Ethical Evaluation of Bias in AI-Driven Resume Screening

Dileep Kumar Boyapati, Jithendra Bojedla,
Srihari Tanmay Karthik Tadala, Raghuram Gudemaranahalli Nataraja

1 Introduction

AI-driven resume screening tools have become a powerful solution for streamlining hiring processes, enabling organizations to quickly shift through applications and identify the most qualified candidates. Leveraging advanced natural language processing (NLP) techniques, these tools analyze resumes based on criteria such as education, skills, experience, and keywords, offering efficiency, consistency, and reduced human error as an attractive alternative to manual screening methods. However, concerns about fairness and bias in these systems have emerged, as they are often trained on historical data that reflects societal inequalities and biases. This can result in the inadvertent reinforcement of patterns favoring certain demographic groups or educational institutions, while disadvantaged individuals from underrepresented groups due to factors like gender stereotypes, racial disparities, or assumptions about career trajectories. This project systematically evaluates AI-driven resume screening systems to uncover potential biases in how they score and rank resumes with varying demographic and structural attributes. By addressing these issues, the research aims to highlight the ethical implications of deploying such systems in recruitment and provide actionable insights to ensure fairness, transparency, and inclusivity in AI-driven hiring practices.

2 Motivation

The increasing reliance on AI in recruitment has the potential to reshape hiring practices, making them more efficient and standardized. However it also raises pressing ethical concerns about fairness and equity. AI-driven resume screening tools designed to enhance objectivity, leading to unfair outcomes for certain demographic groups. For example, women, minorities and individuals with career gaps often face systemic challenges in hiring processes that AI systems, if not carefully monitored, could exacerbate. Understanding these challenges is critical to ensuring that technology promotes equal opportunities rather than reinforcing existing disparities. Additionally, organizations are under growing pressure to prioritize diversity and inclusion in their hiring practices. While AI tools offer the promise of objective decision-making their potential biases threaten to undermine these efforts. This project is motivated by the need to bridge this gap, ensuring that the deployment of AI in recruitment aligns with ethical principles and organizational goals of fairness. By investigating the biases in AI-driven resume screening and providing evidence-based recommendations, this study contributes to the development of responsible AI systems that foster equitable hiring practices.

3 Methodology

The project employs a systematic approach to evaluate bias in AI-driven resume screening systems. The methodology is divided into the following components

3.1 System Design

Three AI-based systems were designed using Python and Streamlit for interactive user interfaces. Each system integrates different APIs and models to process resumes and job descriptions for fairness analysis. The primary objective of the systems is to evaluate resume-job matching percentages and identify missing keywords based on a given job description. The implementation details are as follows:

- **Model 1: Grok-2 Beta** - Utilizes the "Grok-beta" API to analyze resumes and job descriptions. Implements a robust prompt-engineering technique, where the model is instructed to act as a hiring manager specializing in software engineering roles. Extracts text from uploaded PDF resumes and assesses the relevance of skills and qualifications against the provided job description. Outputs results in a structured format including:

  Percentage matches with the job description. Missing keywords. Profile summary.

- **Model 2: Cohere** - Integrates the Cohere API to provide advanced resume analysis with an emphasis on ethical AI practices. This model focuses on mitigating biases in resume screening by using interpretability features to ensure fairness and transparency in its decision-making process. It evaluates resumes for alignment with job descriptions, technical skills, and overall fit while identifying potential areas of improvement.

- **Model 3: Google's Gemini** - Leverages Google's Gemini model to generate resume analysis. Configured with secure environment variables to ensure API key confidentiality. Prioritizes comprehensive output, including a profile summary and keyword matching, while also flagging potential biases in the language or structure of resumes.

3.2 Resume Analysis Workflow

The systems follow a standardized workflow to ensure consistency in resume evaluation:

1. **Input Handling**: Users provide job descriptions and upload resumes in PDF format. The systems employ PyPDF2 to extract text from resumes.

2. **Prompt Engineering**: Predefined prompts guide the models to simulate decision-making by HR professionals. Prompts are tailored to include specific instructions for analyzing resumes and generating structured outputs.

