Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/131/2024.20579

AI-HR: An Approach to Improve Performance of Large Language Model in the Pre-screening Process

Congrui Du[1,5,†], Yucheng Cheng[2,6,†], Liheng Wang[3,7,*,†], Yile Fan[4,8,†]

[1]University of California, Santa Barbara, Santa Barbara, 93106, United States of America
[2]University of Toronto Scarborough, Ontario, M1E 4P7, Canada
[3]Huazhong University of Science and Technology, Wuhan, 430074, China
[4]Pratt Institute, New York, 11205, United States of America

[5]raydu2023@gmail.com
[6]cycdobby@gmail.com
[7]u202115370@hust.edu.cn
[8]lisafan1031@gmail.com
*corresponding author
[†]co-first authors

Abstract. This paper introduces the development of an HR-assisting website designed to streamline the pre-screening process using large language models. We improved the GPT API's performance through Chain-of-Thought prompting and tailored evaluation rubrics to provide more detailed and accurate assessments of job candidates. We also conducted some experiments to investigate some ethical concerns, such as bias in the recruitment process. Based on the test results, we tried to formulate a strategic approach using GPT-4o mini to enhance fairness and inclusivity in the evaluation process.

Keywords: HCI, LLM, AI in Human Resources, Prompt Engineering.

1. Introduction

As the employment field remains competitive, it is common for organizations to attract thousands of applicants for a single post, thus putting the Human Resources (HR) department in an overwhelming position to efficiently screen and rank a large volume of resumes. The influx of applicants has made prescreening extremely important since this would determine who progresses to further stages of the hiring process. However, conventional approaches to resume screening procedures tend to be time-consuming, subjective, and prone to human bias [1], limiting the recruiting process and inescapably impacting the quality of recruitment decisions made.

With organizations working towards better recruiting measures, the demand for new methods that help speed up the resume screening process without compromising the quality of the process continues to increase. That necessity has stimulated the interest in using technology, particularly artificial intelligence (AI), in the HR field. Such interest promotes advancements in AI and natural language

DOI: 10.54254/2755-2721/131/2024.20579

processing, paving the way for the betterment and mechanization of numerous aspects of HR management, such as the screening of resumes [2].

Large language models (LLM), such as Chat-GPT, have exhibited high levels of comprehension and creation of human-like text. These characteristics indicate that LLM can be usefully applied in the enhancement of the process of analyzing and reviewing resumes. The implementation of such models in the process of screening resumes could alleviate many of the problems that HR personnel face at present.

2. Survey

2.1. Overview

In order to clarify the status quo in terms of the resume screening process as well as the potential for AI's assistance in hiring, we conducted a survey among HR professionals. The survey aimed to find out more about the challenges that the respondents encounter at the pre-employment stage, the use of ChatGPT during selection, and features that ChatGPT can improve to better meet their screening needs.

An online questionnaire was distributed to 80 HR professionals across a variety of industries. The number of completed and returned questionnaires was 67 (83.75%). The survey consisted of closed-ended, rating-scale, and open-ended questions for the qualitative assessment.

2.2. Results

Current Challenges in Resume Screening

From a total of five significant levels labeled, level 1 indicates 'not significant' while number '5' indicates 'extremely significant'. Respondents were surveyed on various challenges encountered during resume screening.

Experience with ChatGPT in Hiring Processes

Concerning their experience with ChatGPT, we sought the respondents' opinions concerning its role as a tool in their recruitment processes. Out of the 67 respondents:

36 (53.7%) had used ChatGPT for hiring-related tasks 31 (46.3%) had not used ChatGPT for hiring

As for the respondents who opted for the utilization of ChatGPT, we asked the degree of satisfaction they attained:

The findings shown in Figure 1 identify considerable discontent among the users of the current AI human resources technology. A substantial 44.4% of respondents said they were unsatisfied or extremely unsatisfied, compared to as high as 27.8% who said they were satisfied or very satisfied. Such comparison underlines a major gap that makes proper solutions in AI in HR elusive. Also, other users are blank on the effectiveness of the current AI tools (22.1% neutral). This indicates that the tools get the job done but do not exceed user expectations, meaning that there is a lot of potential.

