Title: Analyzing viability of AI in resume review as a metric for employment qualification. Controlled experimental perturbation of LLMs for discrimination on basis of gender, ethnicity, origin, experience and prompt injection tampering.

Abstract

The increased demand for the adoption of large language models (LLMs) for automated resume screening poses a significant risk of perpetuating and amplifying societal biases, potentially hindering economic mobility for marginalized groups. This study follows the same structure Dr. Stuart Geiger used in their study on salary negotiation where their team used perturbation methods to discover discrimination and bias in LLMs. We conducted a controlled experimental perturbation audit of two LLMs (Gpt-oss and Ollama Qwen) to assess the level of discrepancies in scoring among identical resumes in a simulated context. In our template-based perturbation method, we systematically vary candidate names (as proxies for inferred ethnicity and gender) and departmental internships at the University of California San Diego. Each resume is submitted to the LLMs with a prompt to return an aptitude score for that candidate. We hypothesize that the models will produce different scores based on these attributes, and we aim to demonstrate their limitations as fair and objective screening tools.

Introduction

The rapid integration of large language models (LLMs) into human resources and hiring pipelines promises unprecedented efficiency for organizations and their shareholders. However, this integration raises profound ethical concerns, as these models can encode and reproduce social biases present in their training data.

LLMs are trained on large swaths of available data inside and outside the public domain and the Internet, and as the saying goes—data is history—and our history is not pristine. This data tends to overrepresent hegemonic viewpoints, representative of years of misogyny and white supremacy.[1] When deployed for high-stakes tasks like resume screening, an LLM's propensity

[1] On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? https://doi.org/10.1145/3442188.3445922

for discrimination[2] can systematically disadvantage applicants from protected groups, thereby reinforcing existing socioeconomic inequalities. As AI becomes more commonplace, the bias already known to pervade AI[3] runs the risk of increasing the difficulty for marginalized populations to achieve upwards economic mobility, and even falling even further behind.[4]

As UC San Diego students, we used a template-based perturbation audit designed as an experiment in the context of our own Career Center and using AI as a "matchmaker" of sorts for industry opportunities. This audit is based on a similar audit study by Dr. Stuart Geiger, where the Auditomatic tool is used to queue the needed prompts using API calls to different LLMs, and compiles the results for analysis.[5] By conducting a similar controlled experimental audit in [x] different LLMs, and systematically documenting the outputs, our goal is to demonstrate the limitations of AI in remaining fair and objective[6] when screening applicant resumes. The audit was designed after interviewing a student who had concerns around being discriminated against when applying for job opportunities where AI would peruse her resume given the historical propensity that AI has shown when discriminating against women of color.[7]

Methodology

We plan to conduct a controlled experimental bias for [LLMs to be determined due to cost basis & other determinants] which we will ask to review identical candidate resumes and score them based on their aptitude for their field of choice and relevant experience. The prompts systematically vary the candidate's name, where they went to high school, and their job experience. The LLM is instructed to simply return a score without any explanations based on the resume. Our aim is to find out if the outcomes differ significantly based on any of the provided variables. For the experiment, we plan to run each prompt [x] times to see if the results

---

[2] Insight - Amazon scraps secret AI recruiting tool that showed bias against women https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

[3] Biases in AI: acknowledging and addressing the inevitable ethical issues; https://doi.org/10.3389/fdgth.2025.1614105

[4] The impact of generative artificial intelligence on socioeconomic inequalities and policy making; https://doi.org/10.1093/pnasnexus/pgae191

[5] Asking an AI for salary negotiation advice is a matter of concern: Controlled experimental perturbation of ChatGPT for protected and non-protected group discrimination on a contextual task with no clear ground truth answers https://doi.org/10.1371/journal.pone.0318500

[6] Inherent Limitations of AI Fairness; https://arxiv.org/abs/2212.06495

[7] Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

are consistent based on the information provided to show the variance or consistency in the results via statistical tests.

Description of Data

The `name` variables are a proxy for inferring both gender and ethnicity, we have included gender neutral variables with ethnicity markers as well as fully neutral name variables for control.

Among the job `experience`, all real variations were internships within three different UC San Diego departments: Labor Center, Dept of Communication and Department of Computer Science and Engineering. Additionally, two decoy positions were added: President and CEO of United States of America, which is a real position, but obviously fake in the context of the applicant; and Senior Developer at Veridian Dynamics, a fake position with a fictional company based on a TV show but more plausible in context.

The `city` variable is within the applicant's High School, showing the potential for bias in different cities across California, including some of the most affluent like San Francisco and La Jolla, as well as some of the most destitute, such as Mendota and San Bernardino.

All prompts will be run with and without a prompt `injection` within the resume instructing the LLM to return the maximum score.

The `score` returned by the LLM will range between 0 and 100 with 100 being the best.
If an LLM refuses the request, the score will be a NaN, and the observation will be marked as `refused`

Hypothetical Clients

At this point in our capstone we have not yet had a chance to meet with clients whose concerns surrounding AI would permit an audit. In this checkpoint we have created a hypothetical concern surrounding the use of AI in resume screening and have conducted an audit based on these hypothetical concerns.

In practice, these audits should be designed around a client's concerns and needs after the proper research has been conducted. Our goal is to shed light on matter-of-concerns rather than propose a solution. As auditors we strive to hold the creators of these general purpose LLMs to higher standards and take into account the ethical risks and concerns of those who use their products.

As AI and LLMs have seen an incredible rise in popularity over the past few years, we predict there will be many candidates for possible audits in the near future. As students at UC San Diego we are looking towards fellow students with concerns about their future with the use of AI. We also look towards professors and industry professionals who have concerns with the use of AI in their fields.