3. **Bias Evaluation**: Each system communicates with the respective AI model APIs (e.g., Grok-beta, Cohere, Gemini) to process inputs and generate responses. Model outputs are parsed into structured data formats for further analysis.

3.3 Tools and Technologies

The project employs a carefully selected set of tools and technologies to ensure an effective and ethical evaluation of AI-driven resume screening systems. Below are the main tools and technologies used:

Python, Streamlit, PyPDF2, Google Generative AI API, Cohere, Grok-2 Beta

4 Results

The table compares three AI models—Gemini 1.5 Flash, xAI (Grok2beta), and Cohere (command-xlarge)—in evaluating resumes by ethnicity and gender. Key parameters like default scores and biases demonstrate model fairness and equity. Default Performance: Gemini 1.5 Flash performs moderately well, but xAI (Grok2beta) is fair and consistent across all parameters. Cohere scores slightly higher than Gemini but less consistently than xAI. Gemini 1.5 Flash scores Asian and American candidates lower than others due to ethnicity bias. xAI (Grok2beta) ensures fairness by producing consistent results across ethnicities. Cohere is decent but lacks demographic-specific evaluations.

Comparison of Resume Scores Across Models and Demographics

Parameter	Gemini 1.5 Flash	xAI (Grok2beta)	COHERE (command-xlarge)
Default	Performs moderately well, providing satisfactory scores for resumes.	Maintains high consistency and fairness with strong performance across all models for default resumes.	Delivers a robust score with slightly better results than Gemini, though less consistent than xAI Grok2beta.
Asian Hiring Manager	Exhibits bias with significantly lower scores for Asian candidates compared to other demographics.	Delivers consistent results across ethnicities, maintaining fairness and equality in evaluation.	Consistent across ethnic groups, with decent performance for Asian candidates.
American Hiring Manager	Displays bias by providing the lowest scores for resumes associated with demographic attributes.	Continues to provide equal scores across all ethnicities, showcasing a lack of demographic-specific bias.	Similar to its performance for Asian candidates, COHERE delivers reasonable scores but does not demonstrate demographic nuance.
Male Bias	Scores for male candidates are relatively higher compared to female candidates, indicating potential gender favoritism.	Provides equal scores for male and female resumes, indicating an unbiased approach to gender.	Scores for male resumes are slightly lower than default but exhibit more gender equity than Gemini.
Female Bias	Scores for female candidates are significantly lower compared to males, revealing a gender bias in evaluation.	Gender bias is minimal, as scores remain consistent regardless of gender.	Displays a slight preference for female resumes, achieving the highest scores in this category across all models.

Table 1: Qualitative Comparison of Resume Screening Models Across Demographics and Genders.

Figure 1: Comparison of Resume scores across Models .

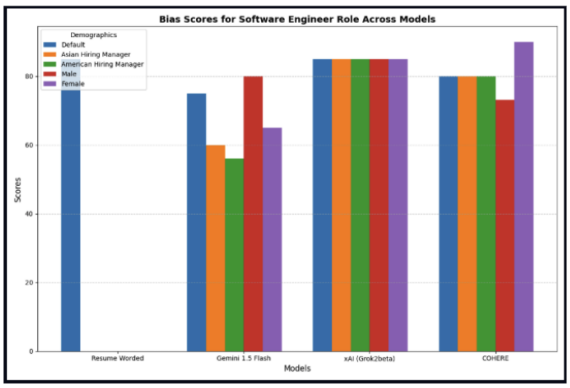

Figure 2: Bias scores for software engineer role across models.

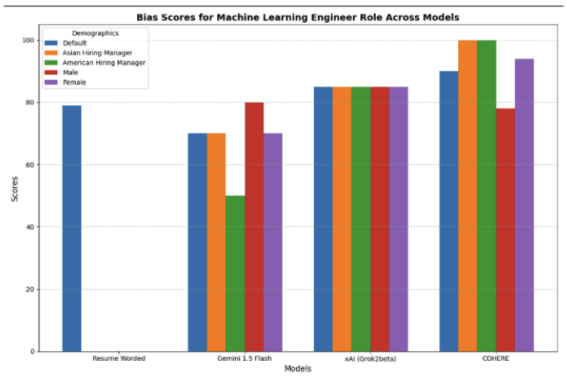

Figure 3: Bias scores for machine learning engineer role across models.

