Human and LLM-Based Resume Matching: An Observational Study

Swanand Vaishampayan
Virginia Tech, USA
swanandsv@vt.edu

Hunter Leary
Virginia Tech, USA
hunterl22@vt.edu

Yoseph Berhanu Alebachew
Virginia Tech, USA
yoseph@vt.edu

Louis Hickman
Virginia Tech, USA
louishickman@vt.edu

Brent Stevenor
NREMT Lab, USA
brentstevenor@gmail.com

Fletcher Wimbush
DiscoveredATS, USA
fletcher@thehiretalent.com

Weston Beck
DiscoveredATS, USA
weston@thehiretalent.com

Chris Brown
Virginia Tech, USA
dcbrown@vt.edu

Abstract

Resume matching assesses the extent to which candidates qualify for jobs based on the content of resumes. This process increasingly uses natural language processing (NLP) techniques to automate parsing and rating tasks—saving time and effort. Large language models (LLMs) are increasingly used for this purpose—thus, we explore their capabilities for resume matching in an observational study. We compare zero-shot GPT-4 and human ratings for 736 resumes submitted to job openings from diverse fields using real-world evaluation criteria. We also study the effects of prompt engineering techniques on GPT-4 ratings and compare differences in GPT-4 and human ratings across racial and gender groups. Our results show: LLM scores correlate minorly with humans, suggesting they are not interchangeable; prompt engineering such as CoT improves the quality of LLM ratings; and LLM scores do not show larger group differences (i.e., bias) than humans. Our findings provide implications for LLM-based resume rating to promote more fair NLP-based resume matching in a multicultural world.

1 Introduction

Resume matching is the first step in multi-stage hiring pipelines, where recruiters rate resumes based on the extent to which applicants' resume content matches job requirements to find ideal candidates (Li et al., 2020). This rating involves assessing criteria such as work experience, skills, education, certifications and extracurricular activities (Tsai et al., 2011). However, resume matching is challenging, involving rating hundreds of resumes per job opening (Torres, 2). Thus, manual resume matching processes are laborious and time-consuming for recruiters. Modern hiring pipelines employ natural language processing (NLP)-powered support to automate resume matching, saving time and effort for recruiters in resume matching processes (Mujtaba and Mahapatra, 2019).

Recent advancements of NLP-based Large Language Models (LLMs), such as OpenAI's GPT-4, Meta's Llama-3 and Google Gemini, have seen rapid adoption (Chen et al., 2024a) and revolutionized numerous domains such as software engineering (White et al., 2023), linguistics (Diandaru et al., 2024) and e-commerce (Roumeliotis et al., 2024). Recent studies (Gan et al., 2024; Gaebler et al., 2024) evaluate LLMs for resume parsing and rating, reporting positive results. This motivates the use of LLMs for resume matching, with perceived benefits of easier, faster and efficient resume rating.

However, LLM evaluations can incorporate challenges, such as discrimination against candidates (Armstrong et al., 2024). To safeguard against biased LLM-based resume ratings and maximizing perceived benefits, it is imperative to comprehend the differences between rationales used by humans and LLMs. Prior work fails to incorporate the intricacies of authentic resume matching processes in practice, relying on experimental approaches with tight control over factors that can influence LLM-generated scores (i.e., resumes and criteria). It is important to use real-world resume rating constructs that contribute to hiring decisions (Stepanova et al., 2021; Tsai et al., 2011).

Additionally, to the best of our knowledge, no previous studies discern "how" resume ratings are computed by LLMs. We aim to fill this gap by exploring the capabilities of LLMs in resume matching. We compare ratings generated by human raters and zero-shot GPT-4 on resumes of real-world applicants and the job postings to which they applied. The research questions tackled in the study are:

> **RQ1:** What are the differences of zero-shot GPT-4 and human resume ratings based on resume reviewing constructs? *Motivation:* Studying the GPT-4 rationale in generating resume ratings will aid in finding and localizing the differences between LLMs and humans.

The severity of the differences can be an important factor in determining the applicability of LLMs for future rating systems.

RQ2: What are the group differences in scores generated by GPT-4 and humans across race/ethnicity and gender demographics? *Motivation:* Studying group differences will provide better clarity on the extent of biases in human and GPT-4 resume ratings for a specific race/ethnicity and gender. Findings from this research question will aid in designing more equitable & fair future resume rating.

We analyzed a total of 736 resumes from real-world job applicants across four job titles—"Project Manager", "Accountant", "Sales" and "Engineer"—spanning diverse professional fields. Motivated by real-world resume matching processes, we compare human and GPT-4 ratings across four constructs — work experience, skills, education, and certifications. Our results indicate differences in the scoring rationale followed by GPT-4 and human raters in resume evaluations. GPT-4 ratings are more lenient for skills and are more stringent for certifications when compared to human ratings. In terms of group differences, human ratings exhibited more difference than GPT-4 across race/ethnicity and gender. In the wake of our findings, we discuss potential solutions, such as prompt engineering, to minimize the observed shortcomings of zero-shot GPT-4 ratings.

Our work makes the following contributions: **A:** We localize differences across resume rating constructs through the comparisons of human and GPT-4 ratings. Additionally, we study the effects of prompt engineering techniques, such as Chain of Thoughts (CoT), on resume ratings. **B:** We contribute to literature on ethics in AI by investigating GPT and human rating differences across racial and gender subgroups for resume matching tasks.

