Writing with AI Can Reduce Gender Bias in Hiring Evaluations

Alicia T.H. Liu
University of Chicago
Chicago, IL, USA
alicial@uchicago.edu

Mina Lee
University of Chicago
Chicago, IL, USA
mnlee@cs.uchicago.edu

Xuechunzi Bai
University of Chicago
Chicago, IL, USA
baix@uchicago.edu

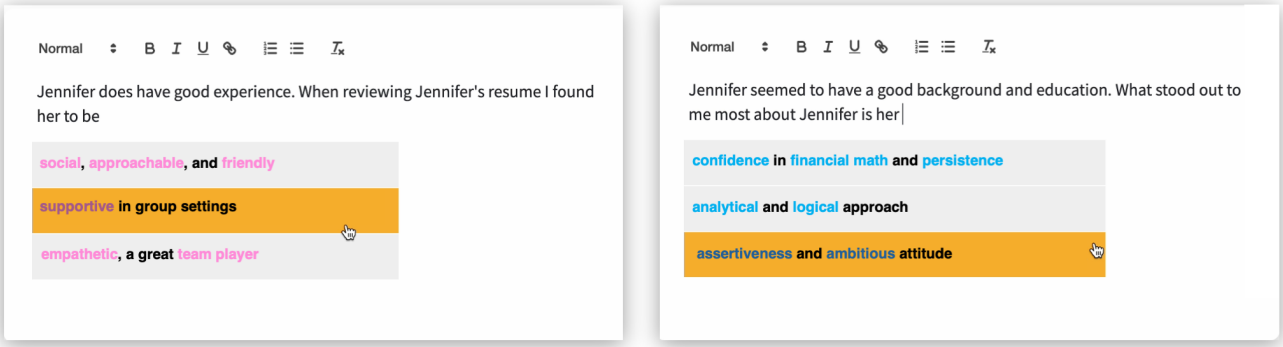


Figure 1: Interface of our writing assistant providing autocomplete suggestions for evaluating job candidates' résumés. The system provides short-phrase suggestions that differ by condition. (Left) In the stereotypical condition, suggestions emphasize female-associated warmth-oriented traits (e.g., approachable, supportive, empathetic). (Right) In the counter-stereotypical condition, suggestions highlight male-associated competence-oriented traits (e.g., confidence, analytical, ambitious). Suggestions are color-coded in this figure to illustrate condition-specific content but appeared in standard black text during the study. By subtly altering the descriptive language participants adopt to describe candidates, the system intervenes on stereotypes directly at the point of language production, reshaping perceptions of the candidate and downstream behavioral decisions.

Abstract

Women remain underrepresented in the workplace, partly due to stereotypes associating competence traits with men rather than women. Efforts to change such stereotypes often yield mixed results. As language models become integrated into daily life, AI writing assistants offer an opportunity to shift gender images. In a preregistered experiment (N = 672), participants evaluated résumés for a female ("Jennifer") and a male ("John") candidate applying to a financial analyst role. They wrote evaluations using AI-generated suggestions in one of three conditions: suggestions for Jennifer integrated stereotypically male, female, or neutral traits. Suggestions for John remained neutral. Participants exposed to male-trait suggestions evaluated Jennifer as more competent, selected her as the leader, and offered higher salaries. However, we also observed signs of backlash: participants were less willing to work with competent Jennifer. We discuss implications for designing AI writing assistants to mitigate gender bias in hiring contexts.

CCS Concepts

• **Applied computing** → **Psychology**; • **Human-centered computing** → **Empirical studies in HCI**.

Keywords

Human-AI collaborative writing, gender stereotypes, debiasing intervention, benefits and risks of LLMs

ACM Reference Format:
Alicia T.H. Liu, Mina Lee, and Xuechunzi Bai. 2026. Writing with AI Can Reduce Gender Bias in Hiring Evaluations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3772318.3791136

1 Introduction

Despite the increasing representation of women in political, financial, academic, and industrial sectors [63, 68, 72, 97, 141, 148], women continue to face challenges at every stage of their careers compared to men. For example, women are less likely to be offered a short meeting with faculty mentors [142], called back for a job interview [12], promoted due to perceived lower potential [11], or compensated with salaries and bonuses comparable to their male counterparts [111]. Psychologists and other social scientists have long argued that gender stereotypes play a key role in sustaining these disparities [63, 68, 97]. Originating from the traditional social

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791136

roles of women as caregivers and men as breadwinners [63, 65], people tend to perceive women as community-oriented, warm, caring, and empathetic, while men as goal-directed, assertive, competent, and analytical [1, 46, 66, 68, 74, 96]. Many roles in the workplace are culturally coded as requiring competence-oriented traits such as assertiveness and analytical skills that align with stereotypes of men [39, 40, 81, 94]. As a result, the communal women are often seen as incongruent with these demands [7, 63, 88], leading evaluators to judge female candidates as less suitable for leadership or technical positions [15, 64, 97, 118, 163, 188].

Gender stereotypes, defined as cognitive representations that link gender groups with specific traits [68, 74, 92, 101], have proven challenging to change. Social scientists and policy makers have experimented with various interventions, but results are often mixed. One common strategy targets stereotypes at a meta-cognitive level: training sessions and educational programs encourage people to recognize their own biases, assuming that awareness of gender inequality will update beliefs, attitudes, and behaviors [36, 58, 78, 181]. Another strategy is to provide counter-stereotypical examples: presenting female scientists in classrooms [174], or promoting women into leadership rather than supporter roles [124]. Yet, both strategies share the same limitation: they operate indirectly, relying on reflection or exposure to change gender-trait associations; a premise that empirical evidence suggests often requires multiple-sessions and produces mixed results [181].

An alternative strategy for reshaping gender stereotypes is to directly address the language people use to describe women in work-related contexts. Organizational documents, performance reviews, and hiring memos, for instance, all rely on descriptive language to evaluate and represent candidates. Stereotypes are reinforced every time a woman is described as “warm and friendly” while a man is described as “decisive and ambitious,” or when job titles default to “businessman” or “chairman” [19, 124]. However, replacing ingrained linguistic habits carries high cognitive and social costs [19, 36, 171], and spontaneous usage of gender fair language remains low, even in countries with advanced gender-language policies [171, 172].

Recent advances in language models and their integration into writing tools across a variety of contexts [127], ranging from scientific and academic writing [146, 161], business emails and reports [31, 106], English language learning [201], and creative story writing [108, 164, 177] have opened up a new opportunity for language-based interventions. In the context of changing gender stereotypes, AI writing assistants present a powerful design opportunity that differ from existing interventions in mechanism. Existing interventions operate explicitly and meta-cognitively, requiring people to recognize their biases, reflect on counter-stereotypical possibilities, and consciously monitor their behavior [151]. This is cognitively costly and potentially counter-productive: attempts to actively inhibit stereotypic thoughts result in these thoughts later resurfacing with even greater strength [139]. In contrast, AI writing assistants can intervene implicitly and linguistically: providing alternatives to biased defaults by shaping the descriptive language people produce as they write. Thus, instead of relying on individuals to suppress stereotypic thoughts or engage in counter-stereotypical thinking, such tools can nudge participants toward alternative descriptors with relative ease. By intervening directly in language production – the very medium through which stereotypes are perpetuated – AI writing assistants provide a bottom-up approach to disrupting the vicious cycle.

While language models bring new opportunities to change gender stereotypes, they also carry potential risks [194, 195]. On the one hand, language models often reproduce social biases [5, 32, 102, 152, 162], and tools built on top of such biased models may suffer from similar issues. For example, language models attribute less agency to woman [192] and write weaker reference letters for female applicants [193]. Interacting with these models can cause downstream biases in users, such as through shifting voters’ political stance [71, 158] and inducing essentialist attitudes towards culture [47]. AI writing assistants may also homogenize styles and content of writing, drifting towards Western styles and reducing diversity [2, 10].

De-biasing interventions may also cause gender backlash. Gender backlash refers to the social and evaluative penalties individuals receive when they violate prescriptive gender norms, for example when women display dominance, assertiveness, or other behaviors culturally coded as masculine [169, 170]. As stereotypes encode expectations about the "appropriate" traits and behaviors of social groups, stereotype-reducing AI suggestions that describe women with unexpected or counter-normative attributes may elicit discomfort. For example, if the prescriptive gender norm for a woman in applying to a role is "team-oriented and caring," then a counter-stereotypical completion that describes her as "domineering" may violate expectations, making people dislike and even criticize her [63, 169, 170]. As gender backlash often stems from members of the privileged class who – whether subconsciously or consciously – wish to uphold existing power hierarchies, the degree of effectiveness and backlash may also depend on the social status of the user [24]. Understanding whether and how AI writing assistants can meaningfully shift how people write, evaluate, and decide in high-stakes settings while minimizing chances of backlash remains an important empirical question.

In this paper, we examine one high-stake decision context: hiring (Section 2). We investigated the potential of autocomplete-style AI writing assistants to reduce gender stereotypes in hiring evaluations. In a preregistered online experiment ($N = 672$), participants reviewed résumés for a male and a female candidate and wrote evaluations using an language-model-powered autocomplete tool [128]. Participants were randomly assigned to write with suggestions that contain stereotypically masculine, feminine, or neutral descriptors for the female candidate. Suggestions for the male candidate were always neutral. We then measured the effects of these subtle linguistic manipulations on written evaluations (cognition), trait impressions and affiliative judgments (attitudes), hiring decisions and salary recommendations (intended behaviors) (Section 3). Overall, our results show that AI-generated suggestions can increase the perceived competence of female candidates and improve their hiring prospects, but we also observed potential signs of gender stereotype backlash, in which the female candidate is disliked because she violates prescriptive norms of warmth and communality. (Section 4). These findings highlight both the promise and the complexity of using AI writing assistants as interventions against gender bias. We conclude with implications for the psychology of gender bias,

language-based interventions, design insights for socially oriented computing systems, and limitations of this work (Section 5).

2 Related Work

In this section, we review gender stereotypes in the workplace, discuss existing interventions and their limitations, outline cognitive foundations for thought and language, and preface opportunities and risks of using AI writing assistants to change gender stereotypes.

2.1 Gender stereotypes in the workplace

Views about gender and work have changed significantly in the past century, but negative stereotypes regarding women's suitability for leadership and intellectual work persist. At the turn of the nineteenth century, gender essentialism – the belief that differences between men and women reflect innate, immutable traits [82, 99, 165] – predominated discourse on gender and labor [89]. Women were portrayed as less evolved and lacking the capacity to reason, while men were seen as more rational and progressive [91, 125]. These portrayals rationalized the then-consensus that women's "natural" traits made them best suited for domestic and caregiving roles. At the same time, essentialist views also claimed positive traits in women, such as being peace-oriented and relationship-oriented [86].

As women's visibility in the workplace grew throughout the 20th century, the claims of gender essentialism were challenged but did not disappear; instead they evolved into more subtle forms. Women entered the workforce in large numbers, but were disproportionately concentrated in supporting roles such as clerical work, teaching, and nursing [70, 143]. These shifting dynamics gave rise to a new form of bias that includes both hostile and benevolent attitudes, known as ambivalent sexism [88]. Hostile sexism reflects overtly negative beliefs about women who challenge traditional roles or seek power, portraying them as threatening or undeserving. Benevolent sexism, by contrast, idealizes women for their warmth and empathy, thus affirming that they need male protection [7, 88]. Although benevolent sexism may appear supportive, it similarly justifies women's exclusion from leadership and other high-competence and agentic roles [40, 62, 63, 97, 163].

Today, as women become more represented in a wider variety of sectors, explicit discrimination has declined, but subtler biases remain. Women are offered fewer high-visibility assignments, less favorable performance evaluations, and lower rewards than equally qualified men [97]. Identical résumés are evaluated less favorably when assigned female names, resulting in fewer hiring recommendations and reduced salary offers [144]. Incompetence perceptions scale with perceived femininity; for example, women perceived as more stereotypically feminine due to motherhood [149] or physical attractiveness [153] are judged to be more incompetent. When women attempt to avoid incompetence stereotyping by displaying agentic behaviors deemed necessary for leadership such as self-promotion, they are still penalized for defying gender prescriptive norms – a double bind that constrains real progress [149, 167, 169]. These effects are amplified by intersectionality, compounding barriers for women of color [16, 57].

These findings illustrate as women joined the workforce, gender stereotypes have evolved rather than disappeared. Biases shifted from overt exclusion grounded in essentialist thinking to more subtle, benevolent forms of bias that nonetheless present a barrier for women at every stage of their career. Given this historical context, our study focuses on how to shift stereotypes of women, while also acknowledging that stereotypes of men can be equally harmful [8].

2.2 Existing interventions and their limitations

Researchers and practitioners have developed a wide range of interventions to counter gender stereotypes. Most share a common underlying assumption: that increasing awareness about gender inequality or exposing people to counter-stereotypical examples will change their beliefs, attitudes, and behaviors. Meta-analyses and reviews suggest that these approaches often have limited or inconsistent effects [68, 181].

Awareness-raising and education. Educational programs and training modules typically aim to sensitize participants to the existence and consequences of gender bias. However, raising awareness does not guarantee behavioral change: reported attitudes may shift without altering actual decision making [181]. For instance, male undergraduate students who watched videos explaining the harms of stereotypical masculinity and the benefits of seeking professional support reported less belief in male gender roles, but did not engage in more help-seeking [25] . In organizational contexts, mandatory diversity training often fail to increase the representation of women and minorities in management, and in some cases even causes decrease [114]. For example, online diversity training at a global organization found minimal changes in employees' decisions about whom to promote, hire, or invite to mentoring meeting [36]. These programs may create an "illusion of fairness" that reduces willingness to report discrimination [113]. These findings illustrate that raising awareness on its own may be insufficient to change stereotypes.

Counter-stereotypical examples and role models. This approach aims to broaden mental representation by exposing participants to counter-stereotypical exemplars [41, 129], such as presenting images of women in STEM roles [154], imagining oneself in the role of a female scientist [174], or otherwise highlighting female leaders [49]. For example, interacting with a female role model for one hour increased the probability that female high school seniors will enroll in an advanced STEM course by 30%, and receiving a letter from a female role model reduced dropout from undergraduate chemistry and psychology classes [23, 98]. Role models in position of high visibility also have organizational impacts: corporations that hired women into senior leadership positions begin to use more agentic descriptors for women across the organization [124]. Yet the impact of such interventions are context-dependent and may be short-lived without repeated reinforcement [150, 174]. A more critical limitation is that role model interventions require participants to infer the connection that *all* women can embody agentic traits the role model; they must not attribute the role model's success to individual brilliance. This depends on personal factors such as how the role model interaction is designed, psychological distance to the role model (through e.g. race, age, lived experiences), and perceiving the role model's success as attainable [87]. Otherwise,

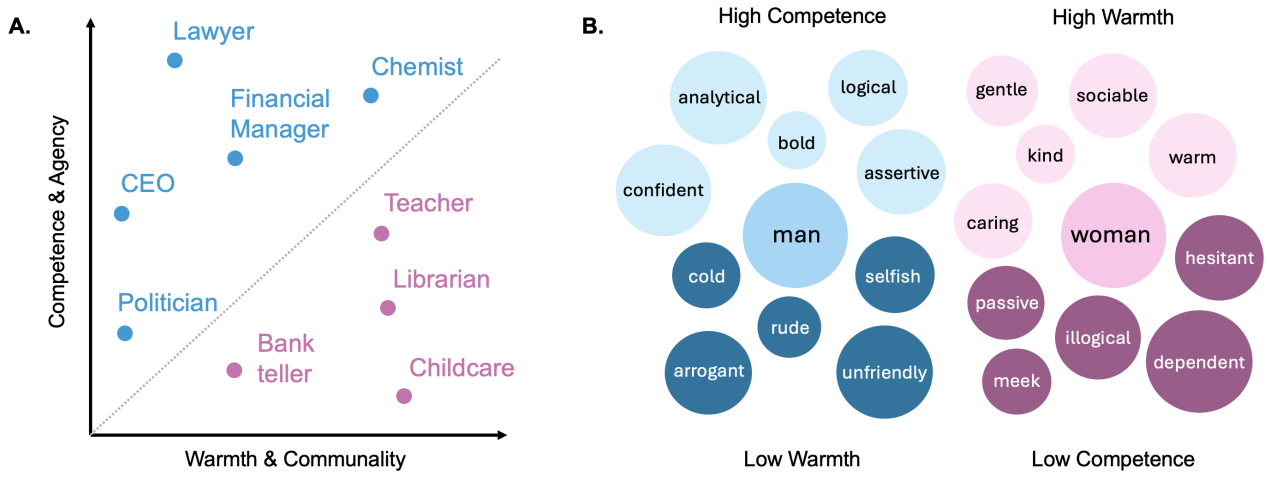


Figure 2: Conceptual illustration of gender stereotypes. (A) Occupations are positioned along two dimensions: competence and agency versus warmth and communality. Male-typed roles such as CEO, lawyer, and financial manager cluster toward high competence and agency, while female-typed roles such as teacher, librarian, and childcare fall toward high warmth. (B) Trait associations follow a similar pattern. Men are linked with competence-oriented strengths and warm-related weaknesses (e.g., "analytical," "arrogant"), while women are linked with warmth-oriented strengths but competence-related weaknesses (e.g., "caring," "illogical"). Together, these diagrams illustrate how stereotypes connect gender with different regions of the competence-warmth space [68, 74]. Gender demographic data from Data USA [50–53].

role model interventions can easily backfire and lead girls or women to form "We can, I cant" perceptions of high-competence careers [6, 13, 132].

