Pilot 1: Validate Dictionary

The goal of the first pilot study was to validate our dictionary of gendered words (i.e., if people's perception of gendered language aligned with our dictionary). Additionally, we sought to explore whether differences in language use in recommendation letters could affect how people evaluate applicants.

Participants. We have 51 participants in total for this study including 11 students in our department and 40 participants recruited through Prolific. All participants are over 18 years old and fluent in English.

Stimuli. We selected letters for this study that varied in language use (more female-associated (Female) or more male-associated (Male)) and quality (Strong or Weak) resulting in four conditions. We use a set of criteria to decide the candidate letters based on 1) the total number of words of a letter (Length); 2) the unique number of female-associated words (FW); 3) the unique number of male-associated words (MW); and 4) the ratio between female-associated words and male-associated words (FW/MW).
The selection criteria for each condition are summarized below:

- **Strong Female:** FW >= 5, FW/ML >= 1.2, Length > 300.
- **Weak Female:** FW >= 2, FW/ML >= 1.2, Length < 200.
- **Strong Male:** MW >= 5, FW/ML <= 0.8, Length > 300.
- **Weak Male:** MW >= 2, FW/ML <= 0.8, Length < 200.

In addition to the criteria above, we also restricted the applicant pool to those with a master's degree. Two of the authors read the candidate letters and selected one letter for each condition. The selected letters were then anonymized where sensitive information such as applicant and recommender names, university names, and email addresses were redacted. Additionally, gendered pronouns (he/him, she/her, etc.) were replaced with gender-neutral pronounces (they/them).

Procedure. Each participant was randomly assigned to one of the four conditions. Participants were asked to read a recommendation letter to answer questions about 1) the perceived gender of the applicant; 2) the perceived gender of the recommender (letter writer); and 3) how competitive the applicant was, based only on the letter, on a scale of 1 (Extremely uncompetitive) to 7 (Extremely competitive). For each question, we also asked participants' confidence about their answers in the form of a slider from 0-100 and asked them to list the words/phrases from the letters that informed each response.

Results. Overall, inferring the gender of the applicant based on the language of the letter is not easy. The accuracy of the perceived applicant gender question for the four letters was 0.22, 0.64, 0.5, and 0.5 respectively. We notice that some participants simply guessed or made judgments based on the statistics in the computer science field, i.e., in general, there are more men in the field. These responses (N=17) were excluded from the analysis. In the subsequent

studies, we clarified that the gender of the applicants and letter writers are evenly distributed so participants would be less likely to make judgments based on the statistics in the field.

We compared the words that participants listed as indicators of the applicant's gender with the words in our dictionary and found that most of the words that participants listed as indicators for female aligned with our dictionary including cooperative, collaborative, hardworking, polite, and dedicated. Participants also mentioned words that were not in our dictionary. For example, initiative and leadership were associated with males. We added these words to our dictionary.
Some male-associated words in our dictionary were perceived differently by the participants. For example, some participants associated excellent, intellectual, and skill with female while some associated with male.

The average competitiveness ratings for the four letters were 4.77, 5.64, 5.80, and 5.00 respectively. We found no significant difference between the competitiveness ratings for Strong letters (mean = 5.36, sd = 1.11) and Weak letters (mean = 5.35, sd = 1.20). There was also no significant difference between the competitiveness ratings for letters with more female language and letters with more male language. However, we did observe that the Female Strong letter was rated slightly lower on average (mean = 5.00, sd = 1.21) compared with the Male Strong letter (mean 5.69, sd = 0.95). We noticed that the Female Weak letter got the second-highest competitiveness rating and a closer look at participants' responses reveals that having published papers is a strong factor that dominates competitiveness perceptions. In the subsequent study, having publications is considered as a factor when selecting letters. We also found that some female-associated words such as team, and loyal were listed as indicators of uncompetitive by participants.