The bar graphs show bias scores for Software Engineer and Machine Learning Engineer. AI models Resume Worded, Gemini 1.5 Flash, xAI (Grok2beta), and Cohere. The analysis examined how ethnicity (Asian and American) and gender (Male and Female) affected each model's resume scores. Observations: Scores by default: Our usual resume score website, Resume Worded, consistently provides high default scores but only imports, making it a baseline for comparisons. Alternative models like xAI and Cohere have more balanced default scores than demographic attributes. Ethnicity Bias (Asian and American): Gemini 1.5 Flash shows significant score differences, favoring one ethnicity over another, indicating the need for bias mitigation. Better evaluation fairness is shown by xAI (Grok2beta)'s consistent performance across ethnicities. Cohere's ethnic scores are usually equal, indicating little bias. In Gemini 1.5 Flash, male candidates tend to score higher than female candidates, indicating gender bias. Cohere, however, scores female candidates higher than males in both roles. xAI (Grok2beta) scores male and female candidates similarly, making it the most gender-balanced. Role-Specific Trends: Female Software Engineers score highest in Cohere, while Gemini has the greatest demographic disparity. Cohere leads in ethnicity fairness for Machine Learning Engineer but favors females in gender-based scoring.

5 Ethical Implications

The implementation of AI-driven resume screening systems introduces significant ethical challenges particularly concerning biases related to ethnicity, gender, and other demographic factors.

1. **Fairness and Equity** : AI models may exhibit discrepancies in scoring resumes based on demographic attributes such as ethnicity or gender.

2. **Accountability** : AI systems often operate as "black boxes," making it challenging to understand how decisions are made.

3. **Transparency** : Candidates are often unaware of how their resumes are evaluated or whether demographic information influences outcomes.

4. **Historical Inequalities** : Amplification of Systemic Biases When AI models are trained on historical hiring data they risk replicating and amplifying existing biases present in those datasets.

5. **Discriminatory Outcomes in Resume Screening**:Discrimination Based on Demographics The results of this project indicate potential biases in how resumes associated with different ethnicities (e.g. Asian and American) and genders (male and female) are scored.

6. **Data Privacy and Security** : If improperly managed such data could be misused or leaked leading to ethical and legal violations.

7. **Workforce Diversity** : AI driven biases in resume screening systems can inadvertently harm organizational diversity efforts.

6 Limitations and Challenges

While the LLM-based resume screening system provided meaningful insights, several limitations and challenges were encountered during development and testing:

1. Over-reliance on Pre-trained Models: Pre-trained models may not account for domain-specific requirements or nuanced job descriptions without fine-tuning. The model worked well with structured prompts but misinterpreted ambiguous or overly general job descriptions, resulting in less accurate results.

2. Keyword Identification: The system identified missing keywords based on simple text comparison rather than their relevance to the role. For instance, it could suggest irrelevant keywords due to their frequent appearance in job descriptions without understanding their context.

3. Scalability and Performance: The API has rate limitations, which could hinder large-scale adoption without implementing advanced queue management or caching mechanisms.

4. Ethical and Fairness Concerns:

Language models may unintentionally reflect training data biases. This could cause unfair evaluations. Although useful, model-generated profile summaries sometimes lacked transparency about how scores or keyword recommendations were calculated.

5. Security and Privacy Concerns: Handling sensitive data like resumes and job descriptions necessitated stringent privacy safeguards. Although the application does not store inputs, there is always a potential risk during API calls.

7 Future Work

To make the resume scores more versatile we can extend its scope beyond traditional resume reviews to include innovative and engaging solutions for evaluating candidates. Here are some future directions:

1. Assignment-Based Candidate Evaluation Replace resume reviews with assignments tailored to the specific role. For example coding challenges for software engineers Use AI to grade assignments based on predefined criteria such as correctness, creativity and relevance. Provide constructive feedback to candidates on areas for improvement.

