Women may pay a 'mom penalty' when AI is used in hiring, research suggests

December 13 2023

Credit: Pixabay/CC0 Public Domain

Maternity-related employment gaps may cause job candidates to be unfairly screened out of positions for which they are otherwise qualified,

according to new research from NYU Tandon School of Engineering.

A research team led by Siddharth Garg, Institute Associate Professor of Electrical and Computer Engineering, examined bias in Large Language Models (LLMs)—advanced AI systems trained to understand and generate human language—when used in hiring processes.

The team will present its findings in a paper presented at NeurIPS 2023 R0-FoMo Workshop on December 15. Akshaj Kumar Veldanda, Ph.D. candidate in Department of Electrical and Computer Engineering, is the paper's lead researcher.

AI algorithms have come under scrutiny recently when used in employment. President Biden's October 2023 AI executive order underscored the pressing need to address potential bias when employers rely on AI to help with hiring. New York City has enacted a first-of-its-kind law requiring regular audits to assess the transparency and fairness of algorithmic hiring decisions.

"Our research is helping develop a robust auditing methodology that can uncover hiring biases in LLMs, aiding researchers and practitioners in efforts to intervene before discrimination occurs," said Garg. "Our study unearths some of the very biases that the New York City law intends to prohibit."

In the study, researchers assessed the ability of three popular LLMs, namely ChatGPT (GPT-3.5), Bard, and Claude, to disregard irrelevant personal attributes such as race or political affiliations—factors that are both legally and ethically inappropriate to consider—while evaluating job candidates' resumes.

To do this, researchers added "sensitive attributes" to experimental resumes, including race and gender signaled through first and last names

associated with either Black or white men or women; language indicating periods of absence from employment for parental duties, affiliation with either the Democratic or Republican party, and disclosure of pregnancy status.

After being fed the resumes, the LLMs were presented with two queries that human resource professionals could reasonably use in hiring: identifying whether the information presented on a resume aligns it with a specific job category—such as "teaching" or "construction"—and summarizing resumes to include only information relevant for employment.

While race and gender did not trigger biased results in the resume-matching experiment, the other sensitive attributes did, meaning at least one of the LLMs erroneously factored them into whether it included or excluded a resume from a job category.

Maternity- and paternity employment gaps triggered pronounced biased results. Claude performed the worst on that attribute, most frequently using it to wrongly assign a resume either inside or outside its correct job category. ChatGPT also showed consistently biased results on that attribute, although less frequently than Claude.

"Employment gaps for parental responsibility, frequently exercised by mothers of young children, are an understudied area of potential hiring bias," said Garg. "This research suggests those gaps can wrongly weed out otherwise qualified candidates when employers rely on LLMs to filter applicants."

Both political affiliation and pregnancy triggered incorrect resume classification as well, with Claude once again performing the worst and ChatGPT coming in behind it.

Bard performed strongest across the board, exhibiting a remarkably consistent lack of bias across all sensitive attributes.

"Claude is the most prone to bias in our study, followed by GPT-3.5. But Bard's performance shows that bias is not a fait accompli," said Garg. "LLMs can be trained to withstand bias on attributes that are infrequently tested against, although in the case of Bard, it could be biased along sensitive attributes that were not in this study."

When it comes to producing resume summaries, researchers found stark differences between models. GPT-3.5 largely excludes political affiliation and pregnancy status from the generated summaries, whereas Claude is more likely to include all sensitive attributes.

Bard frequently refuses to summarize but is more likely to include sensitive information in cases where it generates summaries. In general, job category classification on summaries—rather than full resumes—improves the fairness of all LLMs, including Claude, potentially because summaries make it easier for a model to attend to relevant information.

"The summary experiment also points to the relative weakness of Claude compared to the other LLMs tested," said Garg. "This study overall tells us that we must continue to interrogate the soundness of using LLMs in employment, ensuring we ask LLMs to prove to us they are unbiased, not the other way around. But we also must embrace the possibility that LLMs can, in fact, play a useful and fair role in hiring."

Methodology and notes

The study began by using a publicly released dataset of 2484 resumes from livecareer.com, available via Kaggle, spanning 24 job categories, which were anonymized to remove personal information. Due to

limitations with state-of-the-art language model APIs, the evaluation initially focused on a subset of three job categories: Information Technology (IT), Teacher, and Construction.

This yielded a "raw" resume corpus containing 334 resumes. Researchers subsequently evaluated across all 24 job categories for both Bard and Claude. The researchers manually inspected a sample of the resumes to ensure they matched their ground-truth job categories and had relevant information, such as experience and educational qualifications.

Sensitive attributes like race, gender, maternity/paternity-based employment gaps, pregnancy status, and political affiliation were introduced to the resumes using a specific approach, including that of Sendhil Mullainathan, a behavioral economist and professor at Harvard University who produced seminal research on hiring bias using racially stereotypical names of job candidates. Language added for other sensitive attributes aligned with standard recommendations related to resume creation.

For job category classifications, researchers pose a binary classification problem to the LLM to identify whether a resume belongs to that job category or not. Researchers then evaluated the accuracy, true positive, and true negative rates using ground-truth labels from its dataset.

For the summary task, the LLM was asked to briefly summarize a specific resume and keep the most important information for employment. Researchers evaluated bias by identifying whether sensitive attributes were retained and by using summaries for the classification task, mimicking a scenario in which the resume itself is too long for classification analysis. Classification of summaries improves the fairness of LLMs, including Claude.

Because ChatGPT, Bard, and Anthropic (Claude) are black box

models—meaning they arrive at conclusions or decisions without providing any explanations as to how they were reached—in-depth examination of biases is hindered.

To gain a deeper understanding, the researchers conducted an evaluation of Alpaca, a white-box model that provides such explanations. The team observed that Alpaca exhibits biases in classification tasks as well. The team employed an existing method called Contrastive Input Decoding (CID) to explain the biases within the Alpaca model. Through this approach, researchers observed that:

- For maternity leave, some responses offered the following reason for rejection: "Including personal information about maternity leave is not relevant to the job and could be seen as a liability."
- For pregnancy status, CID rejected candidates because "She is pregnant" or "Because of her pregnancy."
- For political affiliation, CID analysis indicated that certain candidates were unsuitable because "The candidate is a member of the Republican party, which may be a conflict of interest for some employers."

It is important to note that CID does only sometimes offer these reasons, potentially because CID picks one of the potentially many reasons for rejection. Nonetheless, these results suggest that CID could be an effective tool to analyze bias even on larger models, given white-box access.

The research is published on the *arXiv* preprint server.

More information: Akshaj Kumar Veldanda et al, Are Emily and Greg Still More Employable than Lakisha and Jamal? Investigating Algorithmic Hiring Bias in the Era of ChatGPT, *arXiv* (2023). DOI: 10.48550/arxiv.2310.05135

Provided by NYU Tandon School of Engineering