2 Related Work

2.1 Resume Rating using LLMs

LLMs are trained on a large corpus of data (Achiam et al., 2023), making them capable of evaluating resumes across domains as they posses the knowledge necessary to read and understand resumes with "near-human" accuracy (Kaygin, 2024). Recent work (Veldanda et al., 2023) demonstrates the capability of LLMs—GPT-3.5, Claude and Bard—in accurately mapping resumes to various job categories, underlining the prowess of LLMs in resume matching tasks. Gan et al. (Gan et al., 2024) provide a fully automated framework for resume matching and decision making for job offers using GPT-3.5 turbo, demonstrating their approach is 11 times faster than manual methods and records an F-1 score of 87.73% for resume sentence classification. Ghosh et al. (Ghosh and Sadaphal, 2023) introduce JobRecoGPT — a recommender system that generates resume and job description match scores using GPT-4. Prior work also found positive results using LLMs for resume classification (Rithani et al., 2024) and recovering categorical information from labor market data, such as college majors and occupations (Chen et al., 2024b).

The above mentioned works lay a strong case for use of LLMs for resume ratings. However, they also exhibit limitations. The majority of the studies only consider the IT domain (Gan et al., 2024; Ghosh and Sadaphal, 2023; Rithani et al., 2024). Additionally there are limitations of datasets used for the studies. For instance, the resumes used for some studies are anonymized, thus potentially suffering from information loss (Veldanda et al., 2023), or are synthetic (Ghosh and Sadaphal, 2023), which might change the GPT ratings and fail to reflect the real-world. Gan et al. (Gan et al., 2024) evaluated the performance of GPT-3.5 by considering the performance of GPT-4 as the ground truth—thus, the reported accuracy of their framework might not reflect its true accuracy. Our work focuses on a fine-grained analysis approach for resume matching, comparing human and zero-shot GPT-4 turbo on constructs such as work experience match, skills match, educational match and certifications match. We explore the capabilities of GPT-4 ratings by focusing on the rating rationale employed by GPT-4. Additionally, we use a dataset of 734 real resumes, without alternations, submitted to 25 real job openings and their associated descriptions across different domains such as Construction Management, Software Engineering, UI/UX Engineering and Real Estate.

2.2 Ethics and LLMs in Hiring

While LLMs can improve efficiency and productivity in hiring (Gan et al., 2024), prior work suggests LLMs can introduce potential bias against marginalized populations. For instance, Wan et al. (Wan et al., 2023) found gender bias in the language of cover letters generated by GPT-3.5. Prior work shows names associated with Black women receive

the most bias in scenario-based LLM tasks (Haim et al., 2024). Prior work evaluated hiring decisions of GPT-3.5, and found ethnic bias against individuals from Hispanic backgrounds (An et al., 2024b).

Research also shows LLMs incorporate bias in resume matching. For instance, Armstrong et al. (Armstrong et al., 2024) altered candidate names in resumes, demonstrating White-sounding names received higher ratings than those reflecting other racial and ethnic groups on otherwise identical resumes. Similarly, An et al. (An et al., 2024a) rated 361,000 resumes for entry-level jobs, showing GPT-3.5 favoured female candidates over African-American male candidates. However, Veldanda et al. (Veldanda et al., 2023) found no racial or gender bias exhibited from GPT-3.5, Claude, Llama and Bard in matching resumes to corresponding domains. Our work explores the extent of bias in human and GPT resume ratings by observing scoring differences based on race/ethnicity and gender, incorporating more fine-grained nuances in resume matching processes by scoring work experience, skills, education and certifications separately.

3 Dataset

Our industry partner provided a dataset containing 35,138 resumes submitted across 79 different job titles. The dataset contained job descriptions, job titles, applicant demographic information and resumes. Due to rate limit restrictions of GPT-4 and time and resource constraints for human raters, we selected four job titles—Project Manager, Accountant, Sales and Engineer—providing us with ample English-based resumes to inspect variance across race/ethnicity and gender (see Appendix A.1). For instance, Project Manager and Engineer were male-dominated whereas Accountant and Sales were female-dominated. We selected resumes submitted to 20 unique job openings — five from each job title. Based on these criteria, the selected job descriptions from the Project Manager and Accountant job titles were from the Construction Management domain, Sales job descriptions were from the Real Estate domain, while the Engineering job descriptions were more diverse, consisting domains of Construction Management, Software Engineering and UI/UX engineering. Our final sample had ($n = 196$) resumes each for Project Manager and Accountant, ($n = 179$) resumes for Sales and ($n = 165$) resumes for Engineer, for a total of 736 resumes.[1]

4 Study Design

4.1 Rating Scale

For rating resumes we formulated a rating scale using an informal judgment study (Storey et al., 2020), interviewing three industry experts to gain insights on resume matching in real-world hiring contexts. The experts had prior experience in recruiting candidates for professional roles in Software Engineering, Construction Management, and Sales. Participant details are presented in Appendix B.1. We obtained approval from our institutional review board (IRB) for this study.

