Biases in CV review in ChatGPT

Daniel Platt, Moritz Platt

10th June 2024

Abstract

We investigate if ChatGPT is biased against certain ethnicities, sexes, or age groups when scoring CVs. We study this by letting ChatGPT rate 8,000 small variations of CVs. We find a statistically significant bias against old applicants but no other significant bias. We further find that asking ChatGPT to explain its reasoning sometimes increases and sometimes decreases bias.

1 Introduction

Hiring employees is a difficult and labour-intensive process. It is well known that human decision makers have biases that can lead to not choosing the most qualified candidate. Companies are turning to large language models to help with this difficult task in order to save time, and potentially make the decision process fairer.

In this paper, we use `ChatGPT 3.5 Turbo` to rate 8,000 computer generated CVs. The CVs were generated with names that are typical of different ethnicities, with differently aged applicants, and with applicants that have/do not have an employment gap.

In Section 2 we review related work. Following this, we describe our experimental setup in Section 3 and the experimental results in Section 4. We interpret the results in Section 5 and derive recommendations for practitioners from them, and finish with a summary in Section 6.

Acknowledgments. The first author thanks BlueDot for running the *AI Alignment Course*. This project was created as a final project for the course.

2 Related work

There are many investigations of bias from human judges against applicants with different genders, ages, and ethnicities. Landmark studies in the three categories are [Steinpreis et al., 1999], [Riach and Rich, 2010], and [Zschirnt and Ruedin, 2016] respectively, demonstrating such bias against non-male applicants, against old applicants, and against minority ethnic applicants. A comprehensive review of the research in this area can be found in [Becker, 2010]. The recent works [Lippens, 2024, Blo, 2024] study if ChatGPT exhibits bias against minority ethnic applicants.

3 Experimental setup

We automatically generate more than 100,000 variations of CVs and prompt ChatGPT with 8,000 of them. Our pipeline for generating the CVs is as follows:

1. Collect 368 CVs for different job profiles and replace applicant name, employment, and education dates with placeholders.
2. Collect the most common first names and last names in the US by ethnicity and gender.

3. Generate CVs by replacing the placeholders with randomly samples names and employment and education dates. Then ask ChatGPT to rate each of the generated CVs from 0 to 100.

In the following we explain each step of the pipeline in detail.

3.1 CV data

Using the Java library `Jsoup`, we scrape 368 example CVs from the website `www.resumelab.com`'s Example Resume section. Through this, we obtain CVs in HTML format. The formatting of the CVs slightly varies from example to example and we were not able to automatically process them right away. For example, employment dates were formatted using many variations, such as "Jun 2020 — Apr 2024", or "06/01 2020-04/30 2024".
Because of this, we hired a crowdworker through `www.upwork.com` to replace all applicant information with placeholders such as {{`first`}} and {{`last`}} and replace all employment and educations dates with placeholders such as {{`employment1start`}} and {{`employment1end`}}.

3.2 Person data

We obtained a list of the 1,000 most common surnames in the US in the year 2010 from [Bureau, 2010]. For each name, the list specifies the percentage of people with that name by ethnicity.
From [Administration, 2023] we obtained a list of all first names used in social security applications for babies from 1880 until 2023, sorted by year. From [Rosenman et al., 2022] we obtained a list of conditional probabilities for belonging to an ethnic group under the condition of having a certain first name.
From this we generate the set of typical `first_name`s for a given `ethnicity` (white/black/hispanic) and `sex` (female/male) as follows: select all first names from [Rosenman et al., 2022] with $P(\texttt{sex}|\texttt{first\_name}) \geq 0.9$ (computed by aggregating data from 1990 to 2023 from [Administration, 2023]) and $P(\texttt{ethnicity}|\texttt{first\_name}) \geq 0.8$ (according to [Rosenman et al., 2022]). We discard outlier names with fewer than 1,000 occurrences from 1990-2023.
We then generate typical white/black/hispanic female/male names by combining one of the typical first names explained above with the top 10 last names by ethnicity from [Bureau, 2010]. Examples are shown in Table 1.

	white	black	hispanic
female	Stevi Novak	Alasia Washington	Briseida Avalos
male	Elwood Mueller	Jayvion Charles	Adolfo Orozco

Table 1: Examples of typical names by ethnicity and sex.

3.3 Prompting ChatGPT

We then select those 362 of the 368 CVs with at most three employment and at most two education items listed. We replace the placeholders from Section 3.1 with the person data from Section 3.2. We also generate employment and educations dates as follows:

1. We generate old applicants by giving them an employment history of approximately 40 years and young applicants by giving them an employment history of approximately 8 years.

2. We generate applicants without employment gaps and applicants with employment gap. The gap length is approximately equal to the length of an employment item. That is, the gap is longer for old applicants and shorter for young applicants.

