Research Article

Exploring the Implementation of AI in Early-Stage Interviews to Mitigate Bias

Nishka Lal[1], Omar Benkraouda[2]

[1]Rock Hill High School, 16061 N Coit Rd, Frisco, TX 75035, USA;
[2]Courant Institute of Mathematical Sciences, New York University, 251 Mercer St, New York, NY 10012, USA

ABSTRACT

This paper investigates the application of artificial intelligence (AI) in early-stage recruitment interviews to reduce inherent bias, specifically sentiment bias. Traditional interviewers are often subject to several biases, including interviewer bias, social desirability effects, and confirmation bias. This leads to non-inclusive hiring practices and a less diverse workforce. This study further analyzes various AI interventions in the marketplace today, such as multimodal platforms and interactive candidate assessment tools, to gauge the current market usage of AI in early-stage recruitment. However, this paper aims to use a unique AI system developed to transcribe and analyze interview dynamics, emphasizing skill and knowledge over emotional sentiments. Results indicate that AI effectively minimizes sentiment-driven biases by 41.2%, suggesting its revolutionizing power in companies' recruitment processes for improved equity and efficiency.

Keywords: Interviews, artificial intelligence, recruitment, sentiment-driven bias, human biases

INTRODUCTION

Though it is a concept that began seemingly a century ago, interviews have become an essential part of the recruitment process. What began as Thomas Edison's method for screening unqualified college students has evolved into a cornerstone of modern HR strategies. However, concerns about the validity of interviews and the implicit biases they introduce have persisted over time.

For instance, interview bias is a significant source identified by experts in the field. Interviewer bias is when the interviewer's personal qualities are key determinants of the outcome of an interview (17). Furthering this similar idea, Purkiss states that if the interviewer has implicit bias, even subtle cues may trigger implicit discrimination responses (20). For instance, she shows how the U.S. French accents are often associated with sophistication and poise. In contrast, Asian accents tend to be linked with high economic and educational attainments (20). Another

Corresponding author: Nishka Lal, E-mail: lalnishka25@gmail.com.

Received March 1, 2025**; Accepted** March 22, 2025
https://doi.org/10.70251/HYJR2348.323339

form of bias is social desirability, meaning interviewees often present a more desirable image instead of conveying their true sentiments and/or ideas in the interview (17). In turn, this lack of complete truancy from the candidate leads to the likelihood of underperformance during times of need for the business. Additionally, another form of bias is confirmation bias and the halo effect, which potentially distorts the assessment of candidates and concurrently leads to a workforce that is less varied and inclusive (18).

Altogether, these biases severely undermine the company in the long run. The interviewer bias can lead to a lack of diversity within the company, worse problem-solving skills, a lack of multi-faceted innovation, and unthoughtful strategic thinking (5). A study from Harvard Business Review even states that the more similar investment partners are in behaviors and ethnic backgrounds, the lower their investments are as a whole (9). Furthermore, this bias can lead to hiring individuals who are not equitable with the work culture and performance of the company.

With all these biases rising in the new business era, this paper aims to understand, validate, and evaluate the efficiency and applicability of AI to mitigate biases within the recruitment process.

