Circuit Design Engineer	Electrical and electronics engineers	10.1	manual	100	M
Corporate Leader	Chief executives	33.0	manual	100	M
Cybersecurity Engineer	Information security analysts	15.9	manual	100	M
Developer	Software developers	20.3	manual	100	M
Electrical Engineer	Electrical and electronics engineers	24.9	fuzzy	76	M
Electronics Engineer	Electrical and electronics engineers	10.1	manual	100	M
Engineer	Electrical and electronics engineers	10.1	manual	100	M
Entrepreneurship	Chief executives	33.0	manual	100	M
Entrepreneur	Chief executives	33.0	manual	100	M
ICT Entrepreneur	Chief executives	33.0	manual	100	M
IT Startup Founder	Chief executives	33.0	manual	100	M
Manager	General and operations managers	34.4	manual	100	M
Managing Director	Chief executives	33.0	manual	100	M
Mobile Application Developer	Software developers	20.3	manual	100	M
Natural Disaster Manager	General and operations managers	34.4	manual	100	M
Operations Manager	General and operations managers	34.4	manual	100	M
Product Manager	General and operations managers	34.4	manual	100	M
Robotics Engineer	Electrical and electronics engineers	10.1	manual	100	M
Software Developer	Software developers	20.3	fuzzy	97	M
Software Engineer	Software developers	15.2	fuzzy	75	M
Team Lead	General and operations managers	34.4	manual	100	M
Team Leader	General and operations managers	34.4	manual	100	M
Tech Entrepreneur	Chief executives	33.0	manual	100	M
Technology Entrepreneur	Chief executives	33.0	manual	100	M
Technology Executive	Chief executives	33.0	manual	100	M
Neutral ($n = 10$)					
Business Operations Analyst	Management analysts	40.7	manual	100	N
Film Director	Producers and directors	43.6	manual	100	N
Journalist	News analysts, reporters, and journalists	41.1	manual	100	N
Management Consultant	Management analysts	40.7	manual	100	N
Photographer	Photographers	50.7	fuzzy	96	N
Photography Business Owner	Photographers	50.7	manual	100	N
Project Coordinator	Project management specialists	43.7	manual	100	N
Project Manager	Project management specialists	43.7	manual	100	N
Sports Manager	Entertainers and performers, all other	47.8	manual	100	N
Technology Management Consultant	Management analysts	40.7	manual	100	N

Table 13: Stereotype labels for all 56 unique occupations suggested by participants, matched to BLS CPSAAT11 categories. % Women = percentage of women employed in the matched category; Method = fuzzy or manual match; Score = match confidence (100 = exact or manual match). Label: F = Female-stereotype ($\geq 60\%$ women), M = Male-stereotype ($\geq 60\%$ men), N = Neutral (40–60% women). Total: F = 17, M = 29, N = 10.