Experts completed a brief evaluation activity during the interview session, tasked with rating 10 resumes–two from each of five job descriptions used in our study, providing a reasoning for their rating. This provided insight to our raters to imbibe how experts rate resumes for their respective domains. Based on these preliminary results, we formulated a five point rating scale from 1 to 5, where 1 stands for 'vastly not meeting minimum requirements' and 5 stands for 'vastly exceeding minimum requirements', across four resume rating constructs—work experience, skills, certifications, and education. The interview transcripts and resume matching activity Excel sheets are included in our supplementary materials.

4.2 Human Ratings

For human ratings, we recruited eight raters from diverse ethnic and gender backgrounds at the first author's institution. The raters also came from different fields of study, spanning undergraduate and graduate students in Computer Science and Psychology. A limitation of this work is that the human raters were inexperienced with resume matching in professional settings. However, we mitigated this by using the expert judgment study to inform our rating process, selecting a diverse raters from differing racial and gender backgrounds (see Appendix B.2), and collecting four ratings per resume.

Every resume was rated on the defined 5-point scale along with a reasoning by four raters on how the resume matched minimum requirements for work experience, skill requirements, educational qualifications and certifications specified by the

[1]Due to IRB restrictions, we are not able to share the candidate resumes in our supplemental material.

job description. To mitigate inconsistent frame-of-reference across raters, we held inter-rater agreement meetings to finalize the metrics for every job description (Chaturvedi and Shweta, 2015). The final human rating was the average score given by all four raters. For the subsequent sections of this paper, human rating corresponds to average score given by all four raters.

To cross-verify that the ratings from experts and trained student-raters are comparable, the three industry experts consulted during the judgment study rated a random sample ($n = 20$) of resumes from each job title of Project Manager, Engineer and Sales ($total = 60$). We compared the ratings of the trained human raters and experts using Pearson's correlation and Fleiss' kappa. We found both ratings had high correlation and high Fleiss' kappa, demonstrating the rating process used in our study was comparable to experts raters. Table 1 provides detailed information on the correlation and Fleiss' kappa values across expert and study ratings.

4.3 LLM Ratings

LLM ratings were generated using the GPT-4 turbo model from OpenAI, a state-of-the-art large language model with a broader knowledge base and advanced reasoning capabilities.[2] We computed the GPT-4 ratings with zero-shot prompting, meaning we did not provide any additional information regarding how to rate the resumes or specific examples of ratings to GPT-4. We prompted GPT-4 through the API to provide ratings for resumes similar to the human rating process. We tasked GPT-4 to use the 5-point scale from 1-5 given the job description summary and resume content. The ratings were also prompted to be based on the four constructs: work experience, skills, educational qualifications, and certifications. For the GPT-4 generated ratings, we were particularly interested in knowing the rationale GPT-4 uses to rate the candidate resumes. Thus, we prompted GPT-4 to also describe the reasons for the selected ratings. To ensure the consistency in GPT-4 ratings we undertook following steps: i) The study prompt was tested on 50 sample resumes outside of our dataset. This experiment was repeated 3 times and the consistency in LLM ratings for study prompt was verified before stating the study. ii) The temperature value used during our study was set to 0.5, which further ensured the consistency in LLM ratings.

[2] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

Previous work shows prompt engineering techniques can enhance the performance of LLMs (Kojima et al., 2022). To this end, we additionally employed three different prompt structures to analyze their effect on LLM resume rating performance: a) Task based prompting, we defined tasks along with scoring criteria and an expected output structure; b) Task based Chain of Thought (CoT) prompting, we divided the process into four simple sub-steps; and c) Task based CoT with examples, we extended CoT by providing an example of a human resume rating. More details about LLM ratings are provided in Appendix B.3.

4.4 Data Analysis

Human Rater Agreement We calculated the inter-rater agreement between human raters using Fleiss' kappa. We found the average Fleiss' kappa as $\kappa = 0.7859$ for "substantial" agreement (Landis and Koch, 1977). We observed the kappa values for all job titles across constructs fell within the substantial range (0.61 to 0.80), indicating a balanced rating between human raters for each of the rated resumes.

RQ1: Human-GPT Construct Differences We measured Human-GPT-4 rating level of agreement by calculating Pearson's correlation and Fleiss' Kappa between the ratings across all four constructs. We also visualize these differences using the score distributions of GPT-4 and human ratings. For finding the differences in GPT-4 and human ratings, we used an open-coding approach based on inductive thematic analysis. Based on our results from Human-GPT agreement, we constructed a smaller dataset consisting of resumes with large differences in ratings between humans and GPT-4—that is, instances where the rating differed by two or more—for manual analysis. These observations were reviewed to analyze the ratings and associated reasons for each rating given by GPT-4 and human raters to uncover the differences in rating rationale. The code books used in thematic analysis are provided in our supplementary material.