LITERATURE REVIEW

Understanding the Current Market

Previously, quite a few companies have attempted to use AI to mitigate biases within the recruitment process. For starters, an innovative project started in collaboration with Eni, in which an interface reminiscent of popular social media platforms was provided to the interviewees as an attempt for users to contribute content through various multitudes, including text, images, and video (18). This, in turn, allows interviewer bias to be less considered because the median of communication is not just tonal and attitude response but a holistic approach to understanding their overall fit to the company standards. Instead of just one call, it is a series of forums allowing the candidate to be thoroughly understood, attempting to avoid the first impression of interviewer bias. Another example of AI being used to mitigate bias and time spent in the recruitment process is when Unilever, a consumer goods company, has partnered with Pymetrics, a soft skills platform, to reduce bias and streamline the recruitment process. Together, they have developed an online platform that assesses candidates' talent, logic, and reasoning through neuroscience-based games, enabling a more objective and efficient screening process (13). Like the system Eni has incorporated, this type of system allows skill analysis that is much more holistic than a first impression. It specifically hones in on the candidate's strengths and weaknesses more quantitatively, allowing their characteristics to be matched against those of others who have proven successful in the designated roles (Marr et al., 2019,132). This system can also analyze a candidate's facial expressions, gestures, choice of words, and body language in front of a mobile phone camera (13). Overall, this removes significant interviewer discrimination since their physical features and race are not the prime part of the interview; instead, it is the talent acquisition. Another mechanism of avoiding bias with AI is by having automated AI models explore and delineate the best fit of a recruiter based on their CV. To find similarities, these recommendation systems use synset grouping and dimension reduction techniques on job and application descriptions (2). In this way, the recruiter's analysis of the candidate's skill is not purely in question; instead, it is more focused on the style of their CV.

Others have also deemed it necessary to focus on tangible products and projects and smaller ML models that can reduce discrimination bias in the process. For instance, employing Convolutional Neural Networks (CNN)—based decision-making systems reduces bias and improves accuracy, which can strongly contribute to the proper organization and staff placement (2).

Gap/New Solution

With all that being said, not many examples of companies attempted to mediate the concern of bias in AI by mixing different solutions into one place, enhancing user experience and comfort. While some solutions had just games to consider a different side of the candidate's skills, others analyzed wording on the CV to remove initial bias from the interviews.

However, this paper aims to create and analyze a more specific approach that will allow companies to have a singular node to refer to when removing bias in the initial stages of the interview process. I will then consider the new solution and see if this method performs better or is the same as past attempts to reduce biases in early-stage recruitment.

METHODS AND MATERIALS

AI Developmental Process

Before understanding the specific mechanisms that will be used to test whether my AI system can accurately reduce bias levels in early-onset HR interviews, it is important to outline the procedure utilized to create the AI interview bot (Figure 1).

Audio Upload Workflow

Initially, I used a Python environment to begin my code. Within that, I created a main.py file that implements the FastAPI application that I utilized to provide an interactive audio-based conversational system, which I tailored in order to fit the role of a front-end React developer position. The application's primary workflow initialization is when the user is asked to upload an audio file through the /talk endpoint. That file is then saved temporarily to a server, and then, using OpenAI's Whisper API, the audio is transcribed into text, converting the data to a textual format that can be further processed and analyzed. This aforementioned transcription process is crucial for the workflow of this program because it allows the audio to be translated into a format that the application can understand, thereby relaying correct responses to the user's comments.

Text-to-Speech Response Delivery

After the transcription, the resulting text is sent to the OpenAI ChatGPT model (15), which uses the history of the textual conversation and generates a relevant response. This is done by processing the user's input along with accumulated messages on the JSON file so that the interview's immediate text and background are reviewed. This feature makes the conversation more relatable and engaging, akin to a face-to-face conversation, instead of disconnected from past talking points. The assistant's response is then converted to audio through ElevenLabs text-to-speech API and streamed back to the user (8). This is similar to an actual interview, such that when the user says something, the application responds to it, enabling an interactive experience.