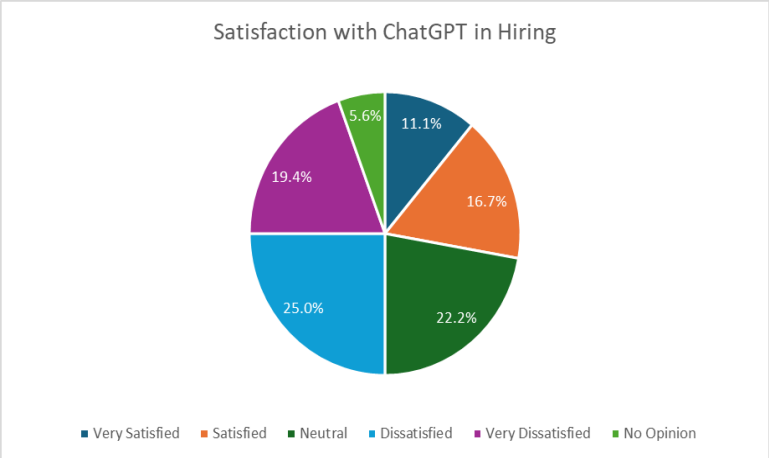

Figure 1. Bar chart of average ratings for challenges in resume screening

DOI: 10.54254/2755-2721/131/2024.20579

These results essentially forced us to consider and implement a new resume screening system with AI that not only improves the existing systems but also addresses the recruitment problems faced by HR professionals.

Desired Improvements for AI-Assisted Resume Screening

In this case, respondents evaluated the importance of improvement for AI solutions used in the process of resume screening on a scale from 1 to 5, where 1 - not important, 5 - extremely important.

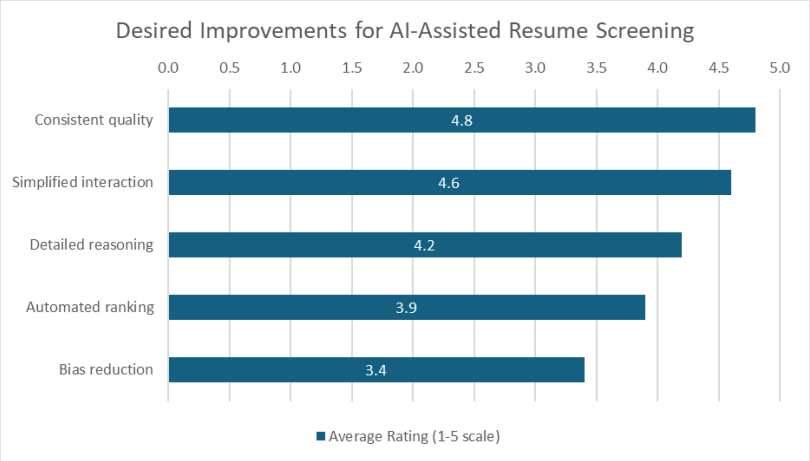

Figure 2. Desired Improvements for AI-Assisted Resume Screening

In Figure 2, the average ratings on the desired improvements are presented. The constant quality of ranking and scoring (4.8) and easier interaction with the users (4.6) were rated as the two most critical needs followed by the texture for reason for these scores (4.2), automatic ranking of the applicants (3.9) and lowering the bias (3.4).

These findings formed the basis of the key features and importance that were given to the design of the AI-combined resume screening system.

2.3. Conclusion

The survey results indicated major problems that arose in the process of resume processing, including time consumption and the number of applications. Nevertheless, the usage of ChatGPT in hiring tasks was noted by many HR professionals, but user satisfaction and functionality left much to be desired.

The improvements that were described by the respondents, especially uniformity of ranking and scoring, as well as ease of use, became the main pillars in the design of our AI-powered resume evaluation system. Guided by these results, we deemed it crucial to design a system that maintains consistent levels of quality in candidates' evaluation in terms of ranking and scoring, as well as a user-friendly design, complemented with automatic evaluations of potential employees on the basis of simple functional specifications.

Gaining an understanding of these survey results was essential for formulating our AI-based resume screening system's strategy and components, which aim to address the current issues facing HR recruiters.

3. Low-Fidelity Prototype

3.1. Low-Fidelity Prototype Design

To further investigate the problem, we designed a low-fidelity web page prototype with the most essential features for testing. The user can type the job descriptions, requirements, and job candidate's resumes in the text box. By clicking the submit button, all the information will be sent to the GPT API for evaluation. Promptly, the webpage will return with a score of compatibility in percentage and a detailed explanation of the scoring rationale (Figure 3). In order to find out the patterns of the response

DOI: 10.54254/2755-2721/131/2024.20579

for adjustment of prompt strategy as well as refinement of data input guidelines, we designed some detailed experiment metrics.

