Gender Bias in Artificial Intelligence-Written Letters of Reference for Otolaryngology Residency Candidates

Grace Young[1], Marianne Abouyared MD[3], Alexandra Kejner MD[4], Rusha Patel MD[5], Heather A. Edwards MD[6], Linda X. Yin MD[7], Janice L. Farlow MD, PhD[1, 2]

[1]Indiana University School of Medicine, Indianapolis, Indiana, USA. [2]Department of Otolaryngology–Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, Indiana, USA. [3]Department of Otolaryngology–Head and Neck Surgery, University of California Davis, Sacramento, California, USA. [4]Department of Otolaryngology–Head and Neck Surgery, Medical University of South Carolina, Charleston, South Carolina, USA. [5]Department of Otolaryngology–Head and Neck Surgery, University of Oklahoma College of Medicine, Oklahoma City, Oklahoma, USA. [6]Department of Otolaryngology–Head and Neck Surgery, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, USA. [7]Department of Otolaryngology—Head and Neck Surgery, Mayo Clinic Rochester, Rochester, Minnesota, USA.

Abstract

Introduction:
- Written letters of reference (LORs) are an important component of the residency application process.
- Human-written LORs have been shown to contain gender-bias.

Knowledge Gap:
- In a previous study, we identified gender bias in AI-written LORs when using prompts with randomly-generated resume variables.
- We sought to investigate whether this bias persisted using **real applicant experiences**, and how this compared to the LORs written by otolaryngology faculty.
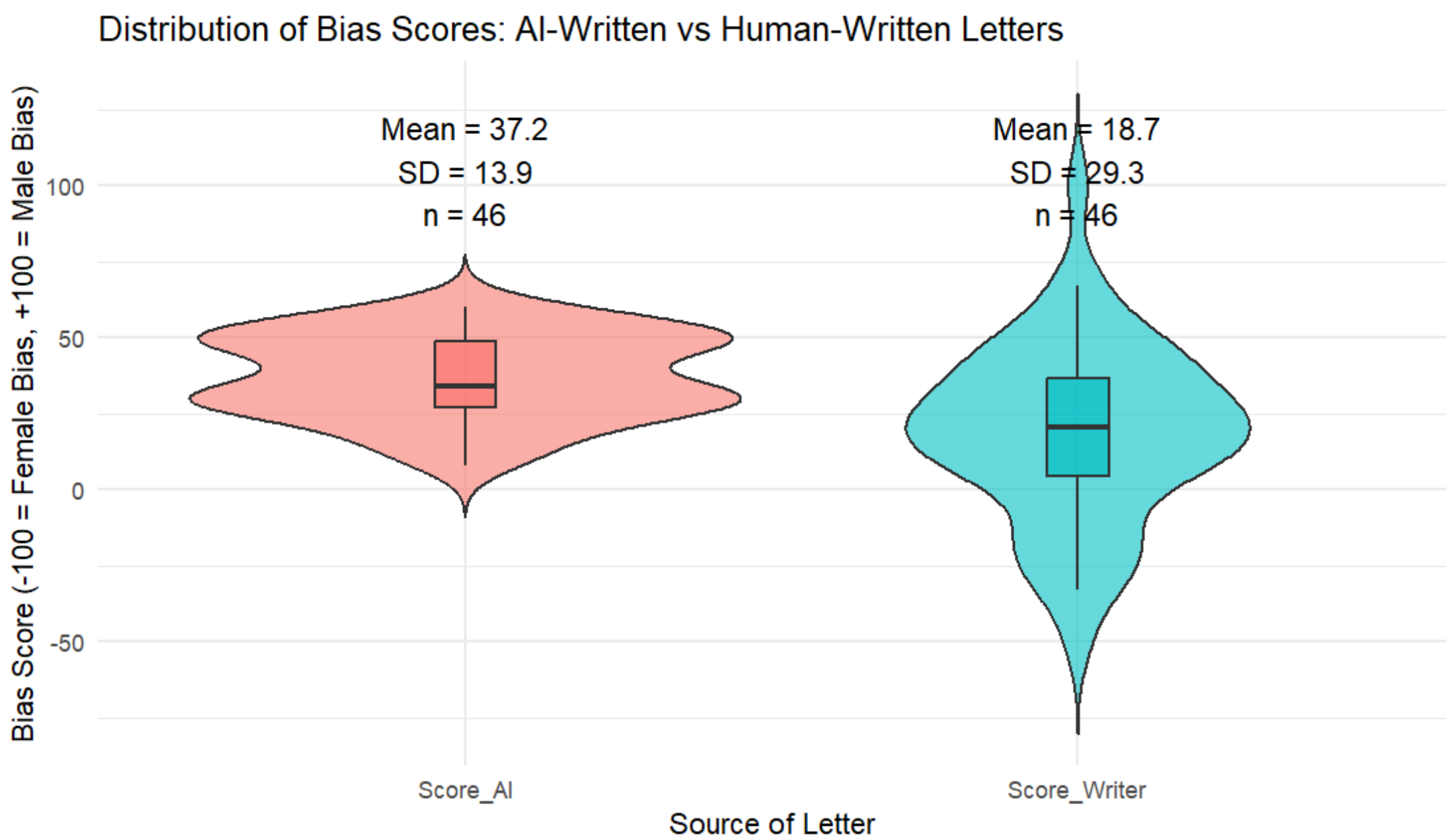
Methodology:
- We obtained 46 LORs for otolaryngology residency applicants written by faculty from 5 different institutions.
- Prompts describing candidate experiences were provided to ChatGPT4.0 in individual sessions.
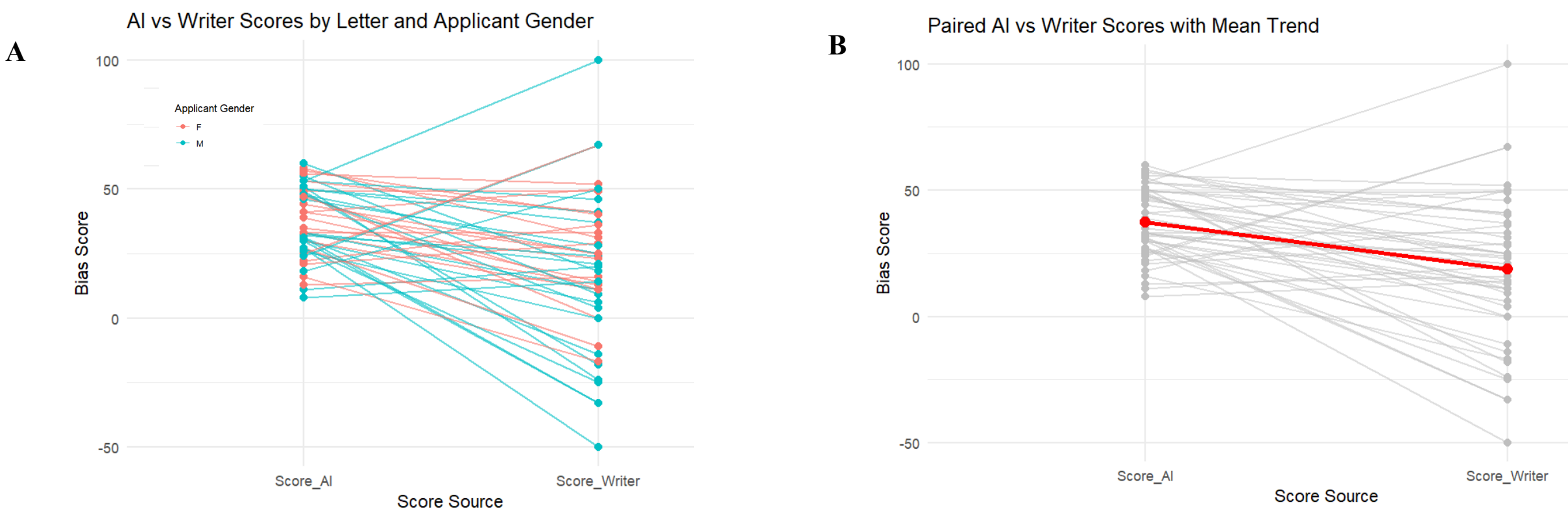- The writer-generated and AI-generated letters were compared using a gender-bias calculator.

Key Findings:
- Both the writer-generated and AI-generated letters exhibited male bias on average (18.7% and 37.2% respectively). We used a paired t-test to determine that the AI-generated letters exhibited significantly higher male bias (t-statistic: -4.27, p-value: 0.0001).
- Independent t-tests did not reveal a significant difference for male versus female applicants for either writer-generated (t-statistic: 1.54, p-value 0.131) or AI-generated letters (t-statistic: 0.14, p-value: 0.892).
- Levene's test indicated AI had significantly lower variability than for writers (Levene's statistic: 11.38, p-value: 0.0011).