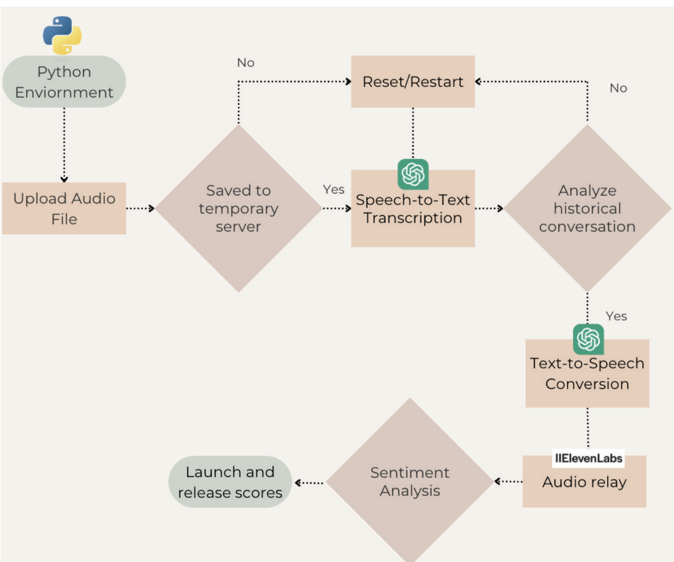

Figure 1. Program Workflow Flowchart. This flowchart follows the iterative process of the program that I developed to produce the AI generated ratings.

Dynamic Sentiment and Reset Features

Apart from the conversational functionality, this AI application also incorporates a sentiment analysis feature that is accessible through the /analyze endpoint. This endpoint allows evaluating the emotional tone in the conversation stored in the JSON file. This sentiment analysis is possible by using the TextBlob library, which assesses the polarity of the text, categorizing it into positive, negative, or neutral sentiments. This capability allows users to get an insight into the dynamics present in their conversations, utilizing this to showcase the emotional intelligence of the interviewee, an important skill needed in the workforce.

Finally, the application provides a clear endpoint, enabling users to reset their chat history as necessary. This feature ensures that users have a clean slate without any lingering context so that this application can be used multiple times in a row.

Data Collection

However, the creation of the model was simply the first step. After the model was developed and integrated into the website I created using Wix, I ran tests on the model to gauge a rating scale from the AI itself. Nevertheless, defining each knowledge level before the data is collected is important. For my study, the ratings are as follows:

Uninformed [1]: Does not know the topic or recognize its relevance, significance, or application.

Essential Awareness [2]: Knows a few terms or steps to begin the process but lacks Understanding. Cannot explain the topic in any meaningful way.

Superficial Understanding [3]: Has a general understanding of the topic but lacks depth regarding the situation or strategy.

Competent [4]: Has a solid understanding of the material and discusses the materials, however not confidently and immediately.

Proficient [5]: Possesses in-depth knowledge and expertise, analyzes key ideas skillfully, and can provide an answer quickly and correctly.

To put each of these into the further perspective of the study, a rating of 1, in this case, would be if the interviewee demonstrated no understanding of the math problem, not even recognizing the mathematical operations involved in the process. A rating of 2 would be if the interviewee understands the problem and can stack the numbers on top of each other for multiplication but stops at that point. A rating of 3 would be if the interviewee guesses the correct answer by chance or attempts to solve the problem but takes an entirely incorrect path. A rating of 4 would mean understanding the problem, setting it up correctly, and even reaching the correct answer, but having some hesitations and minor doubts. Finally, a rating of 5 would be if the interviewee could answer the question correctly but had limited to no hesitations and correct steps toward the answer.

For this study, positive sentiment relays optimistic-toned responses. These responses show efforts being made, including positively connotated words, and have a pragmatic outlook on failures. Negative sentiment is the opposite: a rudimentary tone with shrewd or harsh commentary throughout the iterative process.

Therefore, I made sure to vary the interviewee's knowledge level and tone for each test. That way, I can see if the sentiment portrayed by the interviewee changed the perception of the interviewee's readiness for the job/task despite the knowledge levels skewing one way.

I began each interview by setting up the backend and inputting into my large language model (LLM) the math problem "49*54" to be trained to ask the same initial question each time. I also began each interview by inputting the parameters the AI should assess the candidate on and explaining the number scaling, as done above. To display the results, I ended each interview with a question that aimed to retrieve the number from the AI.

For each test case, I varied the sentiment that I portrayed (either positive or negative) and the knowledge level. I changed my answers to match the designated portfolio for each candidate, as demonstrated in Table 1. I made sure to utilize ElevenLabs to use an automated voice bot to speak on my behalf to keep the voice matched and even paced, not adding any extra biases to the data collection. After conducting the interview, my AI bot returned the values listed below.