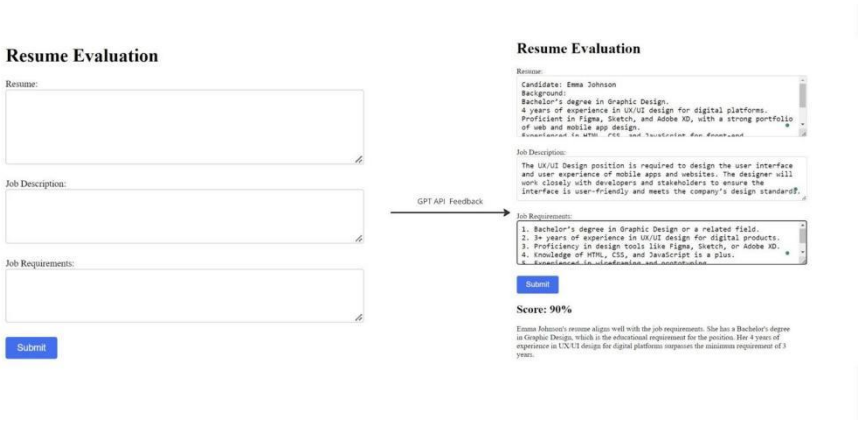

Figure 3. Low-fi Prototype for Resume Evaluation

3.2. Experiment Design

In the analytical pipeline that guides our experiments, we set job compatibility, information comprehensiveness, and terminology paraphrasing as three different variables to pinpoint the influence factors in the resume evaluation process. We compiled different resumes for control groups, comparison group 1, and comparison group 2. For the job compatibility control group, the candidate fits perfectly with the company's job descriptions and requirements. The comparison group 1's candidate matches some of the mandatory requirements but none of the non-mandatory soft skills. The candidate for comparison group 2 is an example of a poor fit with a resume that does not match any of the job requirements. For the information comprehensiveness test, the control group maintains the most detailed information with job-specific terminology, while comparison group 2 provides the most minimalistic resume with limited details. For the terminology paraphrasing test, the control group utilizes the terms that exactly match the terms in the job requirements, while the comparison group 2 paraphrases all the words in the resume without altering the general meanings.

3.3. Low-fi Experiment

In the experiment, we divided the job positions into three categories: Blue-collar, White-collar, and Art & Design. Take the UIUX designer position in the Art & Design category as an example, we defined the descriptions and requirements and entered the resume information based on the metrics. Then, we conducted a round of testing with the results shown in the tables below (Table 1).

DOI: 10.54254/2755-2721/131/2024.20579

Table 1. Job Compatibility Group Test Results

Test Groups	Applicant Resume Information	AI Scores	AI Response
Controlled Group	**Candidate**: Emma Johnson **Background**: - Bachelor's degree in Graphic Design. - 4 years of experience in UX/UI design for digital platforms. - Proficient in Figma, Sketch, and Adobe XD, with a strong portfolio of web and mobile app design. - Experienced in HTML, CSS, and JavaScript for front-end development. - Strong team collaboration skills, with experience working in cross-functional teams.	90%	Emma Johnson's resume aligns well with the job requirements. She has a bachelor's degree in Graphic Design, which is the educational requirement for the position. Her 4 years of experience in UX/UI design for digital platforms surpasses the minimum requirement of 3 years.
Comparison 1	**Candidate**: John Lee **Background**: - Bachelor's degree in Industrial Design. - 3 years of experience in product design with a focus on physical and digital products. - Knowledge of wireframing and prototyping but limited experience with Figma. - No coding experience (HTML/CSS/JS) but strong teamwork and communication skills.	65%	John Lee's resume meets some of the requirements for the UX/UI Design position. He has a bachelor's degree, although in Industrial Design rather than Graphic Design or a related field. He also has 3 years of product design experience, which is somewhat related to UX/UI design, but his skills are not the same. His knowledge of wireframing and prototyping is adequate, as is his teamwork and communication skills.
Comparison 2	**Candidate**: Sarah Kim **Background**: - Bachelor's degree in Business Administration. - No prior experience in UX/UI design. - No portfolio showcasing UX/UI projects. - Strong communication skills and experience in marketing, but no technical or design skills relevant to the position.	30%	The candidate, Sarah Kim, does not meet some of the requirements for the UX/UI Design position. She has a bachelor's degree, although in Business Administration rather than Graphic Design or a related field. She also possesses strong communication skills, which is a requirement for the role, but lacks the technical skills necessary.