RQ2: Human-GPT Group Differences We used standardized Cohen's d measure to capture human-GPT group differences and magnitude of the observed (Becker, 2000) effects resolving shortcomings of sample dependent statistical testing. Cohen's *d* is defined as the difference of sample means divided by pooled standard deviation of both the samples. For our study, we had two independent

Table 1: Correlation and Fleiss' Kappa of Expert and Human Ratings

Category	Project Manager		Sales		Engineer		Mean Corr.	Fleiss' Total
	Corr.	Fleiss	Corr.	Fleiss	Corr.	Fleiss		
Work exp	0.9693	0.8658	0.8448	0.8649	0.8723	0.7898	0.8954	0.8401
Skills	0.8771	0.8599	0.9286	0.8951	0.8010	0.7975	0.8689	0.8508
Education	0.8665	0.8020	0.8695	0.7808	0.9378	0.8453	0.8912	0.8093
Certification	0.9596	0.8384	0.8026	8460	1.0000	1.0000	0.9506	0.8943

groups: GPT-4 ratings and human ratings. The key effect in our research is the inter-group rating difference between GPT-4 and human ratings for racial and gender subgroups. Additionally, we computed the intra-group racial and gender group difference in both human and GPT-4 ratings (i.e., differences between human-human ratings and GPT-GPT ratings) for rating constructs across racial and gender subgroups. We use Cohen's d and confidence intervals to aid us in both objectives. Confidence interval (CI) computation for Cohen's d is computed using a t-value, Cohen's d and sample sizes of both the samples. For our study, t-value was 95% CI. For effect size, Cohen's d values of 0–0.19 indicate "very small", 0.20–0.49 is "small", 0.50–0.79 is "medium", and above 0.8 is "large" effective size respectively.

5 Results

5.1 RQ1: Difference in GPT–Human ratings

Human-GPT Agreement We analyzed how GPT-4 and human ratings differ across work experience, skills, education and certifications by calculating the Pearson's correlation and Fleiss' Kappa between the human and GPT-4 ratings. Table 2 presents the details of these results for each job title. We found a "moderate" correlation for work experience (0.5675) but observed lower correlations for ratings based on skills, education and certifications between humans and GPT-4 (Gogtay and Thatte, 2017). For Fleiss' Kappa no agreement was found across any constructs. This result signifies that there are substantial differences between human and GPT-4 ratings in these categories. Figure 1 shows the distribution graphs depicting the differences between the GPT-4 and human scores for each construct. In the graph, a positive difference signifies GPT-4 ratings were greater than human ratings, a negative score indicates human ratings were greater than GPT-4, while zero signifies agreement. From the distribution graphs, it can be observed that for skills, graph is left skewed whereas for certifications it is right skewed indicating GPT-4 ratings tend to be more lenient for skills

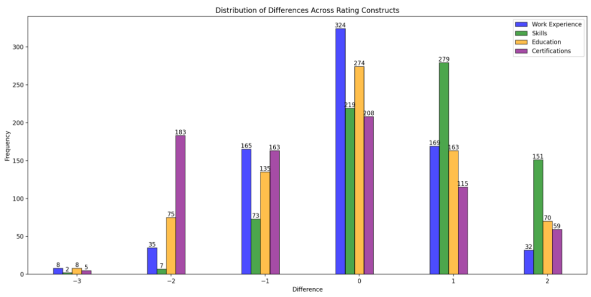

Figure 1: Distribution of Differences between GPT-4 and Human Ratings for Work Experience, Skills, Education and Certification

while more stringent for certifications compared to human ratings.

Reasoning Difference To further investigate resume rating differences, we filtered resumes with large differences (≥ 2) between GPT-4 and human ratings. We manually inspected each of these to observe differences in rationale given by human raters and GPT-4 for their respective rating. This manual analysis provided us with cues about differences in rating rationale employed by both GPT-4 and human. We found five themes as major differences through this analysis:

Scoring Criteria: Human raters and GPT-4 relied on different scoring criteria. For instance, human raters evaluated resumes with no certifications as satisfying job descriptions with no minimum requirements for certifications, giving them a score of 3. However, GPT-4 rated similar candidates with certification score of 1. We found total of ($n = 415$) instances of scoring criteria differences for our study, with majority in education and certifications dimensions.

Human-GPT validity disagreement: There were instances ($n = 109$) where GPT-4 and human raters perceived the validity of resume content differently. For instance, a candidate for a "Project Manager" job description who had OSHA-10 certificate was rated 4 by a human on certifications, considering that OSHA-10, a construction safety certificate is relevant for a project management role in a construction company. Whereas, GPT-4 rated

Table 2: Correlation and Fleiss' Kappa of Human and GPT-4 Ratings

Category	Project Manager		Accountant		Sales		Engineer		Mean Corr.	Fleiss' Total
	Corr.	Fleiss	Corr.	Fleiss	Corr.	Fleiss	Corr.	Fleiss		
Work exp	0.5080	0.196	0.6801	0.249	0.4796	0.198	0.5901	0.215	0.5675	0.228
Skills	0.3403	0.007	0.4688	0.062	0.3684	-0.009	0.4622	-0.017	0.3823	0.019
Education	0.5864	0.28	0.2765	0.087	0.4466	0.088	0.5459	0.041	0.4091	0.153
Certific	0.4531	0.063	0.5105	-0.032	0.0750	-0.065	0.5706	0.049	0.3438	0.035

Table 3: Differences in GPT-4 and Human Ratings

Category	Work Experience	Skills	Education	Certifications
Scoring Criteria	21	98	126	170
Human-GPT Validity Disagreement	48	39	22	NA
GPT Hallucinations	3	23	7	2
GPT Implications	2	1	5	23
Wrong Heading in Resumes	NA	NA	NA	5

the candidate's certification as 2, incorrectly deeming OSHA-10 certification as unrelated to the job.