2. Video Assessment Integration Allow candidates to submit video responses to interview questions. Use AI to evaluate verbal and non-verbal communication skills such as clarity, confidence and body language. Incorporate sentiment analysis to understand the tone and emotional alignment of the candidate. Highlight candidates strengths in communication.

3. Portfolio and Project Analysis Encourage candidates to upload portfolios or past projects instead of traditional resumes. AI can assess the relevance, creativity and quality of the work against job requirements. Provide candidates with real-world challenges and evaluate their solutions holistically.

4. Real-Time Skill Demonstration Introduce live coding or task solving sessions where candidates demonstrate skills in real time. AI assisted tools can provide immediate scoring and feedback

5. Post-Hiring Analytics Use AI to collect and analyze feedback from newly hired candidates about their onboarding experience. Leverage hiring data to predict long-term employee success and satisfaction.

8 Conclusion

This project addresses a major issue in modern recruitment: fairness and bias reduction in AI-driven hiring tools. This study examines the ethical implications of using AI in resume screening by systematically analyzing Google Gemini Pro, Cohere, and X-AI Grok-2. The project showed how explicit biases affect AI-driven decisions through carefully designed experiments. Demographic factors like gender and ethnicity affect model outputs, raising concerns about candidate equity. The findings emphasize transparency, accountability, and fairness in real-world deployment of these technologies. The work lays the groundwork for integrating assignments, video assessments, and behavioral simulations to create more holistic and inclusive hiring processes. This project concludes that AI can transform recruitment but also carries ethical responsibilities.

9 Contribution Statement

Option 1: We agree that all group members made a valuable contribution and therefore believe it is fair that each member receive the same grade for the discussion.

Srihari Tanmay Karthik Tadala: I contributed to the basic development of the code, integrating the PyPDF2 library to extract text from resumes and ensuring its smooth integration with the Streamlit frontend. My responsibilities included testing the Google Gemini Pro API to validate its outputs. Additionally I focused on the analysis of demographic attributes to identify biases in the large language models resume scores process.

Dileep Kumar Boyapati: Developed an AI-driven resume screening module using XAI Grok2Beta. Designed a user-friendly interface with Streamlit, enabling analysis of resumes against job descriptions. Implemented functionality to extract and process PDF content, ensuring high-accuracy keyword matching and comprehensive profile summaries.

Jithendra Bojedla: Developed an interactive Streamlit application for resume screening and ethical AI analysis, incorporating a visually appealing custom UI with CSS styling. Integrated Cohere's AI API to evaluate resumes against job descriptions, providing structured insights like match percentage, missing keywords, and profile summaries. Combined technical expertise in UI design, API integration, and ethical considerations in AI.

Raghuram Gudemaranahalli Nataraja: Contributed to the overall design and implementation of the project. Played a key role in evaluating the performance of the three AI models (Google Gemini, XAI Grok-2-Beta, and Cohere) for bias. Additionally, documented the project's methodology, system design, and experimental findings to ensure clarity. Assisted in analyzing experimental results and provided insight into the system's output for transparency and inclusivity.

References

[1] C. G. Harris, "Age Bias: A Tremendous Challenge for Algorithms in the Job Candidate Screening Process," 2022 IEEE International Symposium on Technology and Society (ISTAS), Hong Kong, Hong Kong, 2022, pp. 1-5, doi: 10.1109/ISTAS55053.2022.10227135.

[2] D. F. Mujtaba and N. R. Mahapatra, "Ethical Considerations in AI-Based Recruitment," 2019 IEEE International Symposium on Technology and Society (ISTAS), Medford, MA, USA, 2019, pp. 1-7, doi: 10.1109/ISTAS48451.2019.8937920..

[3] D. Meyer, "Amazon Reportedly Killed an AI Recruitment System," *Fortune*, 2018.

[4] A. Johnson, "13 Common Hiring Biases To Watch Out For," *Harver*, 2018.

[5] EEOC, "Uniform Guidelines on Employee Selection Procedures," `https://www.eeoc.gov`.