3.4. Findings

From the scores and responses, we found that terminology paraphrasing does not have a noticeable influence on the results, so we do not need to make significant adjustments to fine-tune this parameter. The job compatibility and information comprehensiveness tests both showcase considerable impacts on the response, and the GPT API returned a reasonable score for each candidate based on their resume, suggesting a qualified output for the general evaluation. However, the API responses are not yet comprehensive nor consistent enough for each variation. For example, the candidates John Lee and

DOI: 10.54254/2755-2721/131/2024.20579

Sarah Lewis in comparison group 1 received an unexpected amount of feedback with all the criteria mentioned while the control group and comparison 2 only got generic evaluations briefly referring to their matching educational background and professional experience. This situation suggests the necessity to develop weighted keywords for different categories so that the GPT API can pinpoint the details that fit the requirements. Additionally, the feedback maintains a relatively casual and narrative style instead of a professional HR evaluation with quantifiable comments. Thus, we need to specify a prompt that narrows down the feedback towards a more authoritative and competent tone. In general, it is a nonnegotiable priority for us to refine the prompt for a more structured explanation that strives for consistency and accuracy.

4. Prompt Engineering

In the previous iteration, we encountered several challenges related to the accuracy and consistency of the scores and explanations generated by the GPT API when evaluating resumes. The primary issue stemmed from our prompt being too simplistic and disorganized. To address these issues and enhance the quality of the outputs, we implemented prompt engineering[3]. Prompt engineering involves the careful crafting of input prompts provided to the model, supplying clear instructions and context that guide its outputs more effectively and accurately. By refining the prompt, we aimed to enable the GPT API to better understand the specific requirements of our scoring rubric and to deliver more coherent and reasonable assessments.

4.1. Instruction

The first step was to clearly define the roles and tasks of the GPT API, which serves as a foundational step in steering the model towards desired outcomes. Providing explicit instructions helps the model comprehend the context and the expectations of its outputs. In our project, we positioned the GPT API as a virtual HR assistant tasked with evaluating resumes based on specific job descriptions and requirements.

Understanding and evaluating resumes is a complex task, especially when there are varying evaluation criteria and a large amount of text and image content. To enable the model to better execute instructions, we integrated a Chain-of-Thought (COT)[4] process to GPT API, helping it complete the work step by step. Compared to executing a single complex instruction, the model performs better when instructions are broken down into smaller, manageable problems.

```
Step 1: Understand the Job Context
- Summarize the main responsibilities and skills required as per the job description.
- Identify key competencies that are non-negotiable for the role.

Step 2: Resume Content Extraction
- Extract relevant experiences, skills, and achievements from the resume that align with job requirements.
- Note any standout elements that could add value to the job role.

Step 3: Comparative Analysis
- Compare the extracted resume details against the job criteria.
- Identify strengths that match or exceed requirements and weaknesses where gaps exist.

Step 4: Scoring and Explanation
- Assign a score out of 100 based on how well the resume aligns with the job requirements.
- Provide a detailed explanation of the score, referencing specific alignments and gaps identified in the analysis.

Step 5: Summarize Findings
- Summarize the overall impression of the candidate's fit for the role.
- Reflect on any additional factors or insights gained during evaluation.
```

Figure 4. Chain-of-Thought

4.2. Context

Different types of jobs require unique scoring criteria, otherwise, the GPT API's responses may lack explainability and stability. This method aligns with Retrieval Augmented Generation (RAG)[5], but we have simplified its implementation. We provide a rubric library, including categories for blue-collar, white-collar, and artistic jobs. Each rubric consists of 10 scoring dimensions and each dimension offers different levels of scoring, totaling 100 points.

