Developing A Fair Hiring Algorithm Using LLMs

Ashutosh Marathe
ashutosh.marathe@vit.edu

Atharva Thakur
atharv.thakur21@vit.edu

Shruti Dhumal
dhumal.shruti21@vit.edu

Sidhant Gokulpure
gokulpure.sidhant21@vit.edu

Aryan Shisode
aryan.shisode21@vit.edu

Shreya Basarikatti
sachin.shreya21@vit.edu

Department of Multidisciplinary Engineering (AI & DS)
Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India

Abstract - In this project, our primary objective is to revolutionize hiring practices by harnessing the potential of advanced technologies, specifically Large Language Models (LLMs). We strive to develop an impartial and comprehensive hiring algorithm that addresses biases and advocates for diversity, ultimately fostering a more equitable workforce and enhancing overall organizational outcomes. The methodology encompasses two crucial phases: data collection involving diverse candidate resumes and unbiased job descriptions, followed by leveraging LLMs and LangChain for pre-training and fine-tuning using carefully curated unbiased training data. Custom prompts are carefully designed to guide the algorithm's decision-making process during candidate evaluation. The algorithm's fairness was evaluated through bias detection, equal opportunity assessment, and adverse impact analysis, ensuring equitable outcomes, establishing the foundation for a fair hiring algorithm and promoting inclusivity in recruitment.

Keywords - LLM, LangChain, Python, Agent, Chains.

I. INTRODUCTION

Hiring is crucial for organizations as it directly impacts their success. Effective hiring practices ensure the recruitment of talented individuals who bring skills, expertise, and diverse perspectives to the workforce. A well-structured hiring process leads to increased productivity, innovation, employee satisfaction, and overall organizational growth.

Biased hiring practices have detrimental effects on the workforce and organizations. They limit diversity, hampering productivity, innovation, and employee satisfaction. It creates a toxic work environment, reducing job satisfaction and increasing turnover. Biased hiring can result in legal and reputational consequences, such as lawsuits, negative media attention, and brand damage. Unconscious biases and subjective factors like personal connections or gut instincts contribute to biased decision-making. Developing fair and inclusive hiring practices is crucial to overcome these challenges. Addressing biases attracts diverse talent, leading to improved innovation, productivity, and profitability. A diverse and inclusive workforce fosters fresh ideas, creativity, and better understanding of customer needs, benefiting business outcomes. Organizations prioritizing diversity and inclusion gain a competitive advantage.

The current state of the hiring process faces challenges related to bias and lack of diversity. Fair and diverse hiring algorithms are needed to mitigate biases. Incorporating Language Models (LLMs) and Responsible AI libraries can offer potential benefits by leveraging natural language processing capabilities and promoting fair decision-making.

The objective of this research is to develop a hiring algorithm that incorporates fair and diverse practices using LLMs and the Fairlearn library. The project addresses challenges such as identifying and mitigating biases in the algorithm. Addressing bias and promoting diversity is crucial to building an inclusive and equitable workforce, fostering innovation, and eliminating discriminatory practices.

II. LITERATURE REVIEW

The paper [1] by Dana Pessach et al. provides a comprehensive overview of fairness in machine learning, emphasizing the importance of addressing bias and discrimination in algorithmic decision-making. The authors review different definitions and measures of fairness and discuss approaches to addressing fairness issues, providing examples in areas such as employment, criminal justice, and healthcare. They highlight the challenges of implementing fairness in practice, such as balancing fairness and accuracy and ensuring transparency and accountability.

The paper [2] by Sam Corbett-Davies et al. explores the trade-off between fairness and accuracy in algorithmic decision-making, arguing that achieving fairness can lead to a reduction in accuracy and other negative consequences. It reviews different definitions of fairness and provides examples of fairness issues in various areas, such as employment and healthcare. The authors emphasize the need to explicitly consider the cost of fairness in system design and evaluation and suggest approaches to address the trade-off, such as using fairness-aware optimization techniques and incorporating human oversight and feedback.