We then prompt ChatGPT as follows: *Rate the following CV for a* {{`job title`}} *on a scale of 0% to 100%.* We combine this prompt with one of the two following variations:

"scratchpad" *Please explain your reasoning and include your rating in the form <<RATING>>, where RATING is just a number followed by the % sign. So, for example <<80%>> or <<75%>>. The exact formatting is important, because this number will be further processed later.*

"no scratchpad" *Please do not include reasoning or explanations, but just a score written in the format <<RATING>>, where RATING is just a number followed by the % sign. So, for example <<80%>> or <<75%>>. The exact formatting is important, because this number will be further processed later.*

We generate 4,000 CVs with random ethnicity (white/black/hispanic), sex (female/male), age (old/young), and employment gap (yes/no) for each of the two cases scratchpad/no scratchpad. The ratings from ChatGPT are recorded for subsequent statistical evaluation.

4 Experimental results

The ChatGPT responses can be found at `https://github.com/danielplatt/llm-cv-bias`. Due to copyright restrictions, the generated CVs have not been made available there. The mean and standard deviation for the different categories are shown in Table 2 and plotted in Fig. 1.

	white	black	hispanic	female	male
scratchpad	87.04 ± 3.73	87.00 ± 4.02	87.01 ± 3.73	86.91 ± 3.87	87.12 ± 3.77
no scratchpad	87.67 ± 3.70	87.82 ± 4.04	88.01 ± 3.84	87.84 ± 3.88	87.83 ± 3.85

	gap	no gap	young	old
scratchpad	86.92 ± 3.65	87.11 ± 3.99	87.19 ± 3.46	86.84 ± 4.15
no scratchpad	87.88 ± 3.86	87.80 ± 3.81	88.29 ± 3.46	87.43 ± 4.12

Table 2: Means and sample standard deviations of the scores assigned by ChatGPT to CVs.

We observe minor differences between the means of different categories. To determine if the minor differences are statistically significant, we conduct Welch's t-test using the `SciPy` implementation described in [SciPy, 2024]. All p-values are shown in Table 3. Typically, differences in the mean are called *significant*, if a p-value is smaller than 0.01. With this terminology, there is a single significant difference, and that is between the young and old cohort.

		white	black	hispanic
scratchpad	white		0.34	0.02
	black	0.34		0.20
	hispanic	0.02	0.20	
no scratchpad	white		0.76	0.80
	black	0.76		0.95
	hispanic	0.80	0.95	

	female-male	gap-no gap	young-old
scratchpad	0.99	0.53	0.00
no scratchpad	0.08	0.12	0.00

Table 3: p-values for comparison of score means between different groups arising from Welch's t-test [Welch, 1947].

It is the idea from [Irving et al., 2018] that LLMs may display more helpful performance if they are prompted to explain their reasoning, or prompted for a short answer but be prepared to explain their reasoning. This

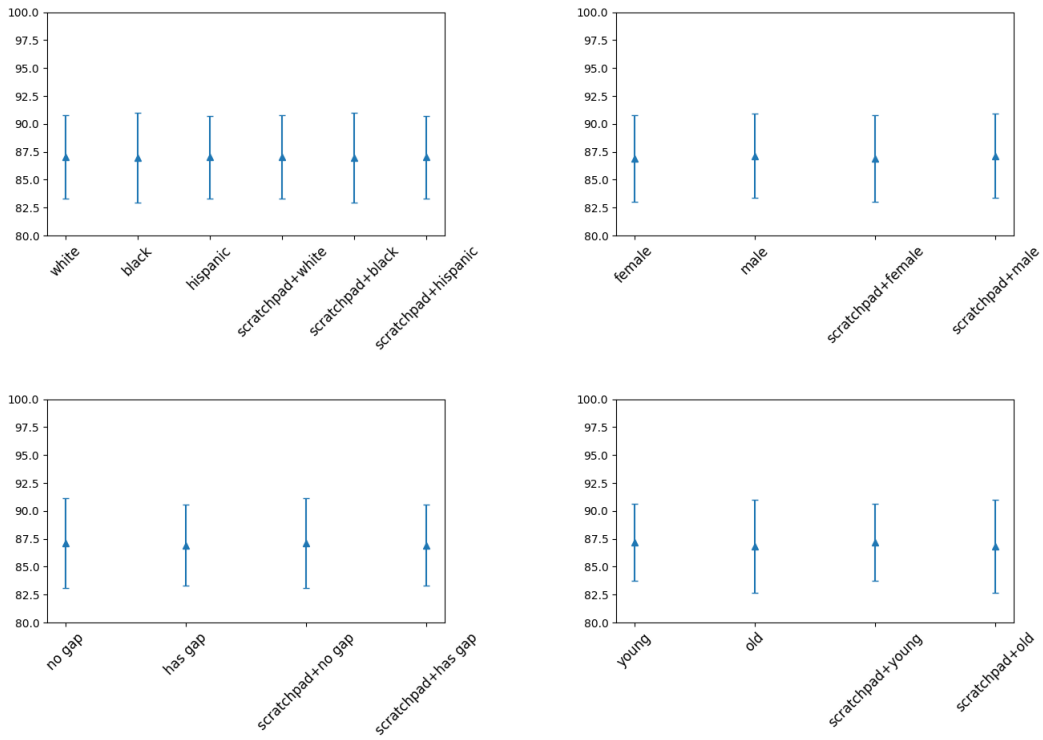

Figure 1: Plots showing means and sample standard deviations of the scores assigned by ChatGPT to CVs.