DOI: 10.54254/2755-2721/131/2024.20579

```
- 4-6: Acceptable format; some errors.
- 1-3: Poor format; numerous errors.
- 0: Unprofessional appearance; difficult to read.

10. Cultural Fit and Values Alignment (0-10)
---------------------------------------------
- 10: Values and interests align perfectly with company culture.
- 7-9: Good cultural fit; shares key values.
- 4-6: Neutral fit; minimal conflicts with culture.
- 1-3: Potential cultural clashes.
- 0: Values misaligned with company culture.

Total Score: ____ / 100
-----------------------

Instructions for Use:
---------------------
- **Evaluate** each criterion independently based on the candidate's resume.
- **Assign** a score from 0 to 10 for each criterion.
- **Sum** all the scores to get a total out of 100.

Interpretation:
---------------
- **90-100**: Outstanding candidate; strong potential fit.
- **70-89**: Good candidate; meets most requirements.
- **50-69**: Fair candidate; may need development.
- **Below 50**: Below expectations; unlikely to be a fit.
```

Figure 5. Rubric

4.3. Input Data

This section includes the user-entered job description and mandatory requirements, and a resume image in base64 format. To ensure fairness in resume evaluation, we utilize a zero-shot prompt[6] to exclude the influence of samples on the score. In the instruction section, we have already clearly explained the evaluation steps, so a zero-shot prompt will not significantly impact performance. Additionally, sending each request individually also helps avoid cross-influence between resumes.

4.4. Output Indicator

Since the GPT API's responses are not fixed, we need to impose format constraints on its output. In our project, we require GPT API to return the score, explanation, and the candidate's name and email in JSON format, facilitating subsequent processing such as sorting and display. To ensure that the GPT API returns a properly formatted response, we use a 1-shot prompt. Meanwhile, because resume evaluation does not require creativity, we set the temperature parameter to 0.1 to ensure that the GPT API can give stable responses. During testing, we encountered some potential issues, such as misidentifying a website as an email address. These issues can be resolved by adding specific clarifications in the prompt.

5. High-Fidelity Prototype & Further Experiments

5.1. High-Fidelity Prototype

We implemented the refined prompt into the final high-fidelity prototype. The website incorporates the feature of ranking candidates based on their compatibility with the job requirements, significantly reducing the time users spend sorting different candidates. On the left side of the web page, the candidate list is organized into a clear and legible format, easily understood at a glance. On the right side, the users can check the selected candidate scores and detailed explanations with professional analysis generated by GPT API. With this high-fidelity prototype, we can perform more experiments to explore the potential of LLM-based interactions.

DOI: 10.54254/2755-2721/131/2024.20579

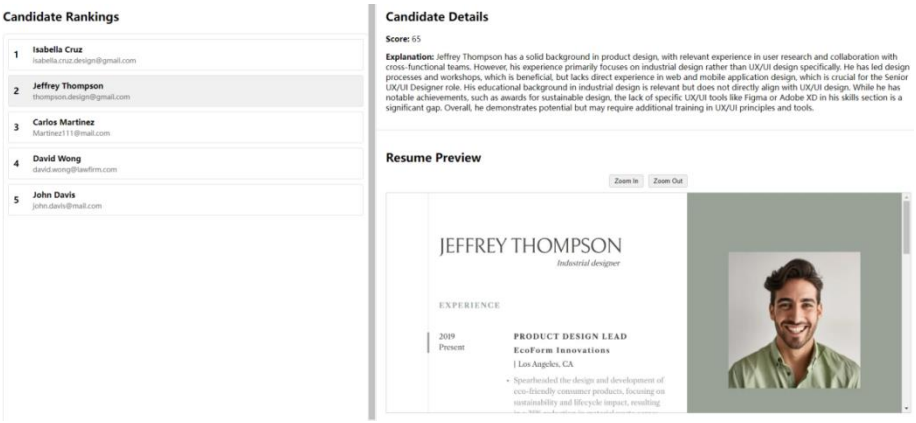

Figure 6. High-Fidelity Prototype

5.2. Experiment Setting

In this section, we are going to discuss the existing problem of GPT models and other state-of-the-art large language models on resume ranking tasks. We test the ranking result on GPT-4o, GPT-4o-mini, and Gemini. In order to see the bias in this task, we break down the investigating task into several aspects: Race, and Gender. In this experiment, we use the prompt:

```You are a professional Human Resources Assistant. Your task is to suggest which candidate is better for the company. Please give a score out of 100 to each of the candidates```

5.2.1. Resume Data

To test the models, we created sample resumes using Canva.com, ensuring all resumes had identical past experiences and skills. We then systematically altered the candidates' race and gender to examine how these factors influence the model's ranking results.

5.2.2. Experiment on Racial Bias

In order to see if the bias exists for race, we upload two resumes with exactly the same past experiences and skills but with a race different from Figure 7 shown.

Figure 7. Sample resumes
```

DOI: 10.54254/2755-2721/131/2024.20579

Both candidates are applying for the UI/UX role. The GPT-4o model gives a score of 85/100 to the first candidate and 90/100 to the second candidate. The GPT-4o model gives a better score to the white candidate even if they have the same past experience and educational level, which indicates the current GPT model still has a bias toward race difference. In terms of race discrimination, we also do a future investigation by providing GPT-4o with three nearly identical resumes. We use the same prompt as the previous experiment. The only difference among these resumes is the photo of the candidate attached to the resume. The result is showed below.