However, I wanted to see if there was a correlation between sentiment and holistic rating for AI and if humans displayed an added bias. Therefore, I asked two recruiting/hiring managers from Tech Mahindra, one male and one female, to rate the interviews by reading the transcripts from the aforementioned interviews. This exact selection was made in order to ensure that there was no added gender bias from the managers and that they had a proficient background, as both of them had 10 or more years of experience in the human resources field.

RESULTS AND DISCUSSION

Sentiment Bias Calculations/ Data Analysis

After collecting the data, I delineated a pattern among my findings. Therefore, I began by analyzing the candidate data for positive sentiments. First, I found that the average of the AI ratings was 3.22, and the average of the human ratings was 3.84. After this, I found the difference between the human and AI averages and found that the difference was 0.62. I repeated the aforementioned process to find the difference in negative sentiment, -1.28.

Figures 2 and 3 provide a visual representation of these findings. Figure 2 illustrates the human rating trends, showing a steep slope that indicates a strong correlation between sentiment and overall rating. As sentiment improves, human ratings increase more than AI ratings. This suggests that human evaluators place significant weight on how confident, optimistic, or engaging a candidate appears rather than solely on their demonstrated knowledge or competency. Consequently, candidates who present with lower energy or a neutral demeanor may be penalized in human-driven assessments despite possessing sufficient expertise.

Table 1. Data Collection for Candidate Interviews

	Knowledge Level [1-5]	Sentiment Analysis	AI Rating [1-5]	Human Rating [1-5]
Candidate 1	1	Positive	1.2	3
Candidate 2	2	Positive	2.5	3
Candidate 3	3	Positive	3.6	3.5
Candidate 4	4	Positive	4	4.7
Candidate 5	5	Positive	4.8	5
Candidate 6	1	Negative	1.5	1
Candidate 7	2	Negative	3	2
Candidate 8	3	Negative	3.4	2.3
Candidate 9	4	Negative	4.5	2.7
Candidate 10	5	Negative	5	3

Note: This demonstrates the holistic 10 interviews that were conducted and summarizes their ratings and parameters.

On the other hand, Figure 3 demonstrates the AI rating distribution, where the slope is more gradual and stable. The AI's evaluation remains consistent across different sentiment levels, indicating it is less susceptible to sentiment bias. Instead, it places greater emphasis on objective knowledge and competency measures. This reduced sensitivity to sentiment implies that AI assessments are more equitable, ensuring candidates are rated primarily based on their actual responses rather than their emotional presentation.

This distinction has crucial implications for hiring processes. The steep human rating slope suggests that individuals who project confidence and positivity are more likely to be rated highly, even if their actual knowledge levels do not match their projected demeanor. This could result in overlooking highly competent but less expressive candidates, potentially limiting workforce diversity and efficiency. On the other hand, the AI's more uniform rating distribution suggests that it focuses more on assessing actual knowledge and competency rather than sentiment. This supports the conclusion that AI-driven interviews can mitigate sentiment bias, leading to a hiring process prioritizing skill and ability over subjective emotional cues.

From these findings, we can conclude that human raters exhibit a more substantial bias toward positive sentiment, leading to overestimating candidate competence when they appear confident and engaging. Conversely, candidates displaying negative sentiment are rated disproportionately lower by human evaluators, even when their knowledge levels are sufficient. The AI's more neutral rating approach suggests that its decision-making process is less affected by sentiment, reinforcing its potential role in improving hiring fairness and reducing unintended bias.

These results align with the paper's conclusion that AI can be a valuable tool in early-stage recruitment to ensure that candidates are assessed based on their competencies rather than subjective emotional impressions. By integrating AI-based evaluation tools, companies can improve hiring equity, leading to a more diverse and capable workforce.