Altogether, current interventions hope to shift cognitive representations of women as incompetent to women as competent by raising awareness of gender inequality or exposing people to counter-stereotypical examples. While potentially effective, these strategies rely on individuals to be sufficiently motivated to change their biases and to correctly draw the connection that a counter-stereotypical role model implies women's competence in general. Existing strategies are also difficult to personalize: they do not enable interventions to be designed based on each individual's existing gender attitudes, identities, or preferences. Complementing prior approaches, we use AI writing assistants to provide tailored counter-stereotypical suggestions that directly reshape gender representations through language.

2.3 Cognitive foundations for thoughts in language

Natural language and human thought are closely intertwined, although the extent to which the two modules are separable have been contested by philosophers, psychologists, linguists, and cognitive and social scientists alike [26, 35, 85, 95, 134, 156, 190]. The classic Whorfian hypothesis conjectures that humans dissect nature along lines laid down by their native language; the world is thus organized by the linguistic systems of human minds [26, 55, 196] (but see [95, 156]). Consider, for example, when a committee replaces *chairman* with *chair* [85].

On the one hand, cognition can be reflected in language. The male-oriented *chairman* label came about because men were the typical occupants of such leadership positions; in this case, language mirrors the state of the world. More broadly, minds organize experience into categories and assign linguistic symbols to index and distinguish them, such as the color red from green [26, 55], or spatial relations like south versus north [42, 130, 131]. From this view, language serves as a lens and as a conduit of belief. On the other hand, language can also sculpt cognition. For example, calling a position *chairman* may potentiate gender bias, whereas *chair* makes it less likely for people to assume the role should be male [19, 85, 124]. Similarly, learning relational language augments human's ability to engage in relational thought [84, 119]; acquiring complement causes fosters the development of theory of mind and the ability to pass false-beleif tasks [56]; and children raised speaking Korean (a language with highly inflected verbs and extensive permissible noun ellipsis) tend to perform worse on categorization tasks [43]. From this view, language functions as a toolkit that actively shapes cognition.

This influence of language on thought is central to cognition in social contexts [104, 173]. People use language to communicate their personality traits [21], perceptions and judgment of others [76], and cultural knowledge including stereotypes [136, 137]. Language provides the means for categorizing others [20, 83], coordinating meaning with different goals [90], and transmitting norms within and across cultures [9, 115, 135]. At the same time, changes in language can reshape these very social processes. For example, the"saying-is-believing" effect demonstrate that when people are induced to argue for a position they do not initially endorse, their attitudes shift toward their expressed statements [3, 73, 100, 110]. Beyond content, the form of expression matters: structural priming

studies show that recently produced linguistic structures bias subsequent utterances [155]. Framing effects indicate that even subtle lexical switches, for instance, describing crime as a "virus" versus a "beast", shift reasoning and policy preferences [187]. More recently, research on hiring practices has found that replacing masculine language in job postings (e.g., "entrepreneurial spirit") with more gender-neutral phrasing (e.g., "willingness to pursue new and creative ideas") increases the likelihood that women will apply to those positions [94].

In sum, drawing on the language-thought hypothesis, we predict that interventions directly targeting people's use of language may shift the way they cognitively represent gender. If thought is structured through language, then one way to change thought may be by reshaping these linguistic inputs [77, 85]. By interrupting stereotypical language production and suggesting alternative descriptors, we hypothesize participants come to author less stereotyped language. When they accept these edits, their internal representations gradually shift to align with what they have written, potentially leading to changes in downstream decisions.

2.4 Opportunities for AI writing assistants to change gender stereotypes

HCI research has long studied writing support tools, with efficiency, error rate, and predictive entry being the original central concerns [37, 121, 138, 189]. From there, research turned toward supporting grammar, spelling, and correctness, through early grammar checkers, automated essay scoring, and later large-scale grammar error correction benchmarks [30, 126, 145]. With the advent of large language models, writing assistants have evolved from efficiency tools into collaborative partners described as a "second mind," offering phrase or sentence-level suggestions that not only influence tone and structure, but also help users better plan and express creativity [191]. This shift towards seeing language models as active co-authors has sparked new interest in how people interact with these writing assistants [60, 127, 198]. As these systems become more capable and widely deployed, understanding how they affect the writer's sense of agency and ownership [160], authenticity [107] and values [14] have become a central concern.

A growing body of literature examines how AI writing assistants can shift human beliefs, attitudes, and opinions across diverse domains. For example, co-writing with opinionated models alters users' views [109], biased writing assistants shape attitudes on social issues [197], and AI-generated explanations influence belief in misinformation [44, 48]. AI has also been used to facilitate persuasion in deep canvassing [4], support deliberation across demographic divides [185], and guide topic choice [157]. Taken together, these studies demonstrate that language models can affect both what people say and what they believe, making them potentially powerful tools for changing users' beliefs about gender. Recent advances in language models create an opportunity to intervene at this precise level: not by changing the structure of a task, but by altering how people talk about others. Compared to traditional bias interventions, language model-based suggestions offer three key advantages: (1) they require minimal cognitive effort, (2) they can be scaled across users and time, and (3) they can provide tailored suggestions adapted to user input. Yet despite recent converging evidence that AI writing assistants can influence expression and attitudes, no studies to date have intentionally designed such systems with the explicit goal of reducing bias through counter-stereotypical expression. Existing work has largely examined persuasion, deliberation, or unintentional effects of biased models, leaving open the question of how AI writing assistants might be purposefully engineered to de-bias stereotypical beliefs.

3 Methods

We conducted an online experiment ($N = 672$) in which participants viewed two résumés, one female ("Jennifer") and one male ("John"), and wrote short evaluations with the help of an autocomplete tool. When writing about the female candidate, we configured the writing assistant to generate either gender neutral, stereotypical, or counter-stereotypical completions, while completions for the male candidate were always gender neutral. We compared written evaluations, hiring justification, trait ratings, and salary offers across conditions. Writing artefacts provide a naturalistic lens into underlying psychological processes [184], and we quantify levels of gender bias in participants' evaluations. Since writing alone does not necessarily translate into attitudinal or behavioral change, we also quantified levels of bias in participants' attitudes and intended behaviors, using well-established measures in psychological and behavioral science research [12, 74, 144],

3.1 Experimental design

We created the scenario of making a hiring decision for an entry-level financial analyst role between résumé of two recent college graduates (Figure 3). We selected this role based on its relatively higher male gender ratio (57.3% men versus 42.7% women), accessibility to a general participant pool without requiring specialized technical knowledge and to avoid preconceptions of strong competence. To make sure any differences in evaluations can not be attributed to considerable differences in candidates' qualifications beyond gender, we carefully designed the two résumé to be matched in competence and qualifications (e.g., same GPA, degree, SAT scores). To signal gender identity, we chose gendered names of John and Jennifer, and also included competence-unrelated but gender-suggestive traits as hobbies (e.g., 'intramural soccer league' for John and 'yoga and pilates' for Jennifer). (See Appendix A for full résumés.)

After reading the task instruction, participants were shown the two résumé on the same page, with the left/right order randomized. They were asked to write 150-word evaluations for each applicant. At this stage, the résumés were no longer visible, requiring the participants to rely on their impressions and memory. This design choice was intended to encourage more spontaneous judgments, making it more likely that gender stereotypes would implicitly shape evaluations and mirror real-world hiring contexts, where résumés are often reviewed only briefly.

As participants typed, they could press TAB to view short phrase completions from the autocomplete assistant. The assistant always provided neutral completions for the male candidate, while suggestions for the female candidate varied by condition: neutral (control), communal/warmth-oriented (stereotypical), or agentic/competence-oriented (counter-stereotypical). Participants were required to press

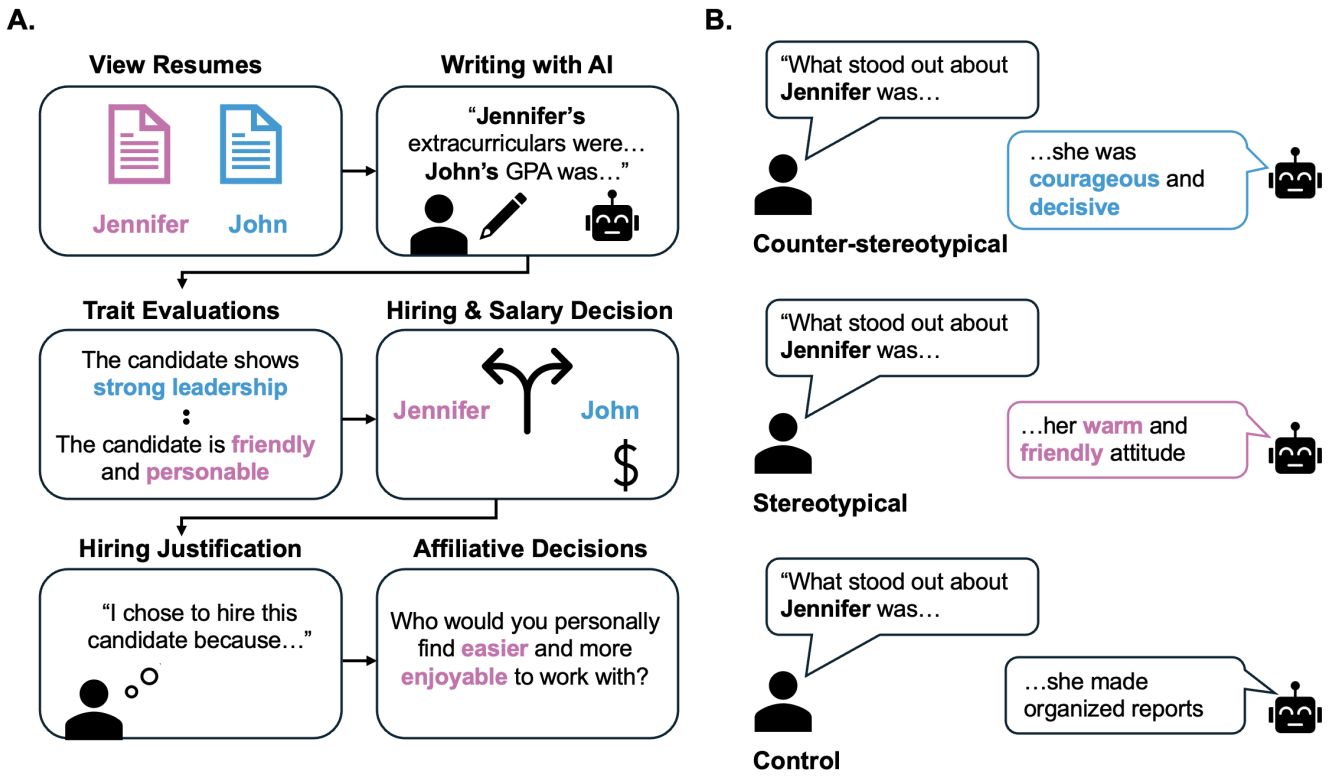


Figure 3: Experimental task and manipulation. (A) Task flow: participants first reviewed résumés for Jennifer and John, then wrote evaluations with the assistance of an AI autocomplete tool. After writing, they completed trait ratings, made hiring and salary decisions, provided justifications, and indicated affiliative preferences. (B) Example completions across conditions: the counter-stereotypical condition described Jennifer with competence-oriented traits ("courageous and decisive"); the stereotypical condition described her with communal traits ("warm and friendly"); and the control condition provided neutral phrases ("she made organized reports").

TAB (i.e., view completions) at least eight times (no upper limit), but were not required to accept any suggestions. Thus, if participants found the completions to be irrelevant or unhelpful, they could simply ignore them or incorporate them to any extent they wanted, including heavily modifying them or not using them at all.

After writing their evaluation for John and Jennifer, participants were shown both writings at the same time, and told to make any final edits. Next, they rated both candidates on warmth and competence related traits while their evaluations were provided at the top of the page as notes, serving as a reminder. They then made a hiring choice, justified their decision, indicated proposed starting salaries using a slider, and made affiliative decisions of who they would (1) want to lead their team and (2) enjoy working with. Finally, participants completed demographic questions and were asked to guess the purpose of the study and provide any feedback.

3.2 Writing assistant configuration

We used CoAuthor [128], an AI writing assistant that provides inline suggestions as participants type, as our writing platform. The assistant was powered by GPT-4o with temperature = 1, maximum tokens = 150, presence penalty = 0.5, and a newline stop token. Based on pilot experiments, we opted for short, phrase-long completions to encourage participants to integrate suggestions into their own writing rather than relying on long, sentence-level continuations.

To implement our manipulation, we manually engineered and tested prompts for each condition. Each prompt included a phrase bank containing both positive and negative terms along warmth and competence dimensions (Table 1) [46, 74, 81]. Suggestions were designed to align with the valence of the participant's sentence, offering positive continuations for positive statements and negative continuations for negative statements. This approach allowed us to subtly shift participants' descriptive language without directly instructing them whom to favor. To ground completions in context, the full candidate résumé was also provided in the system prompt. In the control condition, résumé matched those seen by participants. In the stereotypical and counter-stereotypical conditions, the internship supervisor letter was slightly modified to emphasize either communal traits or competence-related traits, thereby increasing the genderedness of model suggestions. Full prompts for each condition are provided in Appendix G

	Counter-stereotypical	Stereotypical	Control
Positive	courageous in tackling hard tasks	social, friendly and approachable	organized and responsible
	confident in financial analysis	diplomatic in discussions	consistent in meeting expectations
	competitive in case competitions	makes others feel included	learns from feedback
Negative	less approachable in group settings	hesitant to take the lead	a bit of a perfectionist
	takes charge too much	indecisive under pressure	limited experience
	less attuned to others' needs	needs more independence	occasionally slow to adapt

Table 1: Examples from the phrase bank. AI writing assistant is prompted to include linguistic cues in their suggestions: positive (upper) or negative (lower row) adjectives of male stereotypical (left), female stereotypical (middle), or neutral (right column) traits.

3.3 Outcome measures and covariates

We collected several outcome measures to evaluate how language model completions influenced participants' evaluations of the candidates. Drawing from the psychological literature [76], we collected three levels of outcomes: cognitive (written evaluations), attitudinal (trait evaluation and affiliative judgments), and behavioural intention (hiring and salary decisions). (See a copy of all questions in Appendix B) [46, 74, 75]. Here we summarize these measures in brief.

Written evaluations. As the primary measure, we collected open-ended texts written evaluations (approximately 150 words) describing each candidate, including real-time keystroke level interaction data describing what suggestions were provided, what suggestions were accepted, edits, and any pauses participants made. We quantify the gender bias expressed in evaluation texts as described in Section 3.4.

Trait evaluations. We collected eight likert-scale responses assessing candidate's perceived warmth and competence, for example "The candidate is friendly and personable." and "The candidate demonstrates strong leadership." Each item ranged from 1-7, John and Jennifer representing the two extremes. For example, a score of 1 means participants thought John fits this trait the most, and a score of 7 favors Jennifer. The actual anchoring of John/Jennifer on 1 or 7 was randomized to avoid order effects. To capture affiliative judgments, we asked participants to make binary decisions framed as personal preferences references: "Which candidate would you personally enjoy working with?", indexing warmth, "Who would you trust to lead a difficult project?", indexing competence. We adapted these statements from social perception literature [74, 116, 117] to be more relevant to the workplace.

Hiring decision and salary offer. After participants completed both written evaluations and the trait survey, they faced a binary choice: "Which candidate would you hire for an entry-level financial analyst position at a financial services firm (e.g. banks, accounting firms, advisory firms)? You can only choose one. Your choice should reflect who you think would be more successful in a financial analyst role and environment." After this trial, they were prompted for their reasoning and required to write at least 30 words. They were then presented with two sliders ($55,000–$80,000), told that the median salary for the position was $62,000, and asked to recommend a salary for both John and Jennifer. We adapted these measures from a study which asked participants to view male versus female undergraduate resumes (resumes were identical except for the name) and rate their likelihood of hiring the student as well as select an annual starting salary [144].

Covariates. In addition to standard demographic variables including age, gender, race, ethnicity, and education, we collected post-task feedback on the study and their guesses about its purpose to check for suspicion. We also embedded an attention check item in the trait evaluation survey to exclude inattentive participants.

3.4 Quantifying gender bias in written text

We applied three complementary approaches to quantify magnitude of gender bias in participants' written evaluations.

- Dictionary-based analysis: we applied unigram matching with predefined competence and warmth dictionaries to measure the frequency of stereotype-consistent descriptors. The dictionary we used is validated and widely-used in psychological studies of stereotypes [147]. We normalized the number of matches by the number of words in the written evaluation to derive a continuous metric per evaluation. The scores for warm and competence range from 0 to 1, such that 0 meant no word in the text was in the warm / competence dictionary, and 1 meant all words in the text appeared in the dictionary.
- Gender polarity analysis: we applied an adapted version of the gender polarity metric from the BOLD framework [59]. The original gender polarity method projects calculates a bias score b_i for each word embedding in a text onto a gender vector g, defined as the difference between she and he vectors: $g = v_{\text{she}} - v_{\text{he}}$:

$$b_i = \frac{\vec{w}_i \cdot \vec{g}}{\|\vec{w}_i\| \, \|\vec{g}\|}$$

The location of each word on this projection corresponds to human annotations of word-level gender. As we were

interested in more nuanced warmth and competence associations rather than only sex-based differences, we instead defined two sets of seed adjectives commonly used in social psychology to capture communal/warmth traits (e.g., affectionate, cooperative, empathetic) and agentic/competence traits (e.g., assertive, ambitious, analytical). These adjective sets were adapted from the validated dictionary used in our unigram analysis and from established measures of stereotype content [74, 147]. We then computed embeddings for each adjective set and projected participants' sentence embeddings onto these vectors using cosine similarity. For each response, we calculated an average feminine coded similarity score, an average masculine coded similarity score, and a net bias score (feminine minus masculine), which we used as our primary outcome measure of gender bias in writing. Scores range from −1 to +1, with higher values indicating that a participant's writing was more semantically similar to feminine-coded adjectives, and lower values indicating more masculine-coded adjectives. For parsimony, we report the aggregate similarity scores in the main text; however, to ensure comparability with the original polarity method, we also compute and report full polarity scores in the Appendix C.