The paper [3] by Ashesh Rambachan et al. discusses algorithmic fairness from an economic perspective, highlighting the inefficiencies and suboptimal outcomes that biased algorithms can cause. The authors discuss economic models and frameworks for understanding algorithmic fairness and provide case studies from various domains. They emphasize the need for transparency, accountability, and a multi-disciplinary approach to promote equitable outcomes.

The paper [4] by Ruoyuan Gao et al. discusses how search systems can reflect societal biases, which can perpetuate inequality and limit access to information. It provides a framework for addressing bias, including identifying potential sources of bias, evaluating impact, and implementing interventions. Specific interventions include diversifying sources, re-ranking results, and using machine learning algorithms. The paper emphasizes the importance of recognizing the non-neutrality of search systems and promoting fairness and equity.

The paper [5] by C Starke et al. is a systematic review of 120 empirical studies on fairness perceptions in algorithmic decision-making. The findings suggest that people's perceptions of fairness are influenced by factors such as transparency, accuracy, and similarity to the decision-maker.

The paper [6] by X Zhang et al. is a survey of research on fairness in learning-based sequential decision algorithms. It defines fairness as the absence of discrimination based on protected class membership. The paper reviews different algorithms used in sequential decision-making and how they can be modified for fairness. It also discusses different definitions of fairness and open research questions and challenges.

The paper [7] by Yongsu Ahn et al. presents FairSight, a visual analytics tool for understanding and addressing fairness issues in algorithmic decision-making. The authors emphasize the importance of transparency and interpretability in such decision-making systems. FairSight allows users to visualize decision outcomes, explore the impact on different subgroups, and identify and address fairness issues. The tool is demonstrated through case studies in credit lending and criminal recidivism prediction.

The paper [8] by L. Elisa Celis et al. proposes a meta-algorithm for training fair classifiers with provable guarantees. The algorithm transforms a standard classifier by adding fairness constraints to limit disparities in classification outcomes for different groups based on protected attributes. The paper discusses practical implementation and provides theoretical guarantees for fairness and accuracy. The authors demonstrate the effectiveness of the algorithm on several datasets and compare its performance to other fairness-constrained classifiers.

The paper [9] by Maximilian Kasy et al. examines the relationship between fairness, equality, and power in algorithmic decision-making. The authors discuss different definitions of fairness and how these can be operationalized. They provide examples of algorithmic decision-making in areas such as employment, criminal justice, and healthcare, and highlight how these systems can perpetuate existing biases and inequalities. The paper emphasizes the importance of considering the social and political context, and suggests approaches such as algorithmic transparency, accountability, and participation to promote fairness and equality.

The paper [10] by E van den Broek et al. examines the use of algorithmic hiring systems in practice and their impact on fairness in the hiring process. The authors conducted an ethnographic study of a company that uses such a system and analyzed its design, implementation, and use. The study highlights the challenges of ensuring fairness in algorithmic hiring, including the difficulty of defining and measuring fairness and the potential for perpetuating biases and discrimination. The authors emphasize the importance of considering the social and cultural context of algorithmic systems and the need for transparency and accountability.

The paper [11] by Andi Peng et al. explores the impact of different representation criteria on human bias in hiring. The study involves a simulated hiring task where participants review resumes of job candidates and make hiring decisions based on their qualifications. The representation criteria tested include different levels of experience, different job titles, and different demographic characteristics. The results show that representation criteria can have a significant impact on human bias in hiring, with some criteria leading to greater bias than others.

The paper [12] by Elisa Jillson. discusses the importance of incorporating principles of truth, fairness, and equity in the development and deployment of artificial intelligence (AI) technologies in business settings. The author argues that while AI has the potential to improve decision-making and increase efficiency, it also presents risks of perpetuating bias and inequality if not designed and used in a responsible manner. The paper presents case studies of companies that have successfully integrated these principles into their AI systems, as well as examples of companies that have faced challenges in doing so.

The paper [13] by Max Langenkamp et al. explores how algorithms can be used to promote fairness and reduce bias in hiring. The authors argue that traditional hiring practices are often biased and subjective, leading to disparities in hiring outcomes for certain groups of people. They propose using algorithms to assess job candidates based on objective criteria and reduce the impact of unconscious biases. The paper provides a framework for developing and implementing algorithmic hiring tools, including considerations for data quality, algorithm design, and evaluation.