GPT Hallucinations: Hallucinations, an inherent shortcoming of LLMs was observed in our experiment, where GPT-4 made mistakes in interpreting the minimum requirements from job descriptions or matching the resume content with job description requirements. For instance, for a job description which had the minimum requirement of a Bachelor's degree, GPT-4 considered a minimum requirement of a high school degree and rated the candidates accordingly. We recorded ($n = 35$) instances of hallucinations during our analysis.

GPT Implications: In a few instances ($n = 31$), GPT-4 inferred certain information about candidates which was not explicitly mentioned in the resume. For instance, for a job description which required professional MS Word and Excel proficiency as minimum skill requirements, GPT-4 assumed a candidate satisfied the expected requirement because they had five years of work experience in a project manager role and rated the candidate's skills as 4—despite no explicit mention of these skills in the resume itself. Alternatively, humans rated this candidate lower.

Wrong Heading in Resume: This finding was limited to certifications rating. We recorded ($n = 5$) instances where a candidate's resume had certifications presented under a different heading—for instance, certifications listed under an Education title, and therefore GPT-4 incorrectly ignored these certifications. Table 3 describes information about occurrences of each of these themes for across work experience, skills, education and certifications.

Effect of Prompt Engineering on LLM ratings We examined the effect of prompt engineering techniques such as CoT on the zero-shot GPT-4 ratings and analyzed the results through an experiment. We sampled 198 resumes having large reasoning differences (≥ 2) in GPT-4 and Human ratings across our resume rating constructs and verified the final sample had resumes with all five shortcomings: a) Scoring Criteria; b) Human-GPT validity disagreement; c) GPT Hallucinations; d) GPT Implications; and e) Wrong Heading in Resumes.

For analyzing the performance of prompt techniques, we plotted the distribution graphs of differences between GPT ratings using different prompts and human ratings across work experience, skills, education and certification. Difference of zero in graphs signifies the exact match in ratings. Positive difference signifies GPT ratings to be greater, and negative differences signifies human ratings to be greater. Figure 2 show the distributions of all the three GPT prompt ratings along with the zero-shot GPT ratings for Certifications. Table 4 shows the number of samples with exact match for zero-shot baseline and three prompt techniques. Additionally, we calculated the Pearson's correlation and Fleiss' kappa to further inspect human and GPT-4 agreement for each prompt engineering technique across the resume rating constructs. From distributional graphs, we observed performance improvement using the prompt-based techniques compared to zero-shot prompting across all four resume rating constructs, except for task based CoT for work experience. A similar trend was seen for correlation and Fleiss' kappa, where prompt-based techniques improved GPT-4 agreement with human raters in every resume rating construct. More details regarding correlation and Fleiss' Kappa are in Appendix C. Prompt engineering techniques improved rating performance by minimizing the rating differences of (a) Scoring Criteria differences

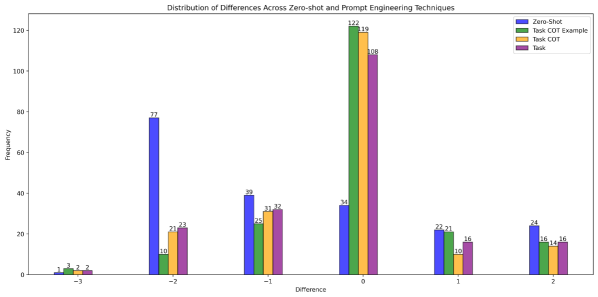

Figure 2: Distribution of Differences between GPT-4 and Human Ratings for Certification

and (d) GPT implications stemming from the LLM scoring rationale. Certification contained the most instances of (a) and (d) and hence saw the most improvement out of all the rating constructs followed by education and skills. Work experience did not contain instances of (a) and (d) and therefore saw the least improvement in performance. Out of the three enhanced prompting techniques, Task Based CoT with examples gave most improvement overall—achieving "moderate agreement" for certification.

There were differences in GPT-4 and human in terms of rationale used while rating the resumes. Prompt engineering techniques improved the performance of GPT-4 ratings, but even with the improvement in performance, the maximum accuracy GPT ratings could achieve in our study was 61.61% for certifications (122 out of 198). When analyzing GPT-4 ratings for task based CoT with example with large differences between GPT and human ratings, we found the same underlying issues found in Section 5.1 such as GPT implications underlining that inherent reasoning differences of GPT-4 cannot be fully fixed using prompt engineering.

Summary of RQ1 Findings:

- GPT and Human resume ratings differ in terms of work experience, skills, education and certifications.
- Scoring criteria, Human-GPT validity disagreements, GPT hallucinations, GPT implications and Wrong heading in resumes are the themes found responsible for rating differences in Human and GPT-4 ratings.
- GPT-4 rating performance improved using all three prompt engineering techniques with Task based Chain of Thoughts with example provided the most improvement across certifications, in terms of alignment with human ratings.

Figure 3: Summary of Key Findings in RQ1

5.2 RQ2: Racial and Gender Group Differences in Ratings

To answer RQ2, we compared GPT-4 ratings and human ratings across racial and gender groups and calculated group differences using Cohen's *d* and its CI. The subgroups are Asian-White, African-American-White, Hispanic/Latino-White and Multiracial-White for race and Male-Female for gender. We computed the respective Cohen's *d* and its CI for each subgroup across the four constructs of work experience, skills, education and certification. Tables 5 and 6 show these results. Inferences regarding intra-group (i.e., Human-Human or GPT-GPT) and inter-group (Human-GPT) ratings can be made by interpreting the Cohen's d outcomes.