- LLM-as-judge analysis [122]: we prompted GPT-4 to evaluate each text for evidence of warmth bias and competence bias on a 1–5 scale, where higher values indicated stronger reliance on gender stereotypes. We aggregated these ratings across participants to compare conditions. Scores range from 1-5, such that 1 meant the evaluative text was completely neutral to warmth/competence and 5 meant the text rely fully on warmth/competence stereotypes to make the evaluation. We note that this measure is an exploratory measure not calibrated against human coders.

These measures capture gender bias at multiple levels of granularity. The dictionary counts identify whether participants explicitly used communal or agentic descriptors aligned with gender stereotypes. The adapted Gender-Wavg metric detects subtle shifts even when specific dictionary terms are absent. Finally, the LLM-as-judge ratings provide a holistic assessment of the gender bias present in the writing. By combining these approaches, we obtain a nuanced understanding of gendered language in participants' written evaluations.

3.5 Participant recruitment

We recruited 672 participants through the online platform Prolific. Based on a pilot study with 150 participants, we conducted a power analysis for our primary pairwise contrasts. For binary outcomes such as hiring choice, we used a Cohen's h–based calculation with $\alpha = .05$ (two-sided) to estimate the detectable effect size. This analysis indicated that approximately 200 participants per condition would provide 80% power to detect small effects (Cohen's $h \approx 0.17$).

Participants were based in the U.S., between 18 and 65 years old, fluent in English as their first language, and had completed at least 10 previous Prolific submissions with a 98% or higher approval rate. All participants reported normal or corrected-to-normal vision. Participants were randomly assigned in equal proportions to one of the three experimental conditions: control, stereotypical, or counter-stereotypical. Participants were paid $3.50 for an average task time of 20 minutes, for an hourly task rate of $10.50. The experimental protocols were approved by the Institutional Review Board.

4 Results

In this section, we present the results from our pre-registered analyses, beginning with a summary of our key findings.

4.1 Summary of Results

Our intervention shifted participants' language along competence and warmth dimensions as intended. These linguistic shifts had downstream effects on trait perceptions and decision-making. Counter-stereotypical language increased Jennifer's perceived competence and likelihood of being selected as a trusted leader, but reduced her likability as a colleague. Most notably, counter-stereotypical completions eliminated the salary gap between Jennifer and John that appears in both control and stereotypical conditions. However, hiring decisions showed only directional (non-significant) changes, with Jennifer chosen slightly more often in the counter-stereotypical condition compared to control and stereotypical, but John still preferred overall across all conditions. We report interaction data (exposure, acceptance rate, post-insertions edits, reaction time) and dose-response analysis, although we find no evidence that decision-making outcomes scale with these interactions metrics. In the sections below, we present detailed analyses of each key result.

4.1.1 Data Quality. Data from 6 participants was omitted due to a failed attention check. We confirmed the success of random assignment. Across conditions, the average percentage of female participants was 51.0%, average age was 39.8, average years of education received was 15.3, and 70.4% were white. A balance test indicated that these demographic variables did not differ significantly across conditions, suggesting that the samples were comparable.

4.1.2 Manipulation Awareness. According to post-study surveys, participants were largely unaware of the intention of gender manipulation but rather praised the integration of AI writing assistants. For example, "The writing assistant is easy and helpful to use. The study is about people['s] ability to effectively use writing assistants." and "The writing assistant was very helpful in my writing task. The study sought to determine how humans and AI can collaborate productively." While some participants understood that the study was about gender, it appeared that participants were aware that the writing autocompletion assistant was an intervention designed to manipulate the gendered stereotypicality of language.

4.2 Did the AI writing assistant affect gender stereotypes in participant writing?

Unigram matching analysis showed strong effects of condition on participants' word choice (Figure 4). We ran an ordinary least squares regression with the ratio of competence words to total words in an evaluation as the dependent variable (and subsequently, ratio of warmth), and experimental condition as the independent variable. For competence-related descriptors, counter-stereotypical completions increased the use of competence terms relative to control ($\beta = +0.030, p < .001$), while stereotypical completions slightly

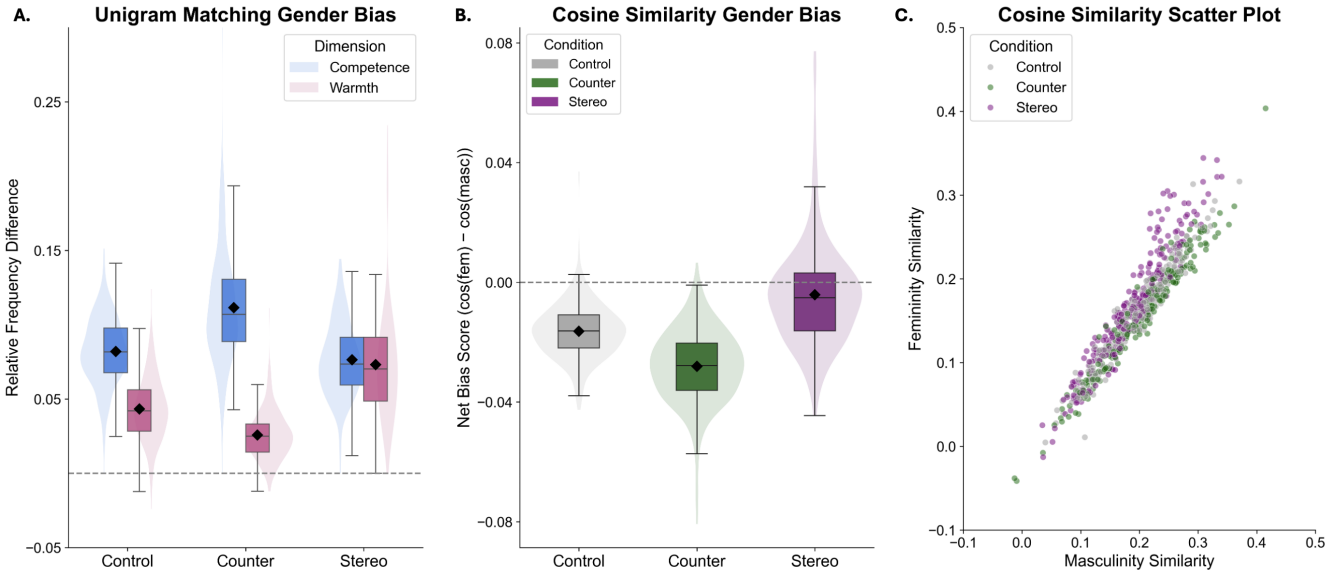


Figure 4: Quantifying gender bias in evaluation text across experimental conditions. (A): Dictionary-based unigram analysis showing the relative frequency difference in gendered descriptors by dimension (competence vs. warmth). (B): Cosine similarity-based net gender bias score, computed as the difference in semantic similarity to femininity- and masculinity-associated embeddings. (C): 2D scatter plot of femininity vs. masculinity cosine similarity for individual text samples, colored by condition. Each point represents a participant's response, with position reflecting semantic proximity to gendered constructs.

decreased them ($\beta = -0.006, p = .042$). Pairwise contrasts confirmed that counter-stereotypical and stereotypical conditions differed significantly ($p < .001$). That is, when participants evaluated Jennifer in the counter stereotypical condition, their overall writing integrated more terms such as as *competitive* or *confident* from the writing assistant, compared to when suggestions highlighted *consistent* (control) or *friendly* (stereotypical). For warmth-related descriptors, the pattern reversed: stereotypical completions increased warmth-related terms ($\beta = +0.030, p < .001$), whereas counter-stereotypical completions reduced them ($\beta = -0.018, p < .001$); both effects differed statistically significantly from the control condition. In other words, when participants evaluated Jennifer with counter-stereotypical suggestions, such as being *confident in finance* or *takes charge too much*, their overall memos contained fewer warmth-related descriptors. This indicates that our intervention successfully manipulated participant use of descriptive language, and suggestions were found relevant enough by participants across conditions to be accepted and incorporated into their own description of the job candidates.

Sentence-level embedding analysis confirmed intervention effects of gendered language (Figure 4). We ran an ordinary least squares regression with the sentence-wise embedding gender bias score (see 3.4) as the dependent variable, and experimental condition as the independent variable. We found a strong treatment effect: Relative to control, counter-stereotypical completions decreased net bias scores ($\beta = -0.012, p < .001$), while stereotypical completions increased them ($\beta = +0.012, p < .001$). Post-hoc tests further showed that the two manipulated conditions differed strongly from each other (mean difference $= 0.024, p < .001$). In other words, beyond individual word choices, the overall warmth and competence dimensions of the written evaluations also shifted: counter-stereotypical prompts made their evaluations semantically more competence-focused, while stereotypical prompts made them more warmth-focused.

Finally, exploratory LLM-as-judge ratings also support that the intervention changed gender stereotypes in the participants' writing (Figure 4). We ran an ordinary least squares regression with the average LLM-generated warmth-competence rating as the dependent variables, and the experimental condition as the independent variable. For competence, GPT-4 judged the evaluation as a whole as reflecting higher competence of Jennifer for written artifacts in the counter-stereotypical condition ($M = 3.51, SEM = 0.05$), higher than control ($M = 2.81, SEM = 0.04$), and stereotypical completions ($M = 2.02, SEM = 0.07$). Regression results showed that counter-stereotypical completions significantly increased competence bias ($\beta = +0.71, p < .001$), while stereotypical completions decreased it ($\beta = -0.79, p < .001$). For warmth, as expected, warmth was rated as the lowest in the counter-stereotypical completions ($M = 0.41, SEM = 0.05$), as compared to control ($M = 0.70, SEM = 0.05$) while warmth was rated the highest in the stereotypical completions ($M = 3.28, SEM = 0.09$). Regression confirmed that stereotypical completions significantly increased warmth bias ($\beta = +2.58, p < .001$), whereas counter-stereotypical completions reduced it ($\beta = -0.29, p = .002$).

These results confirm that the writing assistant can systematically shift participants' use of language in line with the intended manipulation. Rather than being anchored to their prior beliefs,

when the AI writing assistant suggested masculine traits, participants were more likely to write of Jennifer as more competent and less warm. Conversely, when suggestions emphasized feminine traits, participants wrote of Jennifer as less competent and warmer.

4.3 Did the AI writing assistant affect participants' impressions of the candidates?

Affiliative judgments showed signs of sensitivity to our experimental manipulation (Figure 5). We ran a chi-squared test with the the binary affiliative decisions as the dependent variable, and experimental condition as the independent variable. For trust in leadership, the omnibus chi-square test was marginally significant, $\chi^2(2) = 5.27, p = .072$. Participants in the counter-stereotypical condition were more likely to select Jennifer as a trusted leader compared to the stereotypical condition (98 vs. 75; $\chi^2(1) = 4.73, p = .030$; Fisher's exact $OR = 1.56, p = .026$). No reliable difference emerged between the counter-stereotypical and control conditions. Enjoyment of working with the candidates showed a clearer effect ($\chi^2(2) = 10.20, p = .006$). Jennifer was chosen as the more enjoyable colleague more often in the control and stereotypical conditions (149 and 158, respectively) than in the counter-stereotypical condition (126). Pairwise comparisons confirmed that Jennifer's advantage over John was reduced in the counter-stereotypical condition compared to both control ($\chi^2(1) = 4.32, p = .038$) and stereotypical conditions ($\chi^2(1) = 8.90, p = .003$).

Trait ratings showed a similar tradeoff between competence- and warmth-related impressions (Figure 5). Using ordinary least squares regressions with each trait rating as the dependent variable, we observed that counter-stereotypical suggestions marginally increased perceptions of technical skill relative to control ($\beta = 0.217, p = .093$), whereas stereotypical suggestions reduced them ($\beta = -0.279, p = .031$). Effects on other competence-related traits (experience, assertiveness/confidence, leadership) were smaller and not statistically reliable (all $|\beta| < 0.20$, all $p > .15$).

Warmth-related traits showed clearer effects. Counter-stereotypical suggestions reduced perceptions of warmth ($\beta = -0.393, p = .002$) and friendliness ($\beta = -0.304, p = .007$), while stereotypical suggestions increased both warmth ($\beta = 0.353, p = .005$) and friendliness ($\beta = 0.422, p < .001$). Teamwork and respect showed weaker trends in the same direction (teamwork: $\beta = -0.239, p = .051$; respect: $\beta = -0.152, p = .098$), though these effects did not reach significance thresholds. Together, these patterns indicate that counter-stereotypical suggestions increased competence-related impressions slightly while reliably decreasing warmth-related evaluations.

Overall, these results suggest that counter-stereotypical suggestions improved Jennifer's standing as a trusted leader, suggesting shifts in participants' written evaluations carried over into measurable and consistent changes in their attitudes. However, these suggestions simultaneously reduced affiliative judgments, making her appear personally less likeable. While we cannot determine causality from current data, the divergence between competence-related and affiliative judgments is consistent with a backlash-like pattern where woman displaying competence/agentic traits are seen as less warm and likeable.

4.4 Did the AI writing assistant affect participants' hiring decisions and recommended salary?

We first examined whether participants' hiring choices varied by condition. A chi-squared test with candidate chosen (Jennifer vs. John) as the dependent variable and experimental condition as the independent variable revealed no reliable differences across conditions, $\chi^2(2, N = 665) = 1.44, p = .488$. The corresponding effect size was very small (Cramer's $V = 0.046$). As shown in Figure 6, Jennifer was selected by 44% of participants in the counter-stereotypical condition, 41% in the control condition, and 39% in the stereotypical condition, but these differences were not statistically distinguishable.

Salary recommendations showed a clearer treatment effect. In the control condition, participants offered Jennifer significantly lower salaries than John (paired $t(223) = -3.18, p = .002$; Wilcoxon $W = 4181.5, p = .001$). A similar gap emerged in the stereotypical condition (paired $t(222) = -2.26, p = .025$; Wilcoxon $W = 3587.5, p = .007$). By contrast, in the counter-stereotypical condition the difference between Jennifer's and John's salaries was not statistically significant (paired $t(222) = -1.09, p = .278$; Wilcoxon $W = 4854.5, p = .165$).

Overall, we found participants remained more likely to hire John than Jennifer, even when writing with AI suggesting competent traits. However, these suggestions narrowed salary gaps: when nudged to describe Jennifer as competent, participants offered her the same starting salary as John –– an effect we did not see in other conditions.

4.5 Did the AI writing assistant affect how participants explain their decision?

To complement our quantitative analyses, we conducted a qualitative read-through of participants' open-ended hiring justifications. Below, we present illustrative excerpts that highlight how participants reasoned about their choices.

When the assistant suggested competence-oriented descriptors, participants who chose Jennifer framed their decision around analytical skills, leadership, confidence, and her experience as a research assistant:

> "I chose Jennifer because I think she might be a better fit for the analytical criteria. Although she may appear to keep to herself, I see this as a value."

> "I chose to hire Jennifer due to her strong analytical skills and her relevant experience as a research assistant and Financial Analyst intern. Her steady improvement and leadership make her a strong candidate for the job."

> "I chose to hire Jennifer because she demonstrates persistence, accuracy, and confidence—along with strong leadership and mentoring experience. This makes her well prepared to succeed in this role."