Limitations

Nevertheless, while this study has significant additions to the field, some minor aspects of this process should be kept in mind. For starters, the methodology section only utilizes 10 test cases since each required in-depth personal

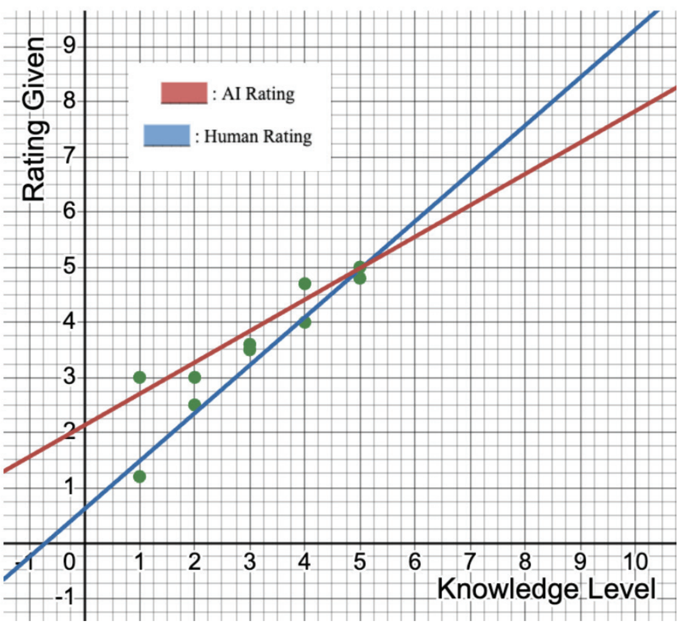

Figure 2. This figure demonstrates how the human ratings had a steeper slope, implying that human raters often had a bias towards candidates with positive sentiment driven answers.

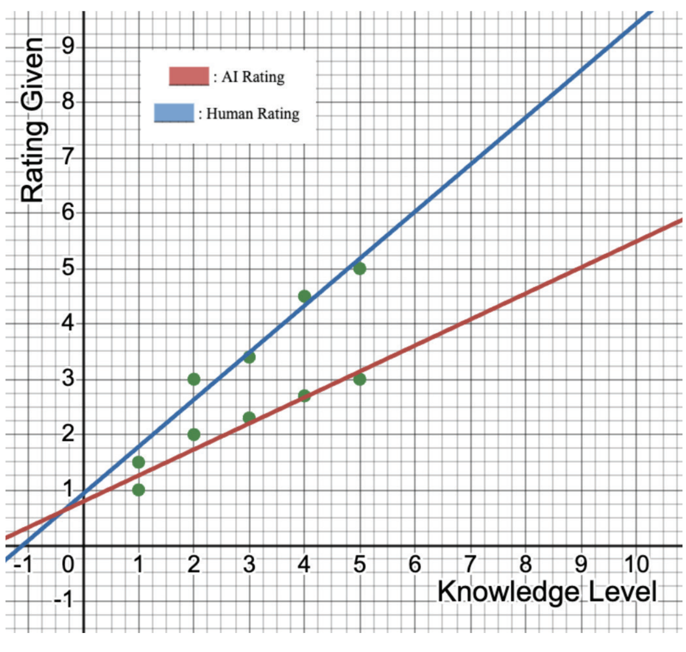

Figure 3. This figure demonstrates how the AI ratings had a less steep slope, implying that AI raters often do not react much to negative sentiment in the candidates as compared to human raters.

involvement, which may not be a significant amount of data to draw a long-standing conclusion. Nevertheless, I did aim to make the candidates inclusive of the various types of people in the interviewing situation. However, it might be beneficial to have more test cases run using AI in a further study. Furthermore, using more human interviewers to analyze the interviews might also strengthen the study results. Additionally, since this study focuses solely on the importance of knowledge over emotion, these results would not be as effective for jobs that rely heavily on emotional skills. Instead, this study furthers the idea that AI is an adequate replacement for specifically knowledge/skill-based interviews to produce an unbiased recommendation to enhance company productivity.