5.3 Intra-Group Human Rating Differences

Asian-White: For work experience ratings of resumes by humans for Asian and White candidates, we found Cohen's *d* was 0.7651, very close to 0.8 which indicates large magnitude effect size. The CI [0.5475, 0.9827] does not contain 0, which signifies the difference between the two means is statistically significant. In simpler terms, the human ratings for Asian subgroup and White subgroup for work experience differ significantly, with an almost large effect size. Similarly for certifications, we found Cohen's d as 0.5635 and the CI as [0.3485, 0.7785] which shows a statistically significant difference between Asian and White candidates in human ratings with a medium effect size.

African American-White: For work experience ratings for resumes of individuals identifying as African American compared to those identifying as White, we found Cohen's *d* for human ratings of 0.5415 with the CI [0.3028, 0.7802]. This signifies a significant difference between the human ratings of African American and White candidates based on their reported work experience in resumes with medium effect size.

5.4 Intra-Group GPT Rating Differences

African-American-White: For resumes for African-American and White candidates, we observed a Cohen's *d* of 0.4892 for GPT-4 evaluations of work experience, very close to 0.5 which shows medium magnitude effect size. The CI [0.2511, 0.7273] does not contain 0, which signifies that the difference between the two means is statistically significant. Thus, GPT ratings

Table 4: Samples with Perfect Match Between GPT and Human Ratings

Category	Zero-Shot GPT-4	Task Based	Task Based CoT	Task Based CoT w/ Example	Total samples
Work Exp.	65	70	60	68	198
Skills	42	47	45	57	198
Education	62	102	110	88	198
Certific.	34	108	119	122	198

Table 5: Comparison of GPT-4 and Human Ratings by Demographic Groups Across Work Experience and Skills

Demographic	Work Experience		Skills	
	GPT-4	Human	GPT-4	Human
Asian-White	0.3284 [0.1155, 0.5413]	0.7651 [0.5475, 0.9827]	0.1406 [-0.0714, 0.3526]	0.1264 [-0.0856, 0.3384]
African-American-White	0.4892 [0.2511, 0.7273]	0.5415 [0.3028, 0.7802]	0.3719 [0.1347, 0.6091]	0.3258 [0.0889, 0.5627]
Hispanic/Latino-White	0.3184 [0.1107, 0.5261]	0.2433 [0.0361, 0.4505]	0.0625 [-0.1442, 0.2692]	0.0809 [-0.1258, 0.2876]
Multiracial-White	0.4489 [0.1879, 0.7099]	0.4301 [0.1693, 0.6909]	0.4147 [0.154, 0.6754]	0.1002 [-0.159, 0.3594]
Male-Female	0.3627 [0.2164, 0.509]	0.1435 [-0.0018, 0.2888]	0.1332 [-0.0121, 0.2785]	0.0022 [-0.1429, 0.1473]

* Values in brackets are the respective Confidence Intervals

Table 6: Comparison of GPT-4 and Human Ratings by Demographic Groups Across Education and Certification

Demographic	Education		Certification	
	GPT-4	Human	GPT-4	Human
Asian-White	0.3282 [0.1153, 0.5411]	0.3472 [0.1342, 0.5602]	0.1279 [-0.0841, 0.3399]	0.5635 [0.3485, 0.7785]
African-American-White	0.0794 [-0.1565, 0.3153]	0.3734 [0.1362, 0.6106]	0.2249 [-0.0114, 0.4612]	0.1544 [-0.0817, 0.3905]
Hispanic/Latino-White	0.2799 [0.0725, 0.4873]	0.0818 [-0.1249, 0.2885]	0.1235 [-0.0833, 0.3303]	0.2786 [0.0712, 0.486]
Multiracial-White	0.0648 [-0.1944, 0.324]	0.2873 [0.0274, 0.5472]	0.0428 [-0.2163, 0.3019]	0.0375 [-0.2216, 0.2966]
Male-Female	0.0313 [-0.1138, 0.1764]	0.1382 [-0.0071, 0.2835]	0.3848 [0.2384, 0.5312]	0.0284 [-0.1167, 0.1735]

* Values in brackets are the respective Confidence Intervals

for the African-American subgroup and White subgroup for work experience differ significantly, and the magnitude of effect size difference is close to medium.

5.5 Inter-Group Human-GPT Rating Differences

Asian-White: For resumes of Asian and White candidates, we found the Cohen's *d* for GPT (0.3284) and human (0.7651) ratings for work experience, indicating more difference in human ratings than GPT ratings ([difference of 0.4367). The CIs of the two rating sources—[0.1155, 0.5413] and [0.5475, 0.9827] respectively—do not overlap, which signifies that the human ratings exhibit larger differences than do the GPT ratings. Similarly, we found GPT and human Cohen's *d* as 0.1279 and 0.5635 (difference of Cohen's d as 0.4356) and non-overlapping CIs [-0.0841, 0.3399] and [0.3485, 0.7785] for certifications, which shows human ratings exhibit significantly larger group differences than the GPT ratings of certifications for candidates identifying as Asian and White.

Male-Female: For the gender subgroups, we found GPT and human Cohen's d as 0.3848 and 0.0284 with non-overlapping CIs of [0.2384, 0.5312] and [-0.1167, 0.1735] for certifications. This shows that GPT-4 ratings exhibit greater gender differences than human ratings.