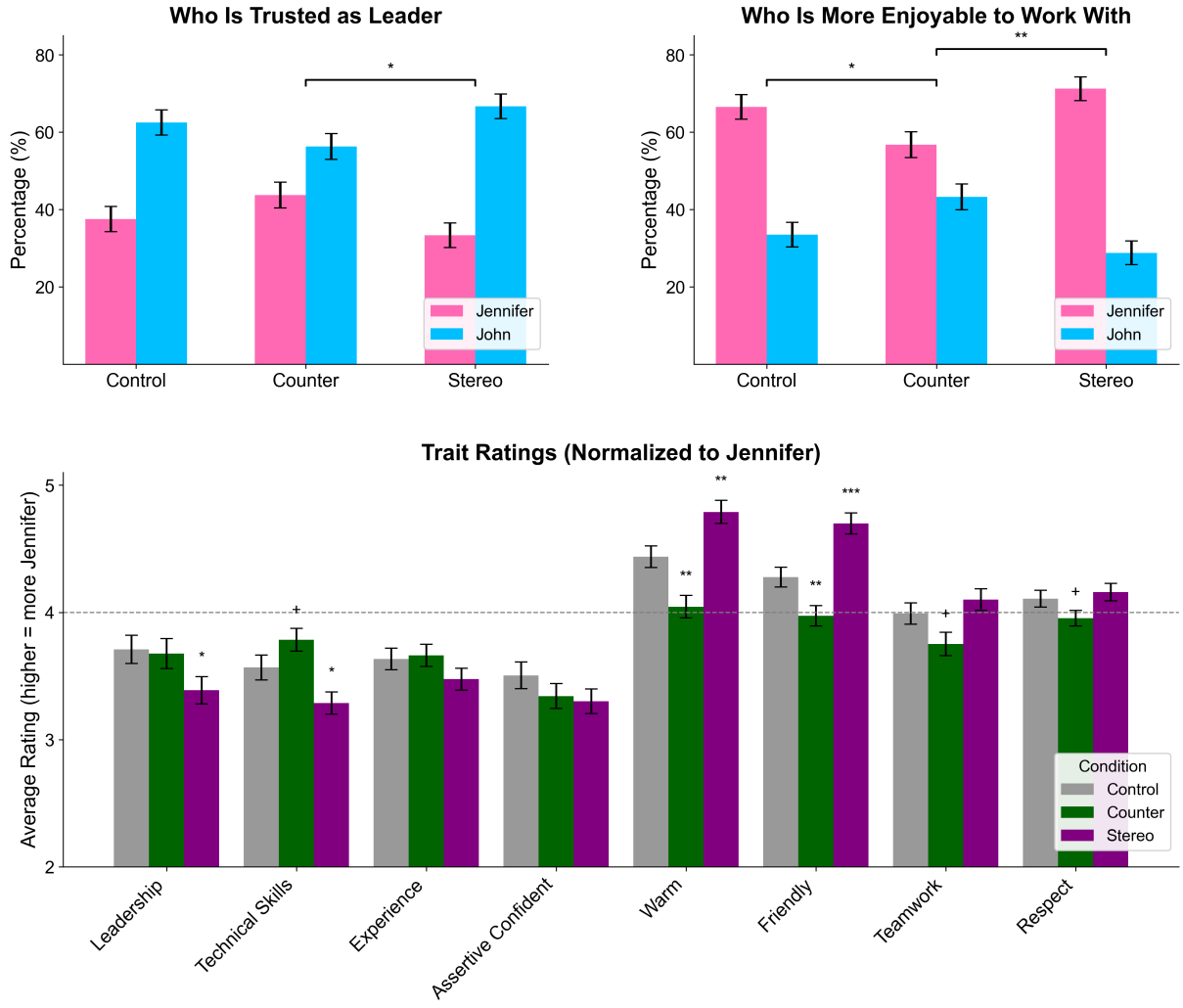


Figure 5: Gendered perceptions of competence and warmth traits across conditions. (Top): Affiliative decision outcomes comparing preferences for "Jennifer" (pink) versus "John" (blue) across control, counter, and stereotype conditions. (A) shows who is trusted as a leader; (B) shows who is rated as more enjoyable to work with. (C) Average trait ratings (normalized relative to Jennifer) across leadership, technical skills, experience, assertive/confident, warm, friendly, teamwork, and respect. Bars represent condition means (gray = control, blue = counter, pink = stereotype) with error bars showing standard error.

When the auto-completion assistant suggested stereotypically feminine descriptors, participants emphasized interpersonal qualities, writing about how they found her personable, and how she might collaborate and uplift teammates.

> "I chose Jennifer because her warm nature is a good boost for morale and to help the team succeed and want to do well. I think she will bring people together and make them want to perform well."

> "I can choose to hire Jennifer because my spirit aligns better with her and I consider her résumé more outstanding and I believe she would do better in this role. Although John isn't bad himself, but I generally prefer Jennifer."

> "I chose Jennifer because I think that for an entry level position she will fit in better with the team. She is warm and friendly and easy to get along with."

Some responses revealed the pathways through which language suggestions may shape downstream judgments. For example, one participant in the competence-oriented condition explicitly tied their hiring choice to earlier trait ratings:

> "I chose to hire Jennifer because when presented with the previous questions, I found myself leaning more towards her side of the scale for most of the questions that are relevant to success in the role."

This reasoning aligns with our finding (Figure 5) that the writing assistant intervention shifted trait impressions. During the trait

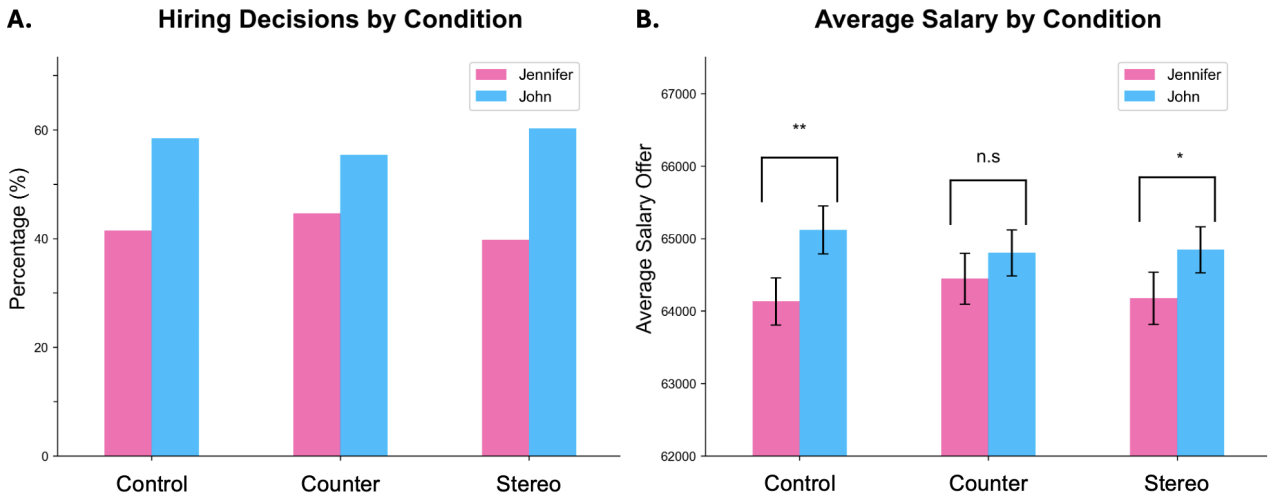


Figure 6: : Hiring and salary decisions by condition. (A): Hiring decisions for Jennifer (pink) and John (blue) across conditions. Jennifer was chosen most often in the counter-stereotypical condition (44.6%), somewhat less in the control (41.5%), and least in the stereotypical condition(39.7%). Despite this directional pattern, chi-square tests indicated that differences across conditions were not statistically significant. (B): Average salary offers for Jennifer and John. In the control condition, Jennifer was offered significantly lower salaries than John (M = 64, 132*vs.*65,116, p = .002). A similar significant gap was observed in the stereotypical condition (M = 64, 174*vs.*64,845, p = .025). In the counter-stereotypical condition, the salary gap narrowed and was no longer statistically significant (M = 64, 446*vs.*64,802, p = .278). These results suggest that hiring preferences were resistant to change, but counter-stereotypical interventions effectively reduced gender disparities in salary recommendations.

perception surveys, the two pieces of writing were displayed on the screen for participants. Thus, for this participant, a plausible interpretation is the gendered language from auto-completions translated into trait perceptions when their were referencing their writing to rate traits, which then impacted their hiring decision.

Another hiring justification in the control condition exemplifies how gendered stereotypes may be the default. One participant who hired John explained:

> "I chose John for this type of industry because I personally believe while Jennifer has the skills necessary ... she is too kind for this type of industry which does need people who are less interested in compromise and collaboration and more interested in authoritative posturing to get the job done. It is really a blessing in disguise for her and as well for him, his rough background would be a good match for this type of job as it is about asserting oneself ..."

Here, the participant interpreted identical résumés in highly gendered terms, construing Jennifer's warmth as a liability and attributing competence to John through imagined biographical details. The contrastive phrasing (*too kind* versus *rough background*) shows how gendered expectation casts the same qualifications in opposite lights. This participant also emphasized *this type of industry*, again showing how role congruity still presents a barrier for women. This response illustrates how the stereotypical traits that participants emphasize shape the very narratives they construct to rationalize their hiring decision.

These excerpts illustrate how the language model shaped the framing of participants' justifications: competence-related language appeared in the counter-stereotypical condition, warmth-related language in the stereotypical condition, and gendered reasoning emerged even in control. These qualitative examples complement our quantitative results by showing, in participants' own words, how subtle differences in language suggestions could influence the way they explained and rationalized their hiring choices.

4.6 Suggestion Exposure and Interaction Behavior

Participants engaged extensively with the autocompletion system. On average, they were shown 12.71 sets of suggestions for John (SD = 7.83) and 13.43 sets of suggestions for Jennifer (SD = 14.35), both well above the minimum exposure requirement. These averages suggest that users generally engaged with the system even beyond the mandated level, though this interpretation should be made cautiously. Acceptance rates were high: participants accepted 77% of suggestions for John (SD = 0.27) and 73% for Jennifer (SD = 0.27). Post-insertion editing was relatively rare, with an edit rate of 10% for John (SD = 0.20) and 12% for Jennifer (SD = 0.23). The resulting texts incorporated a substantial amount of AI-generated language: 41% of words in John evaluations (SD = 0.26) and 39% of words in Jennifer evaluations (SD = 0.26) came from accepted completions. Word counts were similar across candidates (John: M = 166.75, SD = 47.00; Jennifer: M = 165.77, SD = 70.01). None of these engagement metrics differed significantly across candidates

or experimental conditions. Histograms of interaction metric are provided in Appendix E

We conducted dose-response analyses to examine whether individual differences in engagement with the system were associated with downstream judgments. We estimated OLS and logistic regression models treating keystroke-based metrics (suggestions seen, accepted, edited, and reading times) as continuous predictors of Jennifer's salary offer and the likelihood of hiring Jennifer over John. We did not observe consistent or robust associations between these interaction measures and evaluative outcomes. Details are provided in Appendix D.

In an exploratory analysis, we observed a directional difference in the average reaction time spent inspecting suggestions for Jennifer across conditions. Participants in the control condition spent 8,933 ms (SD = 11456 ms) on average reviewing suggestions, compared to 8,207 ms (SD = 5396 ms) in the stereotypical condition and 11,023 ms (SD = 31843 ms) in the counter-stereotypical condition, suggesting that counter-stereotypical suggestions for Jennifer may have prompted additional deliberation or surprise. This increased processing time may indicate that participants were actively reconciling unexpected characterizations with their initial impressions, potentially considering Jennifer in ways that challenged their default stereotypical mental models.

5 Discussion

This study shows that AI-assisted auto-completion writing tools have the potential to change gender stereotypes through direct manipulations of language. Specifically, our intervention systematically altered the descriptive language that participants used for John and Jennifer along competence and warmth dimensions. These linguistic shifts carried downstream effects. Counter-stereotypical completions increased Jennifer's likelihood of being selected as a trusted leader but reduced likability. Hiring decisions, though not statistically significant, followed the expected direction. Relative to herself, Jennifer was chosen the most often in the counter-stereotypical condition and least in the stereotypical condition, while out of the two candidates, John was favored across the conditions. Salary recommendations showed the clearest intervention effect: Jennifer received lower offers than John in control and stereotypical conditions, but the gap disappeared in the counter-stereotypical condition.

5.1 Intervention Effectiveness

Counter-stereotypical completions were successful in eliminating the salary gap between Jennifer and John, but less so in changing the hiring decision. One explanation is that salary judgments are continuous and independent across candidates, in other words, participants could reward Jennifer with higher pay without penalizing John. The hiring decision, by contrast, required choosing one candidate, creating a zero-sum trade-off [54]. This resilience of hiring decisions despite linguistic and attitudinal shifts aligns with prior research on the attitude-behavior gap [78], showing that changing implicit associations and explicit attitudes sometimes fail to translate to behavioral change. Furthermore, even if participants recognized Jennifer's competence, role congruity theory [63] suggests that she have been less likable due to violating gender norms. This is evidenced in participants trusting Jennifer as leader more in the counter-stereotypical condition, but finding her less enjoyable to work with. This pattern potentially indicates gender backlash and may explain why salary decisions changed but not hiring: participants acknowledge her skills with higher pay (while still giving John a directionally higher salary), but still prefer John when forced to pick one candidate due to decreased likability. The salary question was framed such that participants were asked to set salaries "if both candidates were hired," which may have made it easier to hire based on objective skill and experience impression. In contrast, the hiring choice was more subjective and depended on perceived fit, which may have amplified effects of role congruity conflict [63]. Overall, these findings suggest that our interventions can reduce gender bias in continuous and independent evaluations but face greater barriers in zero-sum choices between two candidates.

5.2 AI autocomplete writing assistants as bias mitigation tools

We show that autocomplete writing assistants can influence people's judgments and decisions through subtle changes in descriptive language. Rather than being explicitly opinionated [109, 197], injecting egalitarian statements, or presenting overt counter-stereotypes, our intervention provided short descriptive phrases in ways that participants naturally integrated into their writing. This draws on a well-established principle in social cognition: everyday linguistic habits are a primary pathway through which stereotypes become expressed, reinforced, and made psychologically available [29, 120, 137]. Shifting the linguistic cues people rely on therefore offers a direct and subtle way to influence how they mentally represent others. Unlike interventions that focus on raising awareness of gender inequality or exposing counter-stereotypical role models [181], our work directly manipulated the mediator: language [85]. By changing the words people used to evaluate Jennifer, we were able to at least temporarily alter their cognitive representations. In this sense, our system offers an alternative way to think about stereotype-intervention design.

Our prototype system demonstrates that AI writing assistants can serve as low-cost, scalable stereotype interventions that work by shaping language at the point of production. Importantly, our intervention shifted participants' use of competence and warmth descriptors without changing the overall valence of language, demonstrating that stereotype-relevant traits are malleable through subtle linguistic cues. Because our intervention operates through ordinary word choice rather than overt persuasion, it may avoid some of the reactance and fatigue associated with more explicit approaches such as diversity training [114].

Rather than demanding deliberate monitoring or suppression of biased thoughts, our approach instead provides subtle shifts in language that work at the same level of processing as the stereotypes themselves. Rather than requiring users to detect and override bias in the moment, the system provides an alternative linguistic default to the judgment process.

By leveraging their capability to influence impression-formation, systems like our prototype can be developed for social good in broader contexts. Beyond hiring evaluations, this principle extends to many domains involving social inference based on a limited set

of trait perceptions. For example, two domains in which gender bias is particularly salient are recommendation letters [140] and students' evaluation of professors [18]. The approach could also address stereotypes along other dimensions including race, age, accent, and intersectional identities [16, 45], though each would require tailored interventions accounting for distinct stereotype structures. In female-dominated fields, interventions might counter different stereotypes (e.g., assumptions about men lacking warmth, being selfish, or other devaluation of communal skills).

While our work highlights the potential of linguistic changes in interrupting stereotype bias reproduction, it represents only one point in a broader design space of bias-mitigation strategies. Prior work has explored more explicit interface techniques such as warnings, flagged terms, and explanation-based cues that highlight potentially biased language and suggest alternatives [69, 123]. These approaches make bias visible, but they also have important limitations: they tend to flag only overtly biased wording rather than prompting users to generate counter-stereotypical descriptions, and rely on trained classifiers that may fail to detect biased or harmful language. For example, Microsoft's Inclusivity Suggestion tool has been found to not flag language such as common LGBTQIA+ slurs as sexual orientation bias [69]. They also can mis-classify reclaimed language as biased: for example, the term "cripple" has been reclaimed by the disabled community, but is flagged by Grammarly's inclusive style guide [69]. Furthermore, they require careful design and personalization as users vary widely in preferred levels of explanatory depth, with explanations that feel misaligned or unwarranted sometimes triggering reactance or disengagement [61]. Explicit bias warnings can therefore introduce cognitive load or unintended reactivity. Our findings complement these approaches by showing that less obtrusive, language-level nudges can shift impressions even without explanation and related friction. At the same time, even subtle nudges can sometimes prompt reactivity if users sense that their language is being steered. We do not interpret our findings as robust evidence of such effects, but future systems should remain attentive to these risks when deploying counter-stereotypical suggestions in real-world settings to prevent triggering backlash from gender role incongruity [168, 169].

We note here that the scope of this work is limited to investigating whether AI writing tools can reduce existing stereotype biases in human decision-making. More broadly, we position our work as demonstrating the potential of AI writing assistants in changing stereotypical mental defaults across social perception and evaluation contexts, even those in which decision-making is not involved. We do not make any claims about whether purely human, human and AI in cooperation, or purely AI decision-makers are more or less biased, although recent work suggests that AI displays comparable or greater biases as humans [5, 179].

5.3 Ethical and Practical Considerations

Our intervention raises ethical questions of transparency, user agency, and appropriate deployment contexts. In this section, we situate our work within debates about implicit debiasing interventions and propose deployment models that attempt to balance effectiveness with ethical principles.

5.3.1 Agency and Implicit Interventions. AI-assisted writing systems can be understood along a spectrum of autonomy. At one end, fully human decision-making affords maximal agency but is vulnerable to biases and entrenched maladaptive habits [81, 114]. At the other, fully automated systems may reduce some forms of bias but diminish human control, introduce new algorithmic harms, and remove opportunities for deliberation [103]. Our approach occupies an intermediate point on this continuum: it preserves high user agency by keeping humans fully responsible for judgments, while aiming to nudge linguistic habits in less gender-stereotypical directions.