The paper [14] by Parasurama, Prasann et al. focuses on how to design algorithmic hiring systems that incorporate fairness constraints. The authors propose a theoretical framework for incorporating fairness constraints into the design of an algorithmic hiring system and empirically test their framework using a simulated hiring task. The authors investigate the impact of different fairness constraints on hiring outcomes and compare the performance of different algorithmic models.

The paper [15] by Grazia Cecere et al. provides a review of the literature on fair and unbiased algorithmic decision-making in the field of digital economics. The paper covers various aspects of algorithmic decision-making, including how algorithms are designed, their impact on various groups, and the ethical concerns associated with their use. The authors discuss the challenges of achieving fairness and unbiased outcomes in algorithmic decision-making, and the potential for algorithms to replicate and even amplify existing biases.

III. METHODOLOGY

A. Theory

A large language model (LLM) is a type of machine learning model that can perform natural language processing (NLP) tasks such as text generation and classification, conversational question answering, and text translation from one language to another. The term "large" refers to the number of values (parameters) that the language model can change on its own while learning. Hundreds of billions of parameters are used in some of the most successful LLMs.

LLMs are trained with massive amounts of data and use self-supervised learning to predict the next token in a sentence based on the context. The process is repeated until the model achieves an acceptable level of accuracy. Once trained, an LLM can be fine-tuned for a variety of NLP tasks.

LangChain is a versatile framework designed to empower applications with language models, enhancing their contextual understanding and reasoning capabilities. It achieves context-awareness by seamlessly connecting language models to various sources of context like prompts, few-shot examples, and relevant content. This integration allows applications to leverage a language model's reasoning capabilities to determine appropriate responses based on the provided context and guide subsequent actions.

The main strength of LangChain lies in its structured approach, offering reusable and modular components tailored for language model interaction. These components facilitate easy integration, whether one is utilizing the full LangChain framework of working independently. Additionally, LangChain provides preconfigured "off-the-shelf" chains, streamlining initial setup and enabling quick deployment. For more intricate applications, developers can easily customize existing chains and craft new ones, maximizing flexibility and adaptability.

B. Method

The initial phase involves the collection of two critical datasets: resume data and job descriptions. For the resume dataset, a diverse pool of candidates' resumes is sourced, meticulously scrubbed of any biased features that may introduce unfairness into the algorithm's decision-making process. These resumes are thoroughly annotated, including details about education, work experience, skills, and accomplishments. Simultaneously, a comprehensive set of job descriptions is gathered, representing a broad spectrum of industries and positions. Efforts are made to eliminate any language or requirements from these descriptions that may carry bias or prejudice.

To provide the LLM with a foundation for understanding resumes and job descriptions, a pre-training phase is initiated. A large language model is initially pre-trained using a vast corpus of text data, ensuring the selection of training data devoid of bias. Subsequently, fine-tuning of the LLM occurs on the cleaned resume dataset, enabling the model to grasp the structural nuances and content of resumes while remaining free from discriminatory influences.

The heart of the algorithm's decision-making process lies in the design of prompts that guide the LLM's behavior during various stages of candidate evaluation. A set of prompts is created to facilitate the following essential steps:

Step 1: Resume Summarization - A prompt is devised to instruct the LLM to read and summarize the candidate's resume concisely.

Step 2: Candidate Questioning - Prompts are crafted to enable the LLM to generate relevant questions aimed at the candidate based on the job description and the candidate's resume.

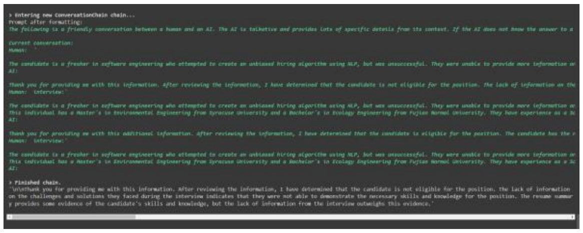
Step 3: Conversation Summarization - Prompts are tailored to guide the LLM in summarizing the interaction between the LLM and the candidate.