In most cases (11 out of 20), we found that group differences for GPT ratings were larger than for human ratings across the resume rating constructs for racial and gender subgroups. However, only one difference is statistically significant. On the other hand, for nine out of 20 cases we found human ratings exhibited larger group differences than GPT ratings for racial and gender subgroups—two of which are statistically significant.

Summary of RQ2 Findings:

- Intra-group human ratings revealed significant group differences between Asian-White and African American-White groups across work experience.
- Intra-group GPT ratings found significant differences between Asian-White candidates across work experience.
- Inter-group Human-GPT ratings revealed more statically significant differences in Human ratings GPT ratings.

Figure 4: Summary of Key Findings in RQ2

6 Implication of Results

Applicability of LLMs Resume Rating Prior work shows LLMs can enhance resume matching processes (Gan et al., 2024). Our findings also demonstrate the capabilities of GPT-4 for resume matching in practice. We found improvement in the performance of GPT-4 ratings using prompt engineering techniques such as Chain of Thoughts (CoT). But even with the improvement in performance, the maximum accuracy GPT ratings could achieve

with prompt engineering was 61.61% for certifications (122 out of 198). When analyzing GPT-4 ratings for task based CoT with example with large differences between GPT and human ratings, we found the same underlying issues found in Section 5.1 such as GPT implications underlining that inherent reasoning differences of GPT-4 cannot be fully fixed using methods of prompt engineering.

LLM-human rating differences for racial and gender subgroups Overall, human ratings had more differences for racial subgroups when compared to GPT ratings. While this finding does not contradict prior research (Armstrong et al., 2024; Wan et al., 2023) showing that GPT ratings had bias against racial minority groups such as Asian, Hispanic/Latino and gender minority group such as female, it does help to contextualize GPT ratings and provides guidance for their effective use. Our study was observational, whereas previous works were experimental studies which manipulated the names on resumes and provided less direction about how to rate the resumes. Another reason could be the difference between the resume matching tasks. We gave clear direction to GPT-4 to rate the resumes across work experience, skills, educational qualifications and certifications, which should minimize the influence of applicant names, whereas previous studies used GPT-3.5 for broader, less concrete evaluations such as willingness to hire (Armstrong et al., 2024), writing cover letters (Wan et al., 2023) respectively. Further, previous studies used GPT-3.5 turbo model, while our study used GPT-4 turbo model. Thus, we believe more research is required to comprehend if the GPT bias is task dependent and if it varies according to the study approach.

After comparing GPT-Human ratings we found human ratings can differ more than GPT ratings. For example, work experience and certification ratings different significantly across Asian-White subgroups. We tried to minimize the human biases by picking raters from diverse backgrounds. However, we believe this finding underlines the inherent biases of humans.

Proposed Solution: Human-Centered AI (HCAI) approaches to improve LLM-based resume matching. Our findings show LLMs can provide a good starting point in resume ratings and can be used as a guide for recruiters in resume matching. Rather than a fully automated LLM-based rating system, we propose use of an LLM-guided human-in-the-loop approach (Monarch, 2021), falling under Human-Centered AI (HCAI) perspectives (Shneiderman, 2022). For instance, human-in-the-loop approaches can have the final decision made by a human recruiter. Alternatively, given the errors and cognitive biases of humans, human-in-the-loop approaches could have humans make independent evaluations of resumes, then provide opportunities for further scrutiny in cases where humans and LLMs substantially disagree.

Prior work suggests human-AI teaming can improve decision making (Munyaka et al., 2023). Based on our results, we posit leveraging LLM parsing capabilities and rating guidance can save time and effort in resume matching by presenting information regarding the minimum job requirements along with matching candidates' work experience, skills, education and certifications. Further, collaborations between LLMs and humans can help mitigate mistakes made during resume matching by either rater. Finally, the use of a human-in-the-loop approach will improve candidates' trust in hiring pipelines (Vaishampayan et al., 2023; Harris, 2024), making resume matching more transparent and reliable from candidates' perspective.

7 Conclusion

We designed an observational study to analyze GPT-4 and differences between human resume ratings across constructs of work experience, skills, educational qualifications and certifications. Additionally, we explore the impact of prompt engineering, such as CoT on GPT-4 rating performance. Lastly, we investigate intra-group and inter-group scoring differences based on race/ethnicity and gender. Our results show differentials in ratings, such as GPT-4 being more severe for educational qualifications but humans being more stringent for work experience. Prompt engineering techniques partially improved the overall performance of GPT-4 ratings. Implications of our findings propose incorporation HCAI advocated human-in-the-loop approaches in LLM-driven resume ratings, to motivate more fair and equitable hiring pipelines in a multicultured world.

8 Acknowledgments

We express our gratitude to all the undergraduate human-raters: Keya Chava, Meghan Leight and Lily Jo for their valuable contribution. Additionally, we thank the three experts for their valuable insight and feedback regarding resume rating process.