Nudges and their benefits. Our intervention can be understood as a form of "nudge," a class of subtle environmental or interface changes that guide behavior without restricting choice [182, 186]. Common examples include making unhealthy foods slightly harder to reach in cafeterias [166] or defaulting to double-sided instead of single-sided printing [67]. We adopt this principle by shifting the linguistic defaults that users encounter during writing. However, we note that the analogy is not perfect: unlike classic nudges that preserve the full action space, counter-stereotypical suggestions *shape* the set of options a user is most likely to see, limiting users from utilizing the full range of AI suggestions that they could have otherwise used in their writing. Future work could address this by showing both neutral and counter-stereotypical suggestions so users may experience the full breadth of possible AI suggestions, thus increasing autonomy while still supporting bias-reduction. Still, we see considerable overlap between our approach and nudges, since both subtly influence how users act while preserving their ability to opt out. In many cases, these forms of support can help people follow through on intentions that are difficult to enact consistently. Just as healthier eating is easier to maintain when the environment reinforces that goal, reducing reliance on gender stereotypes can be easier when the environment provides scaffolds that align with users' stated values. In this sense, our intervention, like many nudges, aims to help users act in ways they already endorse, even when those goals are hard to uphold through conscious effort alone.

Concerns about nudges and user agency. Nudges also raise ethical concerns around autonomy and manipulation. Agency is often defined as the capacity to act according to one's values, intentions, and desires [22, 79], and critics argue that influencing behavior without full awareness can cross the line into coercion [93]. These questions are especially salient for AI writing assistants, where the boundary between tool and collaborator is often ambiguous [127]. Users may be unsure whether the system is merely assisting or subtly shaping their evaluative language, and different users prefer different levels of transparency or explanation.

Mitigating agency concerns. To address these issues, our design preserves key dimensions of human control: users retain full control over when to request AI suggestions, whether to accept, modify, or ignore those suggestions, and how to make final hiring decisions. Suggestions are valence-matched to the user's own wording, and accepting any suggestion is optional. At the same time, we recognize steering adjective choice in writing inevitably introduces tensions around user autonomy, particularly when it influences decision-making in subtle ways. We acknowledge that our intervention involves trade-offs between promoting equity and preserving subjective agency in relation to human-centered AI principles [178].

These tensions highlight the need for carefully considered, context-dependent deployment strategies that balance transparency, user comfort, and debiasing effectiveness. We elaborate these strategies in the following section.

5.3.2 Transparency and Context-Dependent Deployment. A tension in implicit interventions research is that increasing transparency can reduce effectiveness [38, 112], while opacity raises agency concerns. We suggest this tradeoff may be context-dependent, and discuss concrete contexts where our system may be implemented:

- **Training contexts (full transparency)**. Our intervention may be appropriate in bias training programs. Meta-analyses show active, habit-breaking interventions outperform passive awareness-raising [58, 151]. Participants could practice counter-stereotypical language use with our system and be debriefed with full transparency about the intervention's purpose and reflect on changes post-training. Participants could compare their notes with and without using the system, and be encouraged to use such tools afterwards. This maximizes both agency and transparency while potentially supporting internalized change.
- **Organizational deployment (informed opt-in)**. Hiring managers using AI writing assistants in practice could practice informed opt-in: users are notified the system may offer suggestions designed to counteract prevalent gender stereotypes, with the option to reject or stop suggestions at any time. This mirrors existing professional tools (e.g., Grammarly's tone adjustments, inclusivity checkers) where users opt in knowing the tool will shape their expression without knowing which exact moments it occurs. Some effectiveness may be lost through transparency, though emerging evidence suggests certain nudges remain effective even when transparent [27, 38, 133]).
- **Individual use (full transparency and control)**. Users seeking to reduce their own biases could opt in with full transparency, similar to habit formation tools or writing assistants with customizable style preferences [199]. Our system could be augmented with tools enabling reflection [28, 34] and visualization of usage over time [105, 176] to show how linguistic habits have shifted.

5.3.3 Practical implementation. Implementation could follow multiple pathways and stages. Companies could invite employees who already use AI or auto-complete like tools in their daily workflow to integrate our system in their writing. Given the relatively shortness of our suggestions, our system could be integrated in writings of any length, ranging from full-evaluations to quick emails. Organizations could develop or adapt their own models to be aligned with specific DEI commitment or task contexts (e.g., hiring evaluation, performance review, internal presentations) to allow further customization [38]. Regardless of approach, questions of governance remain: who determines which language patterns to counter, who ensures the quality and adherence of these models' behaviors, how models adapt as stereotypes evolve [62], and how responsibility is distributed when biased outcomes occur despite AI assistance [103].

5.4 Limitations and future directions

This study has several limitations, many of which arise from deliberate design decisions made to balance experimental control, realism, and feasibility. Because we position this work as a prototype system, our aim was to illustrate the potential of scalable, language-driven stereotype interventions rather than to deliver a fully production-ready solution. We describe these design choices and the limitations they induce below, and motivate directions for future work.

Study design. Our design manipulated suggestions for the woman candidate while keeping those for the male candidate neutral, reflecting real-world asymmetry in how gender stereotypes disproportionately disadvantage women in male-typed roles, which are often high-prestige, high-power, and high-reward positions. Future work should examine symmetric designs that also vary suggestions for male candidates to assess whether counter-stereotypical nudges operate similarly across gendered expectations. Participants wrote short evaluations under time constraints, which allowed us to maintain experimental consistency and better study the quick, automatic process of stereotyping. But this means there were limited opportunities for richer narrative construction; longer-form writing tasks may reveal different linguistic or cognitive dynamics. We also note that participant gender may influence responses to stereotype-consistent and stereotype-violating information, but the study was not powered to detect such interaction effects; follow-up work should investigate whether men and women differ in how they interpret or react to counter-stereotypical suggestions. Finally, our current design focused on a single career (Financial Analyst); future work should examine how the intervention performs across a broader set of occupations with different gender-typing and evaluative norms.

System Design and Ecological Validity. Our interface provides three autocomplete suggestions in a pop-up box style, following prior work [127, 128] and our pilots, which confirmed three was a balance between providing enough options to choose and not causing information fatigue. We chose to provide short phrases instead of full sentences in Gmail Smart Compose style suggestions to reduce the cognitive load from reviewing suggestions, encourage an active process of integrating suggestions with existing writing, and prevent participants from submitting entirely AI-generated writing [37]. We let the interaction be user initialized (i.e., participants must press TAB to view suggestions) instead of suggestions automatically appearing, to preserve agency. While we did not require participants to accept suggestions, we asked them to view suggestions at least eight times. This ensured enough exposure to counter-stereotypical language but may not reflect naturalistic autocomplete usage. Future work should examine spontaneous acceptance rates and and whether effects persist when users are not required to engage with the tool. Moreover, the realism of autocomplete use in hiring and evaluation settings remains an open question: existing professional writing platforms increasingly incorporate AI-assisted suggestions, but adoption varies widely across organizations and some hiring workflows may not involve writing interfaces where autocomplete plays a central role. As such, our findings speak to the potential influence of autocomplete when it

is used, but full ecological validity will require studying how hiring professionals actually engage withAI suggestions in their real workflows. Better understanding the attributes of populations who refuse AI writing tools altogether could inform whether alternative interventions could better serve them.

Existing Biases in LLMs. AI writing tools may inadvertently worsen gender biases as AI systems often reflect and spread the biases found in their training data [175, 200]. If trained on biased text or not prompt engineered and tested rigorously, they may suggest words or phrases that repeat or reinforce gender stereotypes [5, 32, 102, 152, 162]. From static word embeddings [17, 33] to contextualized embeddings [59, 152], in historical analyses [80], from word-association games [5] to real-world decisions [183, 193], language models have consistently associated women with lower competence and higher warmth, supporter roles not leaders, or to study humanities not science. Such biases, if they appear during interventions, would legitimize stereotypes instead of reducing them.

Cultural Homogenization and Global South Perspectives. Recent work shows that the adoption of LLM-based writing assistants is associated with shifts towards Western norms, resulting in a measurable decline in linguistic diversity, especially in non-English and non-Western contexts [2, 180]; and large-scale analyses of academic writing show convergence of style across regions, which may be exacerbated by AI-writing tools [159]. In Global South contexts in particular, sensitive use requires co-designing interventions with stakeholders to align with local linguistic norms and sociocultural expectations. Approached this way, our prototype system can help counteract inequities without contributing to the erasure of culturally grounded writing styles.

Generalizability. The résumé review task were completed by online participants who were not professional recruiters, and results may not generalize to real organizational contexts. The design focused on binary gender, a single job type, and one AI system, which also limits generalizability. As our intervention is rooted in stereotype content model and the warmth-competence dimension that underlie universal social impression formation, we are hopeful that it may be successfully extended towards other stereotypes (e.g., race, age, nationality). Longitudinal studies in ecologically valid contexts are also needed to test whether repeated exposure to counter-stereotypical completions produces lasting changes in language and judgment, or whether backlash persists.

Together, these limitations point to important opportunities for future work to refine the intervention and test its effectiveness in real-world settings. Our findings establish a proof-of-concept that AI writing assistants can serve as scalable tools for reducing stereotype expression in high-stakes evaluations, providing a foundation for research on how such language-level interventions can advance equity while preserving user agency.

References

[1] Andrea E Abele, Naomi Ellemers, Susan T Fiske, Alex Koch, and Vincent Yzerbyt. 2021. Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological review* 128, 2 (2021), 290.

[2] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 1–21. https://doi.org/10.1145/3706598.3713564

[3] Dolores Albarracín. 2002. Cognition in persuasion: An analysis of information processing in response to persuasive communications. In *Advances in experimental social psychology*. Vol. 34. Elsevier, 61–130.

[4] 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 (2025), 6037.

[5] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences* 122, 8 (2025), e2416228122.

[6] Yael M Bamberger. 2014. Encouraging girls into science and technology with feminine role model: Does this work? *Journal of Science Education and Technology* 23, 4 (2014), 549–561.

[7] Orly Bareket and Susan T Fiske. 2023. A systematic review of the ambivalent sexism literature: Hostile sexism protects men's power; benevolent sexism guards traditional gender roles. *Psychological bulletin* 149, 11-12 (2023), 637.

[8] Orly Bareket and Susan T Fiske. 2025. Lost opportunities: How gendered arrangements harm men. *Proceedings of the National Academy of Sciences* 122, 5 (2025), e2320788122.

[9] Frederic Charles Bartlett. 1995. *Remembering: A study in experimental and social psychology*. Cambridge university press.

[10] Nicolas F Bauer. 2025. Does ChatGPT Increase Language Homogenization? In *KI in Medien, Kommunikation und Marketing: Wirtschaftliche, gesellschaftliche und rechtliche Perspektiven*. Springer, 11–31.

[11] Alan Benson, Danielle Li, and Kelly Shue. 2024. "Potential" and the Gender Promotions Gap. *SSRN Working Paper* (March 2024). Available at SSRN.

[12] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review* 94, 4 (2004), 991–1013.

[13] Diana E Betz and Denise Sekaquaptewa. 2012. My fair physicist? Feminine math and science role models demotivate young girls. *Social psychological and personality science* 3, 6 (2012), 738–746.

[14] Oloff C Biermann. 2022. *Writers want AI collaborators to respect their personal values and writing strategies: a human-centered perspective on AI co-writing*. Ph. D. Dissertation. University of British Columbia.

[15] Monica Biernat and Kathleen Fuegen. 2001. Shifting standards and the evaluation of competence: Complexity in gender-based judgment and decision making. *Journal of Social Issues* 57, 4 (2001), 707–724.

[16] Monica Biernat and Diane Kobrynowicz. 1997. Gender-and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology* 72, 3 (1997), 544–557. https://doi.org/10.1037/0022-3514.72.3.544

[17] 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* 29 (2016).

[18] Anne Boring. 2017. Gender biases in student evaluations of teaching. *Journal of public economics* 145 (2017), 27–41.

[19] Lera Boroditsky, Lauren A Schmidt, and Webb Phillips. 2003. Sex, syntax, and semantics. *Language in mind: Advances in the study of language and thought* 22, 61-79 (2003), 1.

[20] Melissa Bowerman. 1973. Structural relationships in children's utterances: Syntactic or semantic? *Cognitive development and acquisition of language* (1973), 197–213.

[21] Ryan L Boyd and James W Pennebaker. 2017. Language-based personality: A new approach to personality in a digital world. *Current opinion in behavioral sciences* 18 (2017), 63–68.

[22] Michael E Bratman. 2007. Structures of agency. In *Structures of Agency: Essays*. Oxford University Press, 1–31. https://doi.org/10.1093/acprof:oso/9780195187717.003.0001

[23] Thomas Breda, Julien Grenet, Marion Monnet, and Clémentine Van Effenterre. 2023. How effective are female role models in steering girls towards STEM? Evidence from French high schools. *The Economic Journal* 133, 653 (2023), 1773–1809.

[24] Victoria L Brescoll, Tyler G Okimoto, and Andrea C Vial. 2018. You've come a long way… maybe: How moral emotions trigger backlash against women leaders. *Journal of social issues* 74, 1 (2018), 144–164.

[25] Jeff E Brooks-Harris, Martin Heesacker, and Cristina Mejia-Millan. 1996. Changing men's male gender-role attitudes by applying the elaboration likelihood model of attitude change. *Sex roles* 35, 9 (1996), 563–580.

[26] Roger W Brown and Eric H Lenneberg. 1954. A study in language and cognition. *The Journal of Abnormal and Social Psychology* 49, 3 (1954), 454.

[27] Hendrik Bruns, Elena Kantorowicz-Reznichenko, Katharina Klement, Maja Luisa Jonsson, and Bilel Rahali. 2018. Can nudges be transparent and yet effective? *Journal of Economic Psychology* 65 (2018), 41–59. https://doi.org/10.1016/j.joep.2018.02.002

[28] Simon Buckingham Shum, Ágnes Sándor, Rosalie Goldsmith, Xiaolong Wang, Randall Bass, and Mindy McWilliams. 2016. Reflecting on reflective writing analytics: Assessment challenges and iterative evaluation of a prototype tool. In

Proceedings of the sixth international conference on learning analytics & knowledge. 213–222.

[29] Christian Burgers and Camiel J Beukeboom. 2020. How language contributes to stereotype formation: Combined effects of label types and negation use in behavior descriptions. *Journal of Language and Social Psychology* 39, 4 (2020), 438–456.

[30] Jill Burstein and Martin Chodorow. 1999. Automated essay scoring for nonnative English speakers. In *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing-ASSESSEVALNLP'99*.

[31] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[32] Maarten Buyl, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jefrey Lijffijt, et al. 2024. Large language models reflect the ideology of their creators. *arXiv preprint arXiv:2410.18417* (2024).

[33] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[34] Dashiel Carrera and Sang Won Lee. 2022. Watch me write: exploring the effects of revealing creative writing process through writing replay. In *Proceedings of the 14th Conference on Creativity and Cognition*. 146–160.

[35] Peter Carruthers. 2002. The cognitive functions of language. *Behavioral and brain sciences* 25, 6 (2002), 657–674.

[36] Edward H Chang, Katherine L Milkman, Dena M Gromet, Robert W Rebele, Cade Massey, Angela L Duckworth, and Adam M Grant. 2019. The mixed effects of online diversity training. *Proceedings of the National Academy of Sciences* 116, 16 (2019), 7778–7783.

[37] Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. 2019. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2287–2295.

[38] Inyoung Cheong, Alicia Guo, Mina Lee, Zhehui Liao, Kowe Kadoma, Dongyoung Go, Joseph Chee Chang, Peter Henderson, Mor Naaman, and Amy X Zhang. 2025. Penalizing Transparency? How AI Disclosure and Author Demographics Shape Human and AI Judgments About Writing. *arXiv preprint arXiv:2507.01418* (2025).

[39] Sapna Cheryan and Hazel Rose Markus. 2020. Masculine defaults: Identifying and mitigating hidden cultural biases. *Psychological Review* 127, 6 (2020), 1022.

[40] Sapna Cheryan and Gregg A Muragishi. 2025. Removing masculine defaults in the hiring process. *Proceedings of the National Academy of Sciences* 122, 14 (2025), e2501630122.

[41] Sapna Cheryan, Victoria C Plaut, Caitlin Handron, and Lauren Hudson. 2013. The stereotypical computer scientist: Gendered media representations as a barrier to inclusion for women. *Sex roles* 69, 1 (2013), 58–71.

[42] Soonja Choi and Melissa Bowerman. 1991. Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. *Cognition* 41, 1-3 (1991), 83–121.

[43] Soonja Choi and Alison Gopnik. 1995. Early acquisition of verbs in Korean: A cross-linguistic study. *Journal of child language* 22, 3 (1995), 497–529.

[44] Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science* 385, 6714 (2024), eadq1814.

[45] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum* 1989 (1989), 139–167.

[46] Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology* 40 (2008), 61–149.

[47] David Wei Dai, Hua Zhu, and Guanliang Chen. 2025. How does interaction with LLM powered chatbots shape human understanding of culture? The need for Critical Interactional Competence (CritIC). *Annual Review of Applied Linguistics* (2025), 1–22.

[48] Valdemar Danry, Pat Pataranutaporn, Matthew Groh, and Ziv Epstein. 2025. Deceptive explanations by large language models lead people to change their beliefs about misinformation more often than honest explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–31.

[49] Nilanjana Dasgupta and Shaki Asgari. 2004. Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of experimental social psychology* 40, 5 (2004), 642–658.