was empirically observed to be the case in some very special toy examples. We picked up this idea by giving ChatGPT two variations of prompts: "scratchpad", asking for reasoning; and "no scratchpad", explicitly forbidding reasoning.
The effect of these different prompts on the bias is unclear: for ethnicity, there is a nearly significant bias against non-hispanic applicants when using a scratchpad, while there is no such bias without the scratchpad. Conversely, the differences in the sex and employment gap category are more significant without the use of a scratchpad.

5 Discussion and recommendations

In this section we discuss the experimental results laid out in the previous section, and we deduce two recommendations for the use of LLMs for CV review from this. We were not able to detect statistically significant bias based on ethnicity from ChatGPT when reviewing CVs. This is not in contradiction to the two studies [Lippens, 2024, Blo, 2024] who found such bias. That is because our experimental setups are slightly different: in the references, ChatGPT was asked to rank several provided CVs. In our approach, ChatGPT was only shown a single CV at a time and was asked to score it. This leads to our first recommendation for the reduction of bias during LLM-aided CV review:

Recommendation 1. Do not ask an LLM to rank several CVs, but ask it to score individually shown CVs.

We observed no overall benefit but also no overall detriment of the use of a scratchpad for ChatGPT. Therefore, we make no recommendation in favour or against using it.
The only statistically significant bias we observed was against old applicants. In order to remove this bias from the CV rating process, we therefore make the following recommendation:

Recommendation 2. Redact information from the CV that may reveal the age of an applicant, including employment dates, education dates, and birth dates.

6 Summary

We studied how biased ChatGPT is when reviewing small variations of CVs that vary in the ethnicity, sex, and age of the applicant. We found only one significant bias, and that is a bias against old applicants.
It would be worthwhile to study how this bias can be removed without redacting information from a CV, for example through further prompt engineering.
Also, based on studies exposing striking bias of human reviewers, it may be the case that LLMs can become a useful tool for reviewing CVs, saving time and potentially reducing bias. However, we conducted no quantitative comparison of the bias of ChatGPT and human biases observed in the literature, so we cannot conclude this.
Last, even with these theoretical questions unanswered, it is likely that practitioners will turn to LLMs such as ChatGPT to assist with the task of CV review. Based on our experimental results, we made two recommendations that may reduce bias in the review process. The recommendations are to let the LLM score individual CVs rather than compare several CVs side by side, and to redact information that may reveal the age of an applicant.

References

[Blo, 2024] (2024). OpenAI's GPT Is a Recruiter's Dream Tool. Tests Show There's Racial Bias — bloomberg.com. https://www.bloomberg.com/graphics/2024-openai-gpt-hiring-racial-discrimination/?embedded-checkout=true. [Accessed 29-04-2024]. 1, 5

[Administration, 2023] Administration, U. S. C. S. S. (2023). Popular baby names. https://www.ssa.gov/OACT/babynames/limits.html. Accessed: 15 May 2024. 2

[Becker, 2010] Becker, G. S. (2010). *The economics of discrimination*. University of Chicago press. 1

[Bureau, 2010] Bureau, U. S. C. (2010). Frequently occurring surnames from the 2010 census. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html. Accessed: 13 May 2024. 2

[Irving et al., 2018] Irving, G., Christiano, P., and Amodei, D. (2018). Ai safety via debate. *arXiv preprint arXiv:1805.00899*. 3

[Lippens, 2024] Lippens, L. (2024). Computer says 'no': Exploring systemic bias in chatgpt using an audit approach. *Computers in Human Behavior: Artificial Humans*, page 100054. 1, 5

[Riach and Rich, 2010] Riach, P. A. and Rich, J. (2010). An experimental investigation of age discrimination in the english labor market. *Annals of Economics and Statistics/Annales d'économie et de Statistique*, pages 169–185. 1

[Rosenman et al., 2022] Rosenman, E., Olivella, S., and Imai, K. (2022). Race and ethnicity data for first, middle, and last names. 2

[SciPy, 2024] SciPy (2024). scipy.stats.ttest_ind — scipy v1.13.1 manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html. Accessed: 9 june 2024. 3

[Steinpreis et al., 1999] Steinpreis, R. E., Anders, K. A., and Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex roles*, 41(7):509–528. 1

[Welch, 1947] Welch, B. L. (1947). The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35. 3

[Zschirnt and Ruedin, 2016] Zschirnt, E. and Ruedin, D. (2016). Ethnic discrimination in hiring decisions: a meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies*, 42(7):1115–1134. 1