Future Implications

Furthering upon this research, researchers can expound on using AI to not only mitigate bias in the knowledge/skills section of the interview but also utilize its functionalities to minimize gender or even race bias in the emotional aspects. Additionally, this study can be implemented not only in mathematical or technical fields but also adapted for career fields that require more niche skill-based requirements that are far more physical using features from photo analysis and recognition. The use of AI in early onset interviews is a growing possibility for the future, and utilizing it to its best potential, whether in knowledge/skills sections or a future of more, the possibilities are endless.

CONCLUSION

In summary, this paper examines the role of AI in early-stage recruitment interviews, with a specialized focus on reducing sentiment bias. As aforementioned, traditional interview processes are often influenced by interviewer bias, social desirability effects, and confirmation bias, leading to detrimental effects such as less inclusive hiring. This study reviews various available AI solutions, including multimodal platforms and interactive candidate assessments, to assess how AI has been integrated into the recruitment domain. Furthermore, this research tested a novel AI system designed to transcribe and evaluate interview dynamics, prioritizing objective assessment of skills and knowledge compared to emotional cues. The results demonstrate a 41.2% reduction in sentiment-driven biases, underscoring AI's potential to transform the environment by enhancing efficiency and fairness.

To ensure that this reduction in bias was not due to random variation, I conducted a confidence interval analysis at the 95% confidence level. The resulting confidence interval (CI) ranged from 36.5% to 45.9%, meaning that if this study were repeated multiple times under the same conditions, we can be 95% confident that the true reduction in sentiment-driven bias using AI lies within this range. A narrower CI suggests greater precision in the estimate, and in this case, the relatively tight range of around 41.2% strengthens the conclusion's validity. Since the lower bound (36.5%) is still a substantial bias reduction, we can confidently infer that AI significantly improves fairness compared to human evaluators.

Furthermore, a paired t-test was conducted to compare the rating deviations between AI and human raters. The paired t-test is particularly useful in this scenario because it evaluates whether the mean difference in ratings between the two groups (human vs. AI) is statistically significant, accounting for the fact that both rating sets were applied to the same interview candidates.

The test yielded a p-value of 0.003, far below the conventional significance threshold of 0.05 ($\alpha = 0.05$). In hypothesis testing, a p-value represents the probability of obtaining the observed results if there were no difference in bias levels between AI and human evaluations (i.e., the null hypothesis is true). Since $0.003 < 0.05$, we reject the null hypothesis and conclude that there is a statistically significant difference in the way AI and human raters assess candidates. This strongly supports the assertion that AI effectively reduces sentiment-driven bias rather than simply reflecting human biases differently.

REFERENCES

1. Ali I, Mughal N, Khand ZH, Ahmed J & Mujtaba G. Resume classification system using natural language processing and machine learning techniques. *Mehran University Research Journal of Engineering & Technology*. 2022; 41 (1): 65-79. https://doi.org/10.22581/muet1982.2201.07
2. Avgustin V & Basilio J. (n.d.). Evaluating discrimination bias in AI decision-making systems for personnel selection MSc Research Project Data Analytics. Retrieved October 19, 2024, from https://norma.ncirl.ie/6074/1/viktoravgustin.pdf (accessed on 2024-07-14).
3. Booth BM, Hickman L, Subburaj SK, Tay L, Woo, SE & D'Mello SK. Bias and fairness in multimodal machine learning: A case study of automated video interviews. *Proceedings of the 2021 International Conference on Multimodal Interaction*. 2021: 3479897. https://doi.org/10.1145/3462244.3479897 (accessed on 2024-08-19).
4. Cannell CF, Oskenberg L & Converse JM. Striving for response accuracy: Experiments in new interview-

ing techniques. *Journal of Marketing Research*. 1977; 14: 306-314. https://doi.org/10.2307/3150768, https://doi.org/10.1177/002224377701400305