Gender Bias Calculator

- For chemistry and biochemistry job applicants
- Female-associated words
  - Grindstone words (hardworking, reliable)
  - Teaching words (training, teachable)
- Male-associated words
  - Ability words (intellect, capacity)
  - Standout adjectives (best, highest, excellent)
- Scale from -100 (female) to +100 (male)

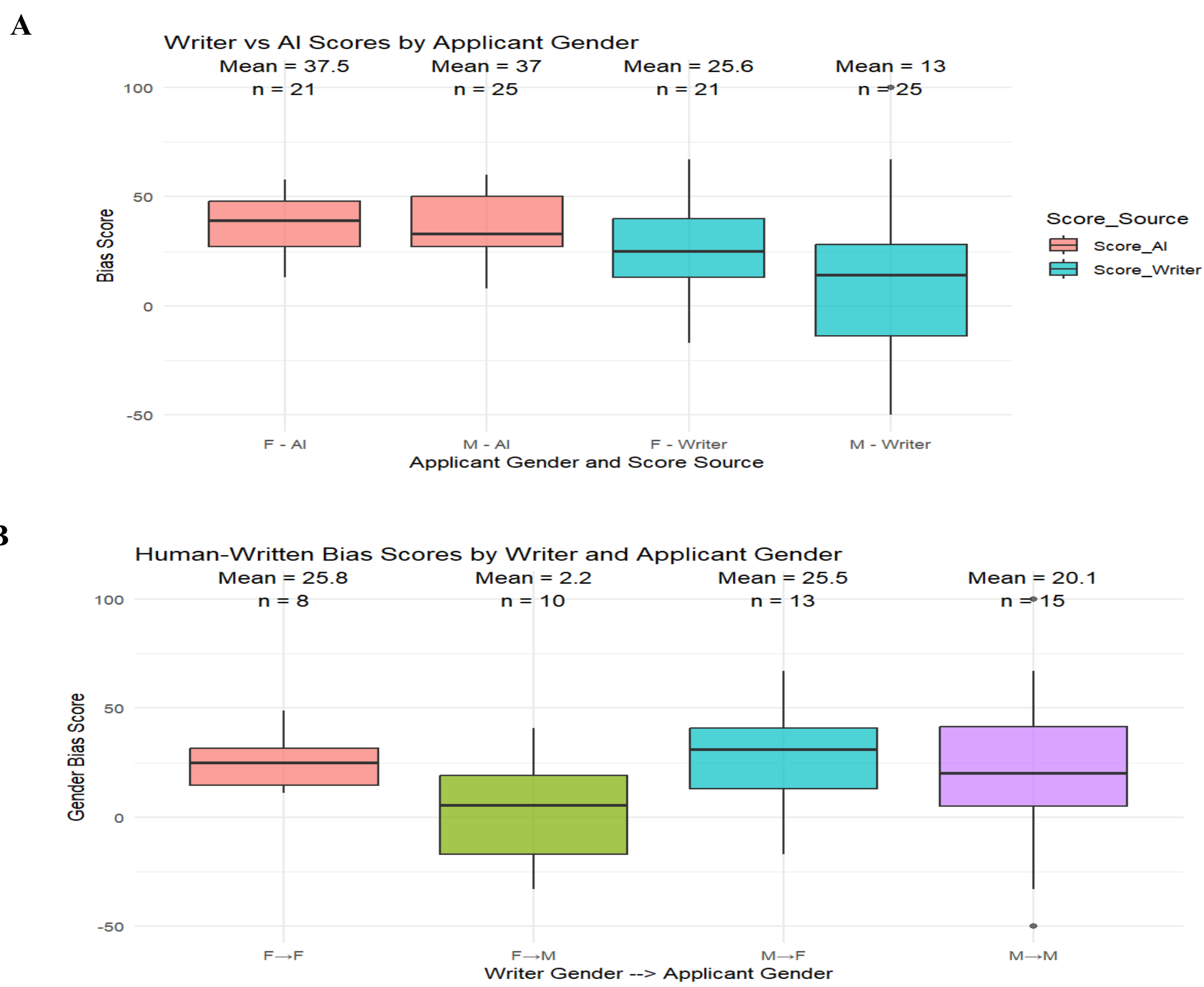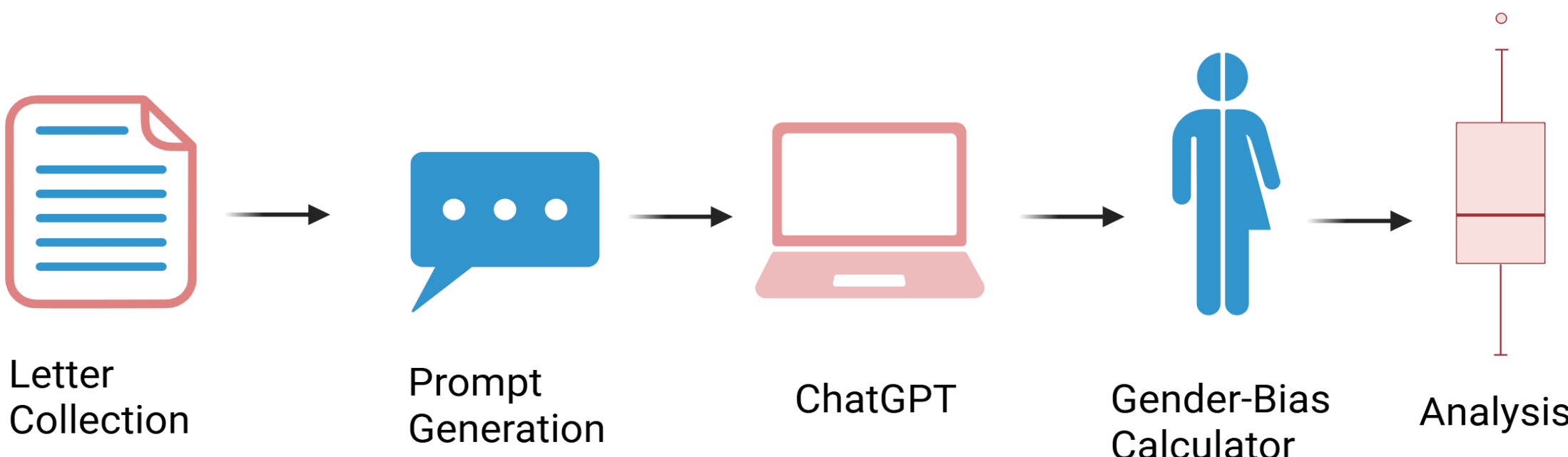| Female-associated words	Male-associated words
	highest
	strongest
	able
student ×2	research ×4
work ×2	remarkable
service	intellect,
reliable	skills, ×3
colleague	resulted
organizational	publications
diligence	excellent ×2
presentations,	insight
education	insightful
presentations	skills
teachable. In	capacity
	intellectual,
	skilful
	bright