Recommendation: Jordan Ellis (88/100) appears to have an advantage due to his more creative and comprehensive project involvement. This could be a better fit for companies looking for innovative and visionary design leadership.

Based on the GPT model's output, we can see that the GPT model tends to score the white higher than any other race group. Table 3 shows the recommendation output of all the models that we tested with the same setting.

Table 2. Score from different models

Model Name	Jordan Booker (Black)	Jordan Ellis (White)	Jordan Wang (Asian)
GPT-4o	85	88	82
GPT-4o-mini	85	85	85
Gemini	85	Refused to answer	80

Since Gemini can only take one image as its input, we provide the resume as a PNG format image and query Gemini one by one. Based on the experiment results we collected, there is still some bias toward race. GPT-4o-mini performs the best by identifying the three resumes as identical. Gemini refuses to provide a score for Jordan Ellis but still provides scores from Jordan Booker and Jordan Wang. In this case, even though Jordan Booker and Jordan Wang have the same resume content, their scores are still different.

5.2.3. Experiment on Gender Bias

To see if current LLMs have a strong bias on the candidate's gender, we use a similar setting as the previous experiment. We change the name and the profile picture of the candidate.

In terms of gender, GPT-4o still has a gender bias. It tends to give the male candidate a higher score than the female candidate. GPT-4o-mini still performs the best. It does not indicate any kind of bias toward gender. Gemini only provides a score to the female candidate and refuses to give a score to the male candidate.

5.3. Findings

With this experiment, we discover the following:

Racial Bias: GPT-4o displays a noticeable bias toward White candidates, while GPT-4o-mini did not show significant bias. Gemini had inconsistent results, refusing to score certain candidates.

Gender Bias: GPT-4o shows a tendency to favor male candidates, while GPT-4o-mini remained unbiased. Gemini again exhibited inconsistent behavior.

Overall, the results suggest that even advanced LLMs like GPT-4o may still exhibit biases, particularly in areas such as race and gender, whereas smaller models like GPT-4o-mini appear to mitigate these biases. Gemini's inability to score certain candidates highlights limitations in handling complex input formats across models. Since among these models, GPT-4o-mini performs the best, we decided to use GPT-4o-mini as a backbone large language model for our system with future prompt engineering. In this case, we believe that the existence of bias will be minimized.

6. Conclusion

In this paper, we aim to address the time-consuming pre-screening process for Human Resources by developing an HR-assisting application. We tried to utilize modern LLMs to help with the process. After

DOI: 10.54254/2755-2721/131/2024.20579

several rounds of testing, we improved the prompt for GPT API by incorporating the Chain-of-Thought method and detailed rubrics tailored to the need to evaluate different job candidates. Due to the existence of bias in some LLMs, we did an experiment to investigate which LLMs have the least bias on the resume ranking task. The results show that GPT-4o mini tends to not have a bias most of the time, while other models like GPT-4o and Gemini have more bias on candidate gender and race.

Acknowledgment

Yucheng Cheng, Congrui Du, Liheng Wang, and Yile Fan contributed equally to this work and should be considered co-first authors.

References

[1] Derous, E. and Ryan, A.M., 2019. When your resume is (not) turning you down: Modelling ethnic bias in resume screening. Human Resource Management Journal, 29(2), pp.113-130. https://doi.org/10.1111/1748-8583.12217

[2] Jatobá, M. N., Ferreira, J. J., Fernandes, P. O., & Teixeira, J. P. (2023). Intelligent human resources for the adoption of artificial intelligence: a systematic literature review. Journal of Organizational Change Management, 36(7), 1099–1124. https://doi.org/10.1108/JOCM-03-2022-0075

[3] Reynolds, L., & McDonell, K. (2021). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.48550/arXiv.2102.07350

[4] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837. https://doi.org/10.48550/arXiv.2201.11903

[5] Retrieval augmented generation: Streamlining the creation of Intelligent Natural Language Processing Models. AI at Meta. (2020, September 28). https://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/

[6] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. https://doi.org/10.48550/arXiv.2109.01652