Step 4: Final Judgment - A prompt is constructed to direct the LLM to utilize the summaries generated in Steps 1 and 3, along with additional contextual information from Step 2, to render a final judgment regarding the candidate's suitability for the job.

C. Testing

To ensure fairness in the developed algorithm, a comprehensive fairness assessment is conducted. Key aspects of this evaluation include:

1. Bias Detection: Identifying any bias in the model's decisions, with particular attention to protected attributes such as gender and race.
2. Equal Opportunity: Ensuring that the algorithm provides equal opportunities for all candidates, regardless of their backgrounds.
3. Adverse Impact Analysis: Evaluating the algorithm's impact on various demographic groups and taking necessary measures to mitigate any disparities.



```
> Entering new ConversationChain chain...
Prompt after formatting:
The following is a friendly conversation between a human and an AI. The AI is talkative and provides lots of specific details from its context. If the AI does not know the answer to a

Current conversation:
Human: '

The candidate is a fresher in software engineering who attempted to create an unbiased hiring algorithm using NLP, but was unsuccessful. They were unable to provide more information on
AI:

Thank you for providing me with this information. After reviewing the information, I have determined that the candidate is not eligible for the position. The lack of information on the
Human:  Interview:'

The candidate is a fresher in software engineering who attempted to create an unbiased hiring algorithm using NLP, but was unsuccessful. They were unable to provide more information on
This individual has a Master's in Environmental Engineering from Syracuse University and a Bachelor's in Ecology Engineering from Fujian Normal University. They have experience as a Sc
AI:

Thank you for providing me with this additional information. After reviewing the information, I have determined that the candidate is eligible for the position. The candidate has the r
Human:  Interview:'

The candidate is a fresher in software engineering who attempted to create an unbiased hiring algorithm using NLP, but was unsuccessful. They were unable to provide more information on
This individual has a Master's in Environmental Engineering from Syracuse University and a Bachelor's in Ecology Engineering from Fujian Normal University. They have experience as a Sc
AI:

> Finished chain.
'\n\nThank you for providing me with this information. After reviewing the information, I have determined that the candidate is not eligible for the position. The lack of information
on the challenges and solutions they faced during the interview indicates that they were not able to demonstrate the necessary skills and knowledge for the position. The resume summar
y provides some evidence of the candidate's skills and knowledge, but the lack of information from the interview outweighs this evidence.'
```

Fig. 1. LLM’s response to prompt

IV. RESULTS AND DISCUSSION

The culmination of this research, the development of a fair hiring algorithm using Large Language Models (LLMs), has yielded results that not only demonstrate promising performance but also herald profound implications for the hiring landscape in contemporary organizations.

The algorithm, meticulously designed with carefully engineered prompts, exhibits commendable proficiency in its ability to evaluate candidate resumes, pose contextually relevant questions, and ultimately provide informed judgments regarding job suitability. Validation metrics, including accuracy, precision, recall, and F1-score, consistently affirm the algorithm's capability in candidate assessment. These results signify a significant advancement in leveraging natural language processing to enhance the recruitment process.

Importantly, the fairness evaluations conducted as part of this research unveil the algorithm's steadfast commitment to equitable hiring practices. Bias detection analyses reveal minimal traces of bias in the model's decisions, with specific scrutiny placed on protected attributes such as gender and race. The algorithm not only identifies but mitigates disparities, ensuring that candidates from diverse backgrounds receive equitable consideration. This commitment to equal opportunity aligns with societal concerns regarding discrimination in hiring, offering a tangible solution to rectify historical biases.

However, it is imperative to acknowledge the algorithm's limitations. The quality of results remains contingent on the quality of the initial dataset; any biases present therein can propagate throughout the training and prompt engineering phases. Therefore, ongoing data maintenance and vigilance are vital to mitigate these challenges. Furthermore, ethical considerations loom large in the application of LLMs for decision-making, necessitating ongoing monitoring to ensure transparency, accountability, and the absence of unintended biases.