[50] DataUSA. 2024. *Chemists & Materials Scientists*. https://datausa.io/profile/soc/chemists-materials-scientists

[51] DataUSA. 2024. *Chief Executives & Legislators*. https://datausa.io/profile/soc/chief-executives-legislators

[52] DataUSA. 2024. *Childcare Workers*. https://datausa.io/profile/soc/childcare-workers

[53] DataUSA. 2024. *Elementary & Middle School Teachers*. https://datausa.io/profile/soc/elementary-middle-school-teachers

[54] Shai Davidai and Stephanie J Tepper. 2023. The psychology of zero-sum beliefs. *Nature Reviews Psychology* 2, 8 (2023), 472–482.

[55] Jules Davidoff, Ian Davies, and Debi Roberson. 1999. Colour categories in a stone-age tribe. *Nature* 398, 6724 (1999), 203–204.

[56] Jill G de Villiers and Jennie Pyers. 1997. Complementing cognition: The relationship between language and theory of mind. (1997).

[57] Eva Derous and Roland Pepermans. 2019. Gender discrimination in hiring: Intersectional effects with ethnicity and cognitive job demands. *Archives of Scientific Psychology* 7, 1 (2019), 40.

[58] Patricia G Devine, Patrick S Forscher, Anthony J Austin, and William TL Cox. 2012. Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology* 48, 6 (2012), 1267–1278. https://doi.org/10.1016/j.jesp.2012.06.003

[59] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. ACM, 862–872. https://doi.org/10.1145/3442188.3445924

[60] Paramveer S Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping human-AI collaboration: Varied scaffolding levels in co-writing with language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[61] Chiara Di Bonaventura, Michelle Nwachukwu, and Maria Stoica. 2025. Wanted: Personalised Bias Warnings for Gender Bias in Language Models. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. 124–136.

[62] Alice H Eagly and Blair T Johnson. 1990. Gender and leadership style: A meta-analysis. *Psychological bulletin* 108, 2 (1990), 233.

[63] Alice H Eagly and Steven J Karau. 2002. Role congruity theory of prejudice toward female leaders. *Psychological review* 109, 3 (2002), 573.

[64] Alice H Eagly and Antonio Mladinic. 1994. Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European review of social psychology* 5, 1 (1994), 1–35.

[65] Alice H Eagly and Wendy Wood. 1999. The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American psychologist* 54, 6 (1999), 408.

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

[67] Johan Egebark and Mathias Ekström. 2016. Can indifference make the world greener? *Journal of Environmental Economics and Management* 76 (2016), 1–13. https://doi.org/10.1016/j.jeem.2015.11.004

[68] Naomi Ellemers. 2018. Gender stereotypes. *Annual review of psychology* 69, 1 (2018), 275–298.

[69] Houda Elmimouni, Anasuya Sharma, Vidushi Manayra, Man Iao Chan, Yifan Feng, Kata Kyrölä, and Jennifer A Rode. 2024. Whose Values Matter in Persuasive Writing Tools?. In *Proceedings of the Halfway to the Future Symposium*. 1–9.

[70] Paula England. 2010. The gender revolution: Uneven and stalled. *Gender & society* 24, 2 (2010), 149–166.

[71] Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. Only a Little to the Left: A Theory-grounded Measure of Political Bias in Large Language Models. arXiv:2503.16148 [cs.CY] https://arxiv.org/abs/2503.16148

[72] Raquel Fernández. 2013. Cultural Change as Learning: The Evolution of Female Labor Force Participation over a Century. *American Economic Review* 103, 1 (Feb. 2013), 472–500. https://doi.org/10.1257/aer.103.1.472

[73] Leon Festinger and James M Carlsmith. 1959. Cognitive consequences of forced compliance. *The journal of abnormal and social psychology* 58, 2 (1959), 203.

[74] Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology* 82, 6 (2002), 878.

[75] Susan T. Fiske and Shelley E. Taylor. 1991. *Social Cognition* (2nd ed.). McGraw-Hill, New York.

[76] Susan T Tufts Fiske and Shelley E Taylor. 2020. Social cognition: From brains to culture. (2020).

[77] Jerry A Fodor. 1975. *The language of thought*. Vol. 5. Harvard university press.

[78] Patrick S Forscher, Calvin K Lai, Jordan R Axt, Charles R Ebersole, Michelle Herman, Patricia G Devine, and Brian A Nosek. 2019. A meta-analysis of procedures to change implicit measures. *Journal of personality and social psychology* 117, 3 (2019), 522.

[79] Harry G Frankfurt. 1971. Freedom of the will and the concept of a person. *The Journal of Philosophy* 68, 1 (1971), 5–20. https://doi.org/10.2307/2024717

[80] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[81] Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology* 101, 1 (2011), 109.

[82] Susan A Gelman. 2003. *The essential child: Origins of essentialism in everyday thought.* Oxford University Press.

[83] Susan A Gelman and John D Coley. 1991. *Language and categorization: The acquisition of natural.* Cambridge: Cambridge University Press.

[84] Dedre Gentner. 1978. On relational meaning: The acquisition of verb meaning. *Child development* (1978), 988–998.

[85] Dedre Gentner and Susan Goldin-Meadow. 2003. Language in mind: Advances in the study of language and thought. (2003).

[86] Carol Gilligan. 1993. *In a different voice: Psychological theory and women's development.* Harvard university press.

[87] Jessica R Gladstone and Andrei Cimpian. 2021. Which role models are effective for which students? A systematic review and four recommendations for maximizing the effectiveness of role models in STEM. *International journal of STEM education* 8, 1 (2021), 59.

[88] Peter Glick and Susan T Fiske. 1997. Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of women quarterly* 21, 1 (1997), 119–135.

[89] Stephen Jay Gould. 1996. *Mismeasure of man.* WW Norton & company.

[90] Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics* 3 (1975), 43–58.

[91] G. Stanley Hall. 1904. *Adolescence: Its Psychology and Its Relations to Physiology, Anthropology, Sociology, Sex, Crime, Religion, and Education.* Vol. I–II. D. Appleton and Company, New York. 589, 784 pages.

[92] David L Hamilton and Jeffrey W Sherman. 2014. Stereotypes. In *Handbook of social cognition.* Psychology Press, 1–68.

[93] Daniel M Hausman and Brynn Welch. 2010. Debate: To nudge or not to nudge. *Journal of Political Philosophy* 18, 1 (2010), 123–136. https://doi.org/10.1111/j.1467-9760.2009.00351.x

[94] Joyce C He and Sonia K Kang. 2025. Debiasing job ads by replacing masculine language increases gender diversity of applicant pools. *Proceedings of the National Academy of Sciences* 122, 7 (2025), e2409854122.

[95] Eleanor Rosch Heider and Donald C Olivier. 1972. The structure of the color space in naming and memory for two languages. *Cognitive psychology* 3, 2 (1972), 337–354.

[96] Madeline E Heilman. 2012. Gender stereotypes and workplace bias. *Research in organizational Behavior* 32 (2012), 113–135.

[97] Madeline E Heilman, Suzette Caleo, and Francesca Manzi. 2024. Women at work: Pathways from gender stereotypes to gender bias and discrimination. *Annual Review of Organizational Psychology and Organizational Behavior* 11, 1 (2024), 165–192.

[98] Sarah D Herrmann, Robert Mark Adelman, Jessica E Bodford, Oliver Graudejus, Morris A Okun, and Virginia SY Kwan. 2016. The effects of a female role model on academic performance and persistence of women in STEM courses. *Basic and Applied Social Psychology* 38, 5 (2016), 258–268.

[99] Gail D Heyman and Jessica W Giles. 2006. Gender and psychological essentialism. *Enfance* 58, 3 (2006), 293–310.

[100] E Tory Higgins and William S Rholes. 1978. "Saying is believing": Effects of message modification on memory and liking for the person described. *Journal of Experimental Social Psychology* 14, 4 (1978), 363–378.

[101] James L Hilton and William Von Hippel. 1996. Stereotypes. *Annual review of psychology* 47, 1 (1996), 237–271.

[102] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature* 633, 8028 (2024), 147–154.

[103] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–16. https://doi.org/10.1145/3290605.3300830

[104] Thomas M Holtgraves and Yoshihisa Kashima. 2008. Language, meaning, and social cognition. *Personality and Social Psychology Review* 12, 1 (2008), 73–94.

[105] Md Naimul Hoque, Tasfia Mashiat, Bhavya Ghai, Cecilia D Shelton, Fanny Chevalier, Kari Kraus, and Niklas Elmqvist. 2024. The HaLLMark effect: Supporting provenance and transparent use of large language models in writing with interactive visualization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* 1–15.

[106] Julie Hui and Michelle L Sprouse. 2023. Lettersmith: Scaffolding written professional communication among college students. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–17.

[107] Angel Hsing-Chi Hwang, Q Vera Liao, Su Lin Blodgett, Alexandra Olteanu, and Adam Trischler. 2025. 'It was 80% me, 20% AI': Seeking Authenticity in Co-Writing with Large Language Models. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–41.

[108] Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *arXiv preprint arXiv:2211.05030* (2022).

[109] 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.* 1–15.

[110] Irving L Janis and Bert T King. 1954. The influence of role playing on opinion change. *The journal of abnormal and social psychology* 49, 2 (1954), 211.

[111] Aparna Joshi, Jooyeon Son, and Hyuntak Roh. 2015. When can women close the gap? A meta-analytic test of sex differences in performance and rewards. *Academy of Management Journal* 58, 5 (2015), 1516–1545.

[112] Janice Y Jung and Barbara A Mellers. 2016. American attitudes toward nudges. *Judgment and Decision Making* 11, 1 (2016), 62–74.

[113] Cheryl R Kaiser, Brenda Major, Ines Jurcevic, Tessa L Dover, Laura M Brady, and Jenessa R Shapiro. 2013. Presumed fair: ironic effects of organizational diversity structures. *Journal of personality and social psychology* 104, 3 (2013), 504.

[114] Alexandra Kalev, Frank Dobbin, and Erin Kelly. 2006. Best practices or best guesses? Assessing the efficacy of corporate affirmative action and diversity policies. *American sociological review* 71, 4 (2006), 589–617.

[115] Yoshihisa Kashima, Simon M Laham, Jennifer Dix, Bianca Levis, Darlene Wong, and Melissa Wheeler. 2015. Social transmission of cultural practices and implicit attitudes. *Organizational Behavior and Human Decision Processes* 129 (2015), 113–125.

[116] Alex Koch, Austin Smith, Susan T Fiske, Andrea E Abele, Naomi Ellemers, and Vincent Yzerbyt. 2024. Validating a brief measure of four facets of social evaluation. *Behavior Research Methods* 56, 8 (2024), 8521–8539.

[117] Alex Koch, Vincent Yzerbyt, Andrea Abele, Naomi Ellemers, and Susan T Fiske. 2021. Social evaluation: Comparing models across interpersonal, intragroup, intergroup, several-group, and many-group contexts. In *Advances in experimental social psychology.* Vol. 63. Elsevier, 1–68.

[118] Anne M Koenig, Alice H Eagly, Abigail A Mitchell, and Tiina Ristikari. 2011. Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological bulletin* 137, 4 (2011), 616.

[119] Laura Kotovsky and Dedre Gentner. 1996. Comparison and categorization in the development of relational similarity. *Child development* 67, 6 (1996), 2797–2822.

[120] Laura J Kray and Leigh Thompson. 2004. Gender stereotypes and negotiation performance: An examination of theory and research. *Research in organizational behavior* 26 (2004), 103–182.

[121] Per Ola Kristensson and Keith Vertanen. 2014. The inviscid text entry rate and its application as a grand goal for mobile text entry. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services.* 335–338.

[122] 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:2408.03907 [cs.CL] https://arxiv.org/abs/2408.03907

[123] Moreno La Quatra, Salvatore Greco, Luca Cagliero, and Tania Cerquitelli. 2023. Inclusively: An AI-based assistant for inclusive writing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 361–365.

[124] M Asher Lawson, Ashley E Martin, Imrul Huda, and Sandra C Matz. 2022. Hiring women into senior leadership positions is associated with a reduction in gender stereotypes in organizational language. *Proceedings of the National Academy of Sciences* 119, 9 (2022), e2026443119.

[125] Gustave Le Bon. 1879. Recherches anatomiques et mathématiques sur les lois des variations du volume du cerveau et sur leurs relations avec l'intelligence. 2 (1879), 27–104. Title in English: *Anatomical and Mathematical Researches into the Laws of the Variations of Brain Volume and their Relation to Intelligence*.

[126] Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. *Automated grammatical error detection for language learners.* Morgan & Claypool Publishers.

[127] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A Alghamdi, et al. 2024. A design space for intelligent and interactive writing assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* 1–35.

[128] 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.* 1–19.

[129] Alison P Lenton, Martin Bruder, and Constantine Sedikides. 2009. A meta-analysis on the malleability of automatic gender stereotypes. *Psychology of Women Quarterly* 33, 2 (2009), 183–196.

[130] Stephen C Levinson. 1996. Language and space. *Annual review of Anthropology* 25, 1 (1996), 353–382.

[131] Stephen C Levinson et al. 1996. Relativity in spatial conception and description. *Rethinking linguistic relativity* (1996), 177–202.

[132] Pauline Lightbody, Gerda Siann, Ruth Stocks, and David Walsh. 1996. Motivation and attribution at secondary school: The role of gender. *Educational Studies* 22,