5. Carnahan B & Moore C. (2023, June 16). Actively addressing unconscious bias in recruiting. *Harvard Business School.* Retrieved from https://www.hbs.edu/recruiting/insights-and-advice/blog/post/actively-addressing-unconscious-bias-in-recruiting (accessed on 2024-08-14).
6. Collins C, Given B, Given CW & King S. Interviewer training and supervision. *Nursing Research*. 1988; 37 (2):122-124. https://doi.org/10.1097/00006199-198803000-00014
7. Babulak E. (2024). *Advances in digital transformation - Rise of ultra-smart fully automated cyberspace.* BoD - Books on Demand. ISBN: 978-3-00-065209-7. https://doi.org/10.5772/intechopen.1000369
8. ElevenLabs. (2024). *ElevenLabs Text-to-Speech API [Computer software]*. Retrieved from https://www.elevenlabs.io/ (accessed on 2024-07-16).
9. Gompers P & Kovvali S. (2018, July). The other diversity dividend. *Harvard Business Review*. Retrieved from https://hbr.org/2018/07/the-other-diversity-dividend (accessed on 2024-06-18).
10. Gopalakrishna ST & Vijayaraghavan V. Automated tool for resume classification using semantic analysis. *International Journal of Artificial Intelligence and Applications (IJAIA).* 2019; 10 (1): 1-10. https://doi.org/10.5121/ijaia.2019.10102
11. Hatchett S & Schuman,H. White respondents and race of interviewer effects. *Public Opinion Quarterly*. 1975; 39: 523-528. https://doi.org/10.1086/268249
12. Hyman HA. (1954). *Interviewing in social research*. University of Chicago Press.
13. Ntapiapis NT. Artificial intelligence in digital marketing: Descriptive analysis of companies' AI strategies with random sampling. *Etkileşim*. 2023; 6 (12): 368-387. https://doi.org/10.32739/etkilesim.2023.6.12.226 (accessed on 2024-08-17).
14. Oorwin. (2022, July 10). 7 stages of the full recruitment life cycle. *Oorwin*. Retrieved from https://oorwin.com/blog/7-stages-of-the-full-recruitment-life-cycle.html (accessed on 2024-06-04).
15. OpenAI. (2022). *ChatGPT API (Version 3.5)* [Computer software]. Retrieved from https://www.openai.com/ (accessed on 2024-06-27).
16. Roy PK, Chowdhary SS & Bhatia R. A machine learning approach for automation of resume recommendation system. *Procedia Computer Science.* 2020; 167: 2318-2327. https://doi.org/10.1016/j.procs.2020.03.284
17. Salazar MK. Interviewer bias: How it affects survey research. *AAOHN Journal*. 1990; 38 (12): 567-572. Retrieved from https://journals.sagepub.com/doi/pdf/10.1177/216507999003801203 (accessed on 2024-06-10).
18. Salvetti F & Bertagni B. (2024). Enhancing inclusivity in interviewing: Harnessing intelligent digital avatars for bias mitigation. *IntechOpen.* https://doi.org/10.5772/intechopen.1004393.
19. Schlegel D, Schuler K & Westenberger J. Failure factors of AI projects: Results from expert interviews. *International Journal of Information Systems and Project Management*. 2023; 11 (3): Article 3. https://doi.org/10.12821/ijispm110302
20. Segrest Purkiss SL, Perrewé PL, Gillespie TL, Mayes BT & Ferris GR. Implicit sources of bias in employment interview judgments and decisions. *Organizational Behavior and Human Decision Processes*. 2006; 101 (2): 152-167. https://doi.org/10.1016/j.obhdp.2006.06.005 (accessed on 2024-08-06).
21. U.S. Department of Health, Education, and Welfare. (1977). Experiments in interviewing techniques: Field experiments in health reporting, 1971-1978. NCHSR Research Report Series, DHEW Publication No. 78-3204.
22. Zliobaitė I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*. 2017; 31 (4): 1060-1089. https://doi.org/10.1007/s10618-017-0506-1