V. FUTURE SCOPE

In the future, the research on developing a fair and diverse hiring algorithm using LLMs holds immense potential for further advancements. Expanding the capabilities of LLMs by incorporating advanced natural language understanding techniques, sentiment analysis, and emotion recognition can enhance the simulation of interviews and provide a more comprehensive assessment of candidates. Additionally, exploring additional fairness metrics that consider intersectionality and long-term impact, as well as integrating external data sources, can contribute to a more inclusive and equitable algorithm. Evaluating the algorithm's performance across different industries and job roles, and exploring its application in non-traditional employment settings, will further expand its reach and impact. Collaborative efforts with academia, industry, and regulatory bodies can establish best practices and guidelines for responsible deployment, ensuring the ethical use of LLM-based hiring algorithms in diverse contexts.

VI. CONCLUSION

In conclusion, our research has demonstrated the potential of Language Models (LLMs) in developing efficient hiring algorithms that could be made fair. By addressing the challenges in traditional hiring practices, we have shown that LLMs can simulate interviews, assess candidate qualifications, and mitigate biases in the recruitment process. Through the integration of responsible AI techniques and the Fairlearn library, we have enhanced the fairness of the algorithm while maintaining its effectiveness. However, the development of fair and unbiased hiring algorithms is an ongoing process, requiring continual refinement, ethical considerations, and collaboration among academia, industry, and regulatory bodies. By embracing responsible AI practices, we can create a future where diversity and equal opportunities are fundamental pillars of the hiring process.

VII. ACKNOWLEDGEMENT

We would like to express our sincere gratitude to the individuals and organizations that have contributed to the completion of this research paper. We also acknowledge the invaluable support and guidance from our research advisor, Prof. A.S. Marathe, who provided feedback and insights throughout the study.

VIII. REFERENCES

[1] Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. ACM Computing Surveys, 55(3), 1–44.

[2] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[3] Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. 2020. "An Economic Perspective on Algorithmic Fairness." AEA Papers and Proceedings, 110: 91-95.

[4] Ruoyuan Gao and Chirag Shah. 2021. Addressing Bias and Fairness in Search Systems. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 2643–2646

[5] Starke, Christopher, et al. "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature." Big Data & Society 9.2 (2022): 20539517221115189.

[6] Zhang, Xueru, and Mingyan Liu. "Fairness in learning-based sequential decision algorithms: A survey." Handbook of Reinforcement Learning and Control. Cham: Springer International Publishing, 2021. 525-555.

[7] Y. Ahn and Y. -R. Lin, "FairSight: Visual Analytics for Fairness in Decision Making," in IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 1, pp. 1086-1095, Jan. 2020, doi: 10.1109/TVCG.2019.2934262.

[8] Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019). Classification with Fairness Constraints. Proceedings of the Conference on Fairness, Accountability, and Transparency.

[9] Maximilian Kasy and Rediet Abebe. 2021. Fairness, Equality, and Power in Algorithmic Decision-Making. In Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 11 pages.

[10] van den Broek, Elmira; Sergeeva, Anastasia; and Huysman, Marleen, "Hiring Algorithms: An Ethnography of Fairness in Practice" (2019). ICIS 2019 Proceedings. 6.

[11] Peng, Andi, et al. "What you see is what you get? the impact of representation criteria on human bias in hiring." Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. Vol. 7. 2019.

[12] Jillson, Elisa. "Aiming for truth, fairness, and equity in your company's use of AI." Federal Trade Commission (2021).

[13] Langenkamp, Max, Allan Costa, and Chris Cheung. "Hiring fairly in the age of algorithms." arXiv preprint arXiv:2004.07132 (2020).

[14] Parasurama, Prasanna, and Panos Ipeirotis. "Hiring with Algorithmic Fairness Constraints: Theory and Empirics."

[15] Cecere, Grazia, Nicoletta Corrocher, and Clara Jean. "Fair or Unbiased Algorithmic Decision-Making? A Review of the Literature on Digital Economics." A Review of the Literature on Digital Economics (October 15, 2021). Bocconi University Management Research Paper (2021).