1 (1996), 13–25.
[133] George Loewenstein, Cindy Bryce, David Hagmann, and Sachin Rajpal. 2015. Warning: You are about to be nudged. *Behavioral Science & Policy* 1, 1 (2015), 35–42.
[134] Gary Lupyan. 2016. The centrality of language in human cognition. *Language Learning* 66, 3 (2016), 516–553.
[135] Anthony Lyons and Yoshihisa Kashima. 2006. Maintaining stereotypes in communication: Investigating memory biases and coherence-seeking in storytelling. *Asian Journal of Social Psychology* 9, 1 (2006), 59–71.
[136] Anne Maass. 1999. Linguistic intergroup bias: Stereotype perpetuation through language. In *Advances in experimental social psychology*. Vol. 31. Elsevier, 79–121.
[137] Anne Maass, Daniela Salvi, Luciano Arcuri, and Gün R Semin. 1989. Language use in intergroup contexts: The linguistic intergroup bias. *Journal of personality and social psychology* 57, 6 (1989), 981.
[138] I Scott MacKenzie and R William Soukoreff. 2002. Text entry for mobile computing: Models and methods, theory and practice. *Human–Computer Interaction* 17, 2-3 (2002), 147–198.
[139] C Neil Macrae, Galen V Bodenhausen, Alan B Milne, and Jolanda Jetten. 1994. Out of mind but back in sight: Stereotypes on the rebound. *Journal of personality and social psychology* 67, 5 (1994), 808.
[140] Juan M Madera, Michelle R Hebl, and Randi C Martin. 2009. Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology* 94, 6 (2009), 1591–1599. https://doi.org/10.1037/a0016539
[141] Sara McLanahan. 2004. Diverging destinies: How children are faring under the second demographic transition. *Demography* 41, 4 (2004), 607–627.
[142] Katherine L Milkman, Modupe Akinola, and Dolly Chugh. 2012. Temporal distance and discrimination: An audit study in academia. *Psychological science* 23, 7 (2012), 710–717.
[143] Ruth Milkman. 1987. *Gender at work: The dynamics of job segregation by sex during World War II*. University of Illinois Press.
[144] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences* 109, 41 (2012), 16474–16479.
[145] Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. *arXiv preprint arXiv:1702.04066* (2017).
[146] Nabi Nazari, Muhammad Salman Shabbir, and Roy Setiawan. 2021. Application of Artificial Intelligence powered digital writing assistant in higher education: randomized controlled trial. *Heliyon* 7, 5 (2021).
[147] Gandalf Nicolas, Xuechunzi Bai, and Susan T Fiske. 2022. A spontaneous stereotype content model: Taxonomy, properties, and prediction. *Journal of personality and social psychology* 123, 6 (2022), 1243.
[148] U.S. Bureau of Labor Statistics. 2025. Labor Force Participation Rate – Women [LNS11300002]. Retrieved from FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/LNS11300002 Accessed September 11, 2025.
[149] Tyler G Okimoto and Madeline E Heilman. 2012. The "bad parent" assumption: How gender stereotypes affect reactions to working mothers. *Journal of Social Issues* 68, 4 (2012), 704–724.
[150] Maria Olsson and Sarah E Martiny. 2018. Does exposure to counterstereotypical role models influence girls' and women's gender stereotypes and career choices? A review of social psychological research. *Frontiers in psychology* 9 (2018), 2264.
[151] Elizabeth Levy Paluck and Donald P Green. 2009. Prejudice reduction: What works? A review and assessment of research and practice. *Annual Review of Psychology* 60 (2009), 339–367. https://doi.org/10.1146/annurev.psych.60.110707.163607
[152] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193* (2021).
[153] Samantha C Paustian-Underdahl and Lisa Slattery Walker. 2016. Revisiting the beauty is beastly effect: Examining when and why sex and attractiveness impact hiring judgments. *The International Journal of Human Resource Management* 27, 10 (2016), 1034–1058.
[154] Alexandra A Phillips, Catherine R Walsh, Korie A Grayson, Camilla E Penney, Fatima Husain, and Women Doing Science Team. 2022. Diversifying representations of female scientists on social media: A case study from the women doing science Instagram. *Social Media+ Society* 8, 3 (2022), 20563051221113068.
[155] Martin J Pickering and Victor S Ferreira. 2008. Structural priming: a critical review. *Psychological bulletin* 134, 3 (2008), 427.
[156] Steven Pinker. 2007. *The stuff of thought: Language as a window into human nature*. Penguin.
[157] Ritika Poddar, Rashmi Sinha, Mor Naaman, and Maurice Jakesch. 2023. AI writing assistants influence topic choice in self-presentation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–6.
[158] Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. arXiv:2410.24190 [cs.CL] https://arxiv.org/abs/2410.24190
[159] Arjun Prakash, Shruti Aggarwal, Jeevan John Varghese, and Joel John Varghese. 2025. Writing without borders: AI and cross-cultural convergence in academic writing quality. *Humanities and Social Sciences Communications* 12, 1 (2025), 1–11.
[160] Janet Rafner, Blanka Zana, Ida Bang Hansen, Simon Ceh, Jacob Sherson, Mathias Benedek, and Izabela Lebuda. 2025. Agency in Human-AI Collaboration for Image Generation and Creative Writing: Preliminary Insights from Think-Aloud Protocols. *Creativity Research Journal* (2025), 1–24.
[161] Christian Rapp, Otto Kruse, Jennifer Erlemann, and Jakob Ott. 2015. Thesis writer: a system for supporting academic writing. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*. 57–60.
[162] Philip Resnik. 2025. Large language models are biased because they are large language models. *Computational Linguistics* (2025), 1–21.
[163] Cecilia L Ridgeway and Shelley J Correll. 2004. Unpacking the gender system: A theoretical perspective on gender beliefs and social relations. *Gender & society* 18, 4 (2004), 510–531.
[164] Melissa Roemmele and Andrew S Gordon. 2015. Creative help: A story writing assistant. In *International Conference on Interactive Digital Storytelling*. Springer, 81–92.
[165] Myron Rothbart and Marjorie Taylor. 1992. Category labels and social reality: Do we view social categories as natural kinds? (1992).
[166] Paul Rozin, Sydney Scott, Megan Dingley, Jena K Urbanek, Hong Jiang, and Mark Kaltenbach. 2011. Nudge to nobesity I: Minor changes in accessibility decrease food intake. *Judgment and Decision Making* 6, 4 (2011), 323–332.
[167] Laurie A Rudman. 1998. Self-promotion as a risk factor for women: the costs and benefits of counterstereotypical impression management. *Journal of personality and social psychology* 74, 3 (1998), 629.
[168] Laurie A Rudman and Kimberly Fairchild. 2004. Reactions to counterstereotypic behavior: the role of backlash in cultural stereotype maintenance. *Journal of personality and social psychology* 87, 2 (2004), 157.
[169] Laurie A Rudman and Peter Glick. 2001. Prescriptive gender stereotypes and backlash toward agentic women. *Journal of social issues* 57, 4 (2001), 743–762.
[170] Laurie A Rudman and Julie E Phelan. 2008. Backlash effects for disconfirming gender stereotypes in organizations. *Research in organizational behavior* 28 (2008), 61–79.
[171] Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in psychology* 7 (2016), 154379.
[172] Sabine Sczesny, Franziska Moser, and Wendy Wood. 2015. Beyond sexist beliefs: How do people decide to use gender-inclusive language? *Personality and Social Psychology Bulletin* 41, 7 (2015), 943–954.
[173] Gün R Semin. 2000. Agenda 2000—Communication: Language as an implementational device for cognition. *European Journal of Social Psychology* 30, 5 (2000), 595–612.
[174] Reut Shachnai, Tamar Kushnir, and Lin Bian. 2022. Walking in her shoes: Pretending to be a female role model increases young girls' persistence in science. *Psychological Science* 33, 11 (2022), 1818–1827.
[175] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326* (2019).
[176] Antonette Shibani, Ratnavel Rajalakshmi, Faerie Mattins, Srivarshan Selvaraj, and Simon Knight. 2023. Visual Representation of Co-Authorship with GPT-3: Studying Human-Machine Interaction for Effective Writing. *International Educational Data Mining Society* (2023).
[177] Muhammad Shidiq. 2023. The use of artificial intelligence-based chat-gpt and its challenges for the world of education; from the viewpoint of the development of creative writing skills. In *Proceeding of international conference on education, society and humanity*, Vol. 1. 353–357.
[178] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction* 36, 6 (2020), 495–504. https://doi.org/10.1080/10447318.2020.1741118
[179] Aleksandra Sorokovikova, Pavel Chizhov, Iuliia Eremenko, and Ivan P Yamshchikov. 2025. Surface Fairness, Deep Bias: A Comparative Study of Bias in Language Models. *arXiv preprint arXiv:2506.10491* (2025).
[180] Zhivar Sourati, Farzan Karimi-Malekabadi, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Jackson Trager, Ala Tak, Meng Chen, Fred Morstatter, and Morteza Dehghani. 2025. The shrinking landscape of linguistic diversity in the age of large language models. *arXiv preprint arXiv:2502.11266* (2025).
[181] Rebecca Stewart, Breanna Wright, Liam Smith, Steven Roberts, and Natalie Russell. 2021. Gendered stereotypes and norms: A systematic review of interventions designed to shift attitudes and behaviour. *Heliyon* 7, 4 (2021).
[182] Cass R Sunstein and Richard H Thaler. 2003. Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review* 70, 4 (2003), 1159–1202. https://doi.org/10.2307/1600573

[183] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689* (2023).
[184] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
[185] Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, et al. 2024. AI can help humans find common ground in democratic deliberation. *Science* 386, 6719 (2024), eadq2852.
[186] Richard H Thaler and Cass R Sunstein. 2003. Libertarian paternalism. *American Economic Review* 93, 2 (2003), 175–179. https://doi.org/10.1257/000282803321947001
[187] Paul H Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PloS one* 6, 2 (2011), e16782.
[188] Ruth Van Veelen, Belle Derks, and Maaike Dorine Endedijk. 2019. Double trouble: How being outnumbered and negatively stereotyped threatens career outcomes of women in STEM. *Frontiers in psychology* 10 (2019), 150.
[189] Keith Vertanen, Haythem Memmi, Justin Emge, Shyam Reyal, and Per Ola Kristensson. 2015. VelociTap: Investigating fast mobile text entry using sentence-based decoding of touchscreen keyboard input. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 659–668.
[190] Wilhelm Von Humboldt. 1999. *Humboldt:'On language': On the diversity of human language construction and its influence on the mental development of the human species*. Cambridge University Press.
[191] Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. " It Felt Like Having a Second Mind": Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *Proceedings of the ACM on human-computer interaction* 8, CSCW1 (2024), 1–26.
[192] Yixin Wan and Kai-Wei Chang. 2024. White Men Lead, Black Women Help? Benchmarking and Mitigating Language Agency Social Biases in LLMs. *arXiv preprint arXiv:2404.10508* (2024).
[193] 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. arXiv:2310.09219 [cs.CL] https://arxiv.org/abs/2310.09219
[194] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).
[195] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 214–229.
[196] Benjamin Lee Whorf. 1956. Language, thought, and reality: selected writings of….(Edited by John B. Carroll.). (1956).
[197] Sterling Williams-Ceci, Maurice Jakesch, Advait Bhat, Kowe Kadoma, Lior Zalmanson, and Mor Naaman. 2025. Biased AI Writing Assistants Shift Users' Attitudes on Societal Issues. https://doi.org/10.31234/osf.io/mhjn6_v2
[198] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and Ray Lc. 2022. AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In *Joint proceedings of the ACM IUI workshops*, Vol. 10. CEUR-WS Team, 1–11.
[199] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 841–852.
[200] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).
[201] X. Zhao. 2022. Leveraging Artificial Intelligence (AI) Technology for English Writing: Introducing Wordtune as a Digital Writing Assistant for EFL Writers. *RELC Journal* 54, 3 (2022), 890–894. https://doi.org/10.1177/00336882221094089 Original work published 2023.

A Resume Stimuli

We provide the full résumé stimuli used in this study (see Figure 7). Following established practices in gender bias research [144], résumés were designed to reflect high but slightly ambiguous competence, allowing for variability in participant evaluations. For example, although both résumés included awards and 2 years of experience as a research assistant, the GPAs were slightly below average (3.1).

B Candidate Evaluation Questions Shown to Participants

The following questions were presented after participants completed their written evaluations of both candidates. Items were adapted from prior work on gender bias in hiring and assessed perceptions of competence, warmth, and hire-ability (workplace affiliative decisions, hiring choice, salary recommendation). Following [144], we anchored salary recommendations to a realistic industry range $55,000–$80,000 to increase ecological validity.

Trait Impressions

Based on your impressions, rate which candidate fits each statement better. Your evaluations have been provided above to help you rate. 1 = John fits more, 7 = Jennifer fits more, 4 = Equal or Unsure. *(Note: This scale is flipped depending on which resume is shown first).*

Leadership and Competence Traits

(1) The candidate demonstrates strong leadership.
(2) The candidate has the logical and technical skills to succeed.
(3) The candidate has the experience necessary for this role.
(4) Attention check: Please select "5" for this question.
(5) The candidate is assertive and confident.

Warmth and Interpersonal Traits

(1) The candidate is warm, approachable, and easy to work with.
(2) The candidate is friendly and personable.
(3) The candidate values teamwork and contributes to group success.
(4) The candidate demonstrates respect and consideration for colleagues.

Affiliative Decisions

(1) Who would you trust to lead a difficult project? (*binary choice*)
(2) If both were your colleagues, who would you personally find easier and more enjoyable to work with? (*binary choice*)

Hiring Decision

Which candidate would you hire for an entry-level financial analyst position at a financial services firm (e.g., banks, accounting firms, advisory firms)? You can only choose one. Your choice should reflect who you think would be more successful in a financial analyst role and environment. **Please select one candidate.**

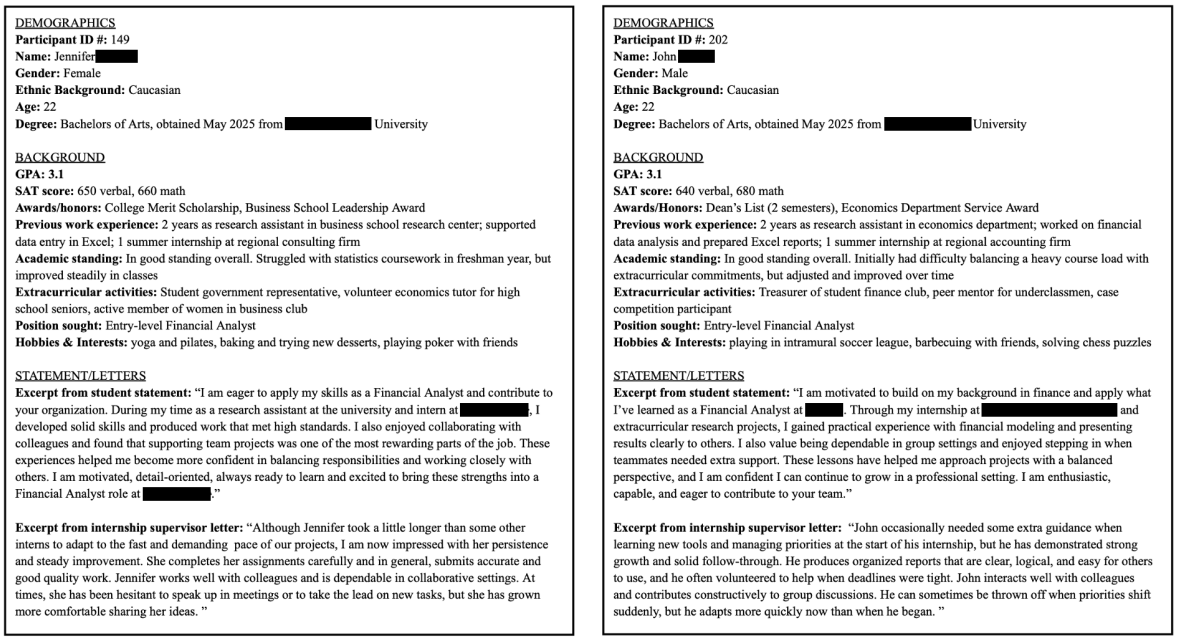


Figure 7: Resume stimuli used in the study.

Open-Ended Justification

Please briefly explain why you chose this job candidate. You may refer to specific phrases, impressions, or overall qualities of the two candidates.

Salary Decision

Entry-level financial analysts at financial services firms typically earn between $55,000 and $80,000 per year, with a median starting salary of $62,000. What is the amount you think each candidate should be paid based on their experience, skills, and competence? *(sliding scale for both candidates with range $55,000–$80,000)*

C Unaggregated Cosine Similarity Gender Bias Score

For parsimony, the main text reports aggregate net bias scores (feminine minus masculine similarity). Here, we present the disaggregated cosine similarity scores for warmth and competence dimensions separately (Figure 8), following the text polarity method from the BOLD framework [59].

D Keystroke Dose-Response Analysis

We conducted dose–response analyses to examine whether individual differences in interaction behavior with the AI writing assistant are associated with downstream evaluations. Five continuous predictors were tested in separate ordinary least squares models predicting Jennifer's salary and in logistic regression models predicting the likelihood of hiring Jennifer over John.

(1) *Suggestions shown* measured how many autocomplete suggestions participants saw.
(2) *Acceptance rate* measured the proportion of suggestions that participants accepted.
(3) *AI post-edit rate* measured the proportion of AI-generated text that were edited after being accepted.
(4) *AI share final* measured the total proportion of the final written evaluation that consisted of AI-generated text.
(5) *Reaction time* measured the average time spent looking at the autocomplete suggestions after pressing TAB.

Across models, effect sizes were small and inconsistent (see Tables 2, 3, and 4). The analyses do not provide evidence that the amount of engagement with the AI writing assistant, as measured by suggestions shown, acceptance rate, AI post-edit rate, AI share final, or mean reading time, systematically influenced downstream judgments. This result is consistent with the possibility that simply encountering counter-stereotypical language is sufficient to shift evaluations.

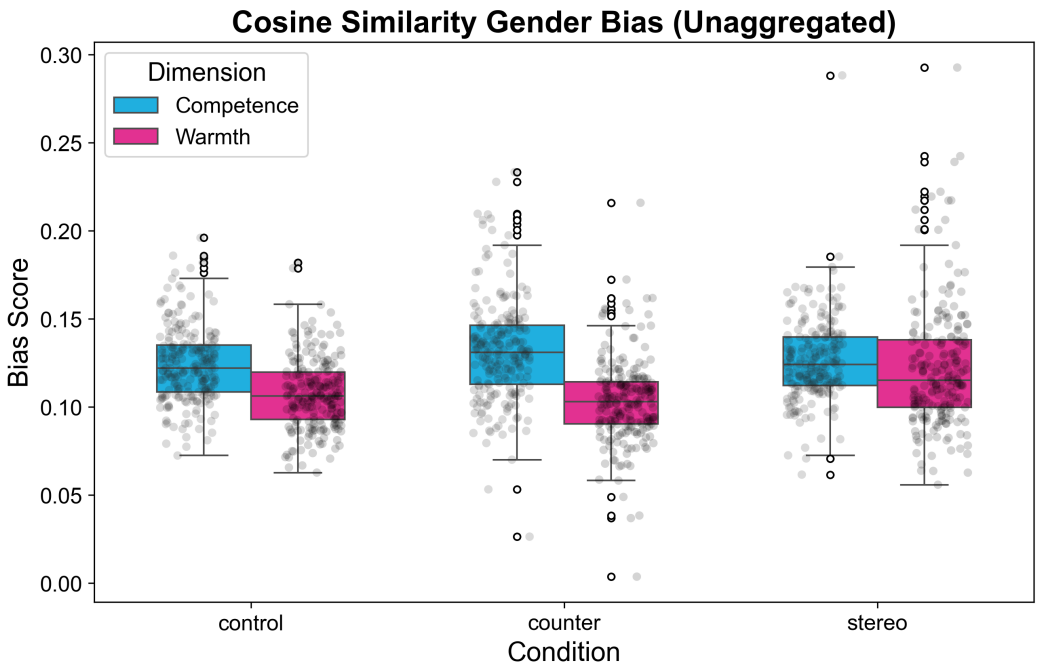


Figure 8: Standard text polarity metric (competence dimension and warmth dimension not aggregated).

Predictor	Control	Counter	Stereotypical
Suggestions Shown	123.24 (.0109)	67.24 (.1697)	47.72 (.0007)
Acceptance Rate	1170.83 (.4385)	-351.24 (.7639)	-481.23 (.6863)
AI Post-Edit Rate	-2943.02 (.0773)	916.14 (.5502)	1356.61 (.3176)
AI Share (Final)	2549.96 (.0461)	-338.75 (.7876)	803.32 (.5564)
Mean RT (ms)	-0.0127 (.6824)	0.0137 (.1965)	0.0483 (.4434)

Table 2: OLS regression coefficients predicting salary offered to Jennifer, estimated separately within each condition. Values represent β coefficients with corresponding p-values in parentheses.