Distribution of Bias Scores

Fig. 1: Violin plot comparing AI-written to human-written letters. The variance is significantly lower for the AI-generated letters, suggesting improved consistency between letters, but not elimination of gender bias.

Fig. 2: A. Spaghetti plot maps the scores of real letters to AI-generated counterparts. We used a paired t-test to determine that the AI-generated letters exhibited significantly higher male bias (t-statistic: -4.27, p-value: 0.0001). **B.** Mean trendline between paired scores.

Methods

Letter Collection → Prompt Generation → ChatGPT → Gender-Bias Calculator → Analysis

Fig. 3: Methods. Illustration of the workflow used to generate bias scores.

Sample Prompt

"*Write a letter of reference for Michael, who is a 4th year medical student applying to Otolaryngology residency. He attended multiple conferences, was involved in research in the department, obtained a BS in Biology at the University of XXX and graduated with Honors, spent 2 years as a research assistant...*"

Results

Fig. 4: Results. A. Data graph comparing AI-letter scores (red) with human-written letter scores (blue), broken down by applicant gender. All categories show male bias on average, with no significant differences by gender. **B.** Graph of bias scores for human-written letters, broken down by gender of the writers and of the applicants. ANOVA shows no statistically significant difference among the four groups.

Conclusions

- Most letters exhibited male bias, including all ChatGPT letters.
- AI letters were significantly more male-biased than human-written letters. Understanding why the LLM demonstrated male bias is beyond the scope of this investigation. One may postulate that ability words tended to co-occur more frequently with the research and scholarly activities performed by the candidates.
- No significant difference was found between genders for either human-written or AI-written letters.
- The variance of the AI letters was significantly lower than for human-written letters. Reassuringly, the use of AI may help reduce discrepancies between genders, as it does not demonstrate a difference when given a male or female name.
- It is not yet clear how gender discrepancies in letters of recommendation may impact candidate selection.

References

Farlow, J.L., Abouyared, M., Rettig, E.M., Kejner, A., Patel, R. and Edwards, H.A. (2024), Gender Bias in Artificial Intelligence-Written Letters of Reference. Otolaryngol Head Neck Surg, 171: 1027-1032. https://doi.org/10.1002/ohn.806.

Schmader T, Whitehead J, Wysocki VH. A Linguistic Comparison of Letters of Recommendation for Male and Female Chemistry and Biochemistry Job Applicants. Sex Roles. 2007;57(7-8):509-514. doi: 10.1007/s11199-007-9291-4. PMID: 18953419; PMCID: PMC2572075.