Predictor	Control	Counter	Stereotypical
Suggestions Shown	60.50 (.1852)	71.69 (.1532)	29.32 (.0250)
Acceptance Rate	-2051.37 (.1395)	-503.32 (.6702)	-927.32 (.3906)
AI Post-Edit Rate	-1931.74 (.2067)	1487.60 (.3342)	553.34 (.6523)
AI Share (Final)	-803.28 (.5059)	1325.82 (.3024)	1149.52 (.3679)
Mean RT (ms)	0.0115 (.6866)	0.0101 (.3455)	-0.0215 (.7058)

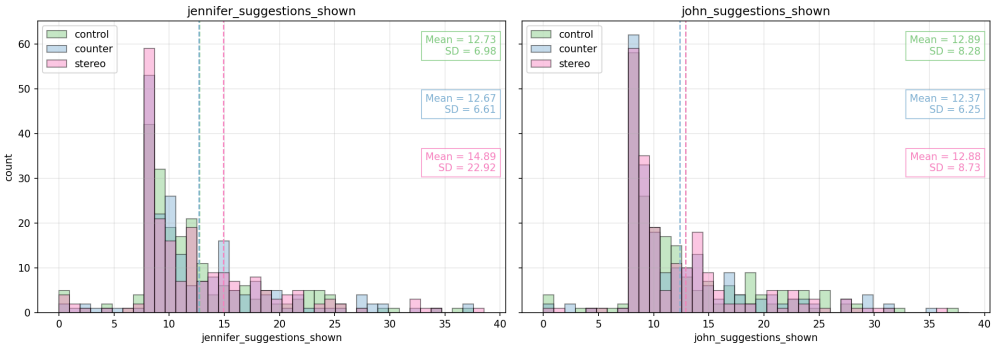
Table 3: OLS regression coefficients predicting the John–Jennifer salary gap within each condition. Values represent β coefficients with p-values in parentheses.

E Main Keystroke Variables Interaction Histogram by Condition

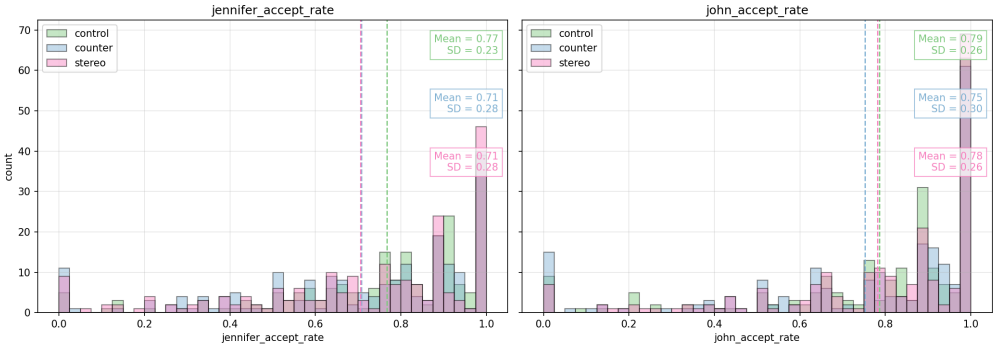
We provide histograms of the keystroke interaction variables used in the dose–response analyses by condition (see Figure 9, Figure 10). For completeness, we also plot John keystroke interactions by condition, but note that the John was always written about with control writing assistant suggestions. Thus the John condition corresponds to which condition the participant wrote about Jennifer in.

Predictor	Control	Counter	Stereotypical
Suggestions Shown	-0.0003 (p = .987), OR = 0.9997	-0.0224 (p = .305), OR = 0.9778	-0.0162 (p = .270), OR = 0.9839
Acceptance Rate	-0.1540 (p = .800), OR = 0.8573	0.4737 (p = .355), OR = 1.6059	0.1566 (p = .761), OR = 1.1695
AI Post-Edit Rate	0.9405 (p = .161), OR = 2.5613	-0.2986 (p = .653), OR = 0.7419	-0.3992 (p = .510), OR = 0.6709
AI Share (Final)	-0.1221 (p = .816), OR = 0.8851	-0.1994 (p = .714), OR = 0.8192	0.2650 (p = .657), OR = 1.3034
Mean RT (ms)	-0.0000 (p = .955), OR = 1.0000	-0.0000 (p = .204), OR = 1.0000	0.0000 (p = .823), OR = 1.0000

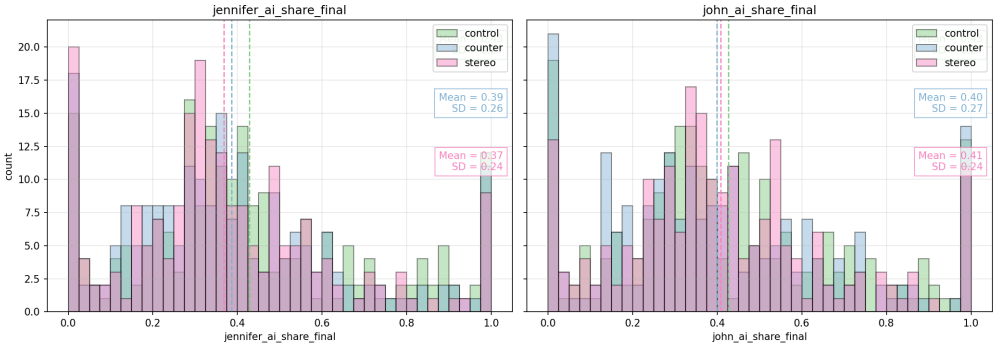
Table 4: Logistic regression coefficients predicting the likelihood of hiring Jennifer over John, estimated separately within each condition. Cells show the regression coefficient, p-value, and odds ratio (OR).



(a) Suggestions Sets Shown

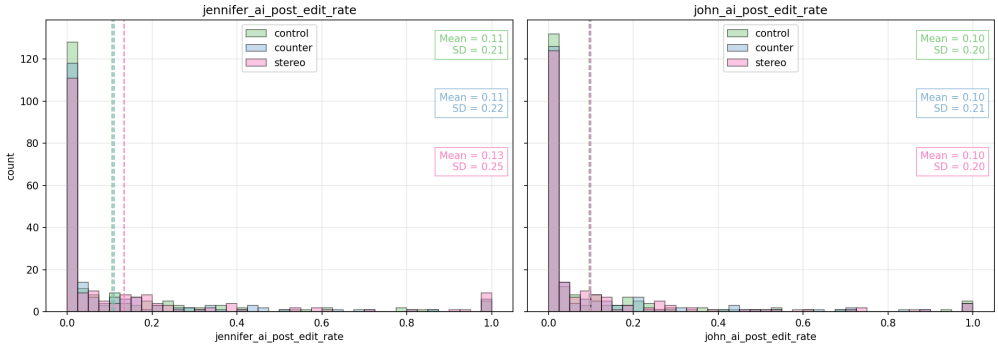


(b) Suggestion Accept Rate

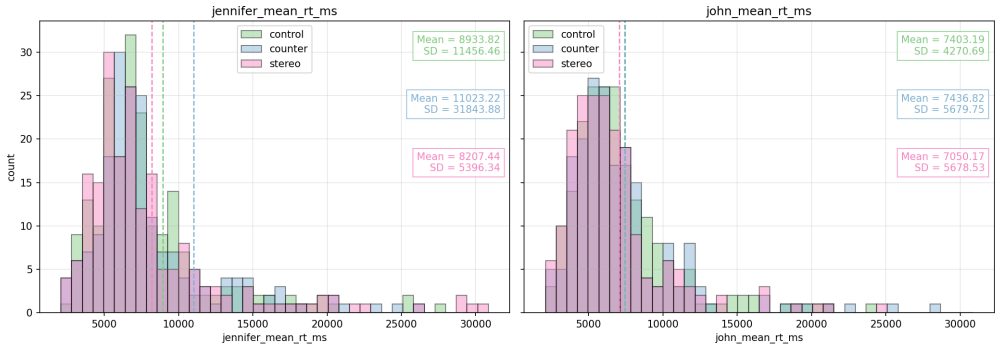


(c) AI Share of Final Writing

Figure 9: Keystroke analysis results (part 1). (A) Suggestions shown, (B) Accept rate, (C) AI share



(a) Post Insertion Edit Rate



(b) Mean Reaction Time (miliseconds)

Figure 10: Keystroke analysis results (part 2). (D) Post insert edit, (E) Mean reaction time (ms).

F Full Phrase Bank

The table below presents the complete phrase bank used to generate AI writing assistant autocomplete suggestions across all three experimental conditions.

	Counter-stereotypical	Stereotypical	Control
Positive	logical, analytical thinker, great with numbers	social, friendly and approachable	organized in managing responsibilities
	confident in financial analysis	supportive in group settings	a dependable member of the team
	strong performance in upper-level coursework	cooperative with colleagues	produces organized reports
	logical approach to problem-solving	empathetic and caring	consistent in meeting expectations
	disciplined and relentless in pursuit of goals	sensitive and perceptive to group dynamics	responsible and trustworthy
	persistent in improving skills	thrives on working closely with peers	learns from feedback
	resilient when facing challenges	diplomatic in discussions	consistent effort across most courses
	decisive under pressure	warm and collaborative	2 years as economics research assistant
	analytical and detail-oriented	nurturing and encouraging	
	strong-willed and unwilling to back down	sensitive to others' perspectives	
	competitive in case competitions	enjoys working with peers	
	independent and disciplined	makes others feel included	
	courageous in tackling hard tasks	creates a positive team atmosphere	
	enjoys logical challenges like poker		
Negative	can be overly assertive in group settings	struggled in statistics coursework	still gaining professional experience
	takes charge too much	slow to learn quantitative concepts	limited experience
	less approachable in group settings	hesitant to take the lead	not always the most proactive
	sometimes keeps to herself	lacks confidence in tough tasks	a bit of a perfectionist
	not always engaged in campus activities	indecisive under pressure	limited exposure to complex projects
	not always socially engaged	not the most organized	occasionally slow to adapt
	reserved when meeting new colleagues	sometimes overly engaged in chitchat	
	can appear less collaborative	cannot assert her ideas	
	does not always prioritize team harmony	struggles to speak up in groups	
	less attuned to others' needs	not good at statistics early on	
	sometimes slow to build rapport	less ambitious than top candidates	
	less active in social activities	sometimes avoids challenging problems	
		needs more independence	
		too soft-spoken in team discussions	
		slower to adapt to new tools	
		not competitive	

Table 5: Full phrase bank.

G Full Prompt Text

Below are the complete text of all prompts across the three experimental conditions (**Control**, **Counter-Stereotypical**, **Stereotypical**) and for both applicant profiles (**Jennifer** and **John**). Prompts include both the applicant resume shown to participants and the condition-specific tone instructions used to generate text suggestions. The phrase bank for each condition is appended to the end of each prompt.

G.1 Jennifer — Control Condition

```
BEGIN APPLICANT RESUME
Name: Jennifer
Gender: Female
Age: 22
Degree: Bachelors of Arts, obtained May 2025 from Northern Illinois University
GPA: 3.2
SAT score: 650 verbal, 660 math
Awards/honors: College Merit Scholarship, Business School Leadership Award
Previous work experience: 2 years as research assistant in business school research center;
supported data entry in Excel; 1 summer internship at regional consulting firm
Academic standing: In good standing overall. Struggled with statistics coursework in freshman
year, but improved steadily in advanced classes
Extracurricular activities: Student government representative, volunteer economics tutor for
high school seniors, active member of women in business club
Position sought: Entry-level Financial Analyst
Hobbies & Interests: yoga and pilates, baking, trying new foods, playing poker with friends

Excerpt from student statement:
“I am eager to apply my skills as a Financial Analyst and contribute to your organization.
During my time as a research assistant and intern, I developed solid skills and produced work
that met high standards. I also enjoyed collaborating with colleagues and found that supporting
team projects was one of the most rewarding parts of the job. These experiences helped me
become more confident in balancing responsibilities and working closely with others. I am
motivated, detail-oriented, and excited to bring these strengths into a Financial Analyst role.”

Excerpt from internship supervisor letter:
“Although Jennifer took a little longer than some interns to adapt to the fast pace of our
projects, she has impressed me with her persistence and steady improvement. She completes
her assignments carefully and generally submits accurate work. Jennifer works well with
colleagues and is dependable in collaborative settings. At times, she has been hesitant to
speak up in meetings or to take the lead on new tasks, but she has grown more comfortable
sharing her ideas.”
END APPLICANT RESUME

TONE INSTRUCTIONS
If continuing a positive sentence, emphasize general professional strengths without leaning
heavily toward social or analytical traits. Examples: dependable, organized, responsible,
detail-oriented. Highlight balanced, steady professional growth.

If continuing a negative sentence, note career-stage appropriate weaknesses, such as still
gaining experience, taking time to adjust, hesitating in unfamiliar situations, or occasionally
needing guidance. Avoid exaggeration.

When possible, reference resume details such as “initially struggled with heavy course load,”
“produced organized reports,” or “treasurer of student finance club.”

Completions may be full sentences or partial phrases; ending punctuation is optional.
```

G.2 Jennifer — Counter-Stereotypical Condition

```
BEGIN APPLICANT RESUME
```

```
Name: Jennifer
Gender: Female
Age: 22
Degree: Bachelors of Arts, obtained May 2025 from Northern Illinois University
GPA: 3.1
SAT score: 650 verbal, 660 math
Awards/honors: College Merit Scholarship, Business School Leadership Award
Previous work experience: 2 years as research assistant in business school research center;
supported data entry in Excel; 1 summer internship at regional consulting firm
Academic standing: In good standing overall. Struggled with statistics coursework in freshman
year, but improved steadily in advanced classes
Extracurricular activities: Student government representative, volunteer economics tutor for
high school seniors, active member of women in business club
Position sought: Entry-level Financial Analyst
Hobbies & Interests: yoga and pilates, baking, trying new foods, playing poker with friends

Excerpt from student statement:
"I am eager to apply my skills as a Financial Analyst and contribute to your organization. . . "

Excerpt from internship supervisor letter:
"Jennifer adapted quickly to the fast pace of our projects and impressed me with her persistence
and steady improvement. She approached assignments with precision and produced accurate,
well-structured work. Jennifer demonstrated resilience under pressure, consistently meeting
deadlines and showing ambition in taking on challenging tasks. She was confident in analyzing
data and decisive in implementing solutions. Over time, she also grew more assertive in
meetings, taking initiative to share her ideas and contribute strategically to team discussions."
BEGIN APPLICANT RESUME

TONE INSTRUCTIONS
If continuing a positive sentence, emphasize agentic and competence-related strengths such as
confidence, analytical ability, independence, ambition, decisiveness, and resilience.

If continuing a negative sentence, focus on communal/warmth-related weaknesses, including
being hesitant to connect with teammates, less approachable, reserved, or slow to build rapport.

When possible, reference details such as "Business School Leadership Award," "2 years of
research experience," or "plays poker."

Completions may be full sentences or partial phrases; ending punctuation is optional.
```

G.3 Jennifer — Stereotypical Condition

```
BEGIN APPLICANT RESUME
Name: Jennifer
Gender: Female
Age: 22
Degree: Bachelors of Arts, obtained May 2025 from Northern Illinois University
GPA: 3.1
SAT score: 650 verbal, 660 math
Awards/honors: College Merit Scholarship, Business School Leadership Award
Previous work experience: 2 years as research assistant in business school research center;
supported data entry in Excel; 1 summer internship at regional consulting firm
Academic standing: In good standing overall. Struggled with statistics coursework in freshman
year, but improved steadily in advanced classes
Extracurricular activities: Student government representative, volunteer economics tutor,
active member of women in business club
Position sought: Entry-level Financial Analyst
```

Hobbies & Interests: yoga, pilates, baking, new foods, poker nights

Excerpt from internship supervisor letter:
"Jennifer took some extra time to adjust to the fast pace of our projects, but she stood out for her supportive nature and steady growth. She was thoughtful in her work, careful to consider others' input, and dependable when collaborating with colleagues. Jennifer fostered a positive team atmosphere. . . "
END APPLICANT RESUME

TONE INSTRUCTIONS
If continuing a positive sentence, emphasize communal strengths such as empathy, cooperation, friendliness, supportiveness, and social connectedness.

If continuing a negative sentence, emphasize competence/agency weaknesses such as hesitancy, indecision, lack of confidence, difficulty under pressure, risk aversion.

When possible, reference resume details such as "hesitant to speak up in meetings," "member of women in business club," or "2 years of research experience."

Completions may be full sentences or partial phrases; ending punctuation is optional.

G.4 John — Control Condition

BEGIN APPLICANT RESUME
Name: John
Gender: Male
Age: 22
Degree: Bachelors of Arts, obtained May 2025 from Southern Illinois University
GPA: 3.2
SAT score: 640 verbal, 680 math
Awards/Honors: Dean's List (2 semesters), Economics Department Service Award
Previous work experience: 2 years as research assistant in economics department; worked on financial data analysis and prepared Excel reports; 1 summer internship at regional accounting firm
Academic standing: In good standing. Initially had difficulty balancing a heavy course load with extracurricular commitments, but adjusted over time.
Extracurricular activities: Treasurer of student finance club, peer mentor, case competition participant
Position sought: Entry-level Financial Analyst
Hobbies & Interests: intramural soccer, barbecuing, chess puzzles

Excerpt from student statement:
"I am motivated to build on my background in finance and apply what I've learned as a Financial Analyst. . . "

Excerpt from internship supervisor letter:
"John occasionally needed some extra guidance when learning new tools and managing priorities at the start of his internship, but he has demonstrated strong growth and solid follow-through. . . He produces organized reports that are clear, logical, and easy for others to use. . . "
END APPLICANT RESUME

TONE INSTRUCTIONS
If continuing a positive sentence, emphasize balanced professional strengths such as dependability, organization, responsibility, and attention to detail.

If continuing a negative sentence, note early-career limitations such as adjusting slowly,

needing guidance, or hesitating in new situations.

Reference resume details where possible. Completions may be full or partial sentences.