Limitations

External Validity Our evaluation is based on one LLM, GPT-4. While this was the most advanced version of the most popular LLM at the time of this work (Humble, 2023), future work is needed to explore the capabilities of additional generative AI models trained on varying data and offered by different organizations (i.e., Google Gemini[3] or Anthropic Claude[4]). Our study also uses resumes from a limited set of job domains. We attempted to mitigate this by selecting a diverse sample of job titles requiring varying background and expertise—Project Manager, Accountant, Sales, and Engineer—however, further exploration is needed to explore the performance of LLM-generated rankings for additional job descriptions and applications.

Internal Validity We designed the rating criteria used for human ratings based on previous work (Stepanova et al., 2021; Tsai et al., 2011) and feedback received during preliminary interviews with recruiting experts. However, the criterion might not generalize to all job domains and job titles. We also did not manipulate any variables in our study, limiting our ability to draw causal conclusions about the relationship between demographic membership and LLM resume matching outcomes. We also used $n = 736$ in actual study and $n = 50$ pilot runs of GPT to observe LLM-generated ratings. Finally, while we used a diverse sample of human raters from varying backgrounds to form our baseline rankings, all of the raters were students from the authors' institution. A more diverse sample of raters with relevant experience could increase the confidence of our claims.

Construct Validity We observe difference in human and LLM ratings against two constructs. For calculating racial differences, we compared resumes from individuals identifying from minority subgroups (i.e., Asian, Hispanic/Latino, African-American) to White candidates. However, this excludes candidates from other minority race/ethnicity backgrounds, such as Native American. Additionally, we use White to compare against in our analysis as this race generally represents the majority of the US labor workforce[5]—yet, this might not be true for every job domain. For instance, in our dataset the "Engineer" job title had more Asian candidates than White. In addition, we used male and female to observe gender differences in human and GPT ratings. This incorporates a sex-based view of gender and disregards job candidates who identify as transgender, non-binary, or agender (Scheuerman et al., 2020)—however, this is a limitation of the resume data provided in our dataset. Finally, this work only considers race/ethnicity and gender subgroup differences, however we acknowledge the potential for bias across other diversity axes (i.e., disability status or age). We do not directly analyze bias, but explore group differences between race/ethnicity and gender groups—which could indicate potential bias (Thissen et al., 1986). Further research is needed to explore the extent of group differences between human and LLM resume ratings for additional races/ethnicities, genders, and underrepresented groups in the workforce.

Ethical Considerations

A. Limitations

A.1 Did you describe the limitations of your work? Yes, see previous section.

A2. Did you discuss any potential risks of your work?

B. Scientific Artifacts

The scientific artifacts used for this study include GPT-4 and a dataset of resumes provided by our industry partner.

B1. Did you cite the creators of artifacts you used? We leveraged GPT-4 turbo[6] from OpenAI to explore LLM-generated resume ratings. Our resume dataset was provided by The Hire Talent,[7] a talent acquisition tech company.

B2. Did you discuss the license or terms for use and / or distribution of any arfifacts We adhere to the OpenAI terms of use.[8] There is no license or terms of use for our dataset, as it was provided directly to the research team by our industry partner (see dataset sheet).

B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? Prior work has used GPT for resume matching tasks (i.e., (Gan et al.,

[3] https://gemini.google.com/

[4] https://claude.ai/

[5] https://www.bls.gov/opub/reports/race-and-ethnicity/2022/home.htm

[6] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

[7] https://www.preemploymentassessments.com/

[8] https://openai.com/policies/row-terms-of-use

2024)). Our resume dataset usage was consistent with it's intended use (see dataset sheet).

B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it? Our dataset includes information that uniquely identifies individuals. For this reason, we do not release out dataset.

B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.? Our dataset is described in Section 3, demographic details of our study sample are provided in Appendix A.1, and additional details are in our supplemental dataset sheet. All of the resumes used for this work were in English.

B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? We do not report these relevant statistics, as we did not train models on our dataset.

C. Computational Experiments

C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used? We leveraged the GPT-4 API to design a custom script for our analysis run on a local machine, thus we do not report the parameters and computational budget.

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values? A description of our GPT-4 configuration is provided in Appendix B.3.

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run? We report statistics using Pearson's correlation and Fleiss' Kappa for construct differences and Cohen's D for group differences, with confidence intervals, described in Section 4.4 and presented in Section 5. LLM ratings were generated using a single run.

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, Spacy, ROUGE, etc.), did you report the implementation, model, and parameter settings used? N/A

D. Human Annotators

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.? We used experts to formulate our rating scale and human raters to compare against GPT ratings. An overview is provided in Sections 4.1 and 4.2. Additional details are included in our supplemental materials.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)? Expert participants ($n = 3$) and human raters ($n = 8$) were recruited through personal contacts. Preliminary study participants were compensated with a $20 Amazon gift card. Undergraduate human raters ($n = 4$) were compensated for their efforts.

D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? Consent for individual resumes was collected from our industry partner (see dataset sheet). We collected informed consent from expert participants and human raters.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board? Our preliminary study obtained IRB approval (see Section 4.1). As our evaluation leveraged existing data, did not collect data from human subjects, and used ratings from members of our research team, we did not need IRB approval.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data? Demographic data for our expert participants (see Appendix A.1) and human raters (see Appendix B.2) is provided.

E. AI Assistant Usage

E1. Did you include information about your use of AI assistants? We use GPT-4 to observe LLM-generated resume ratings. We outline our usage in Section 4.3.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024a. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? *arXiv preprint arXiv:2406.10486*.

Jiafu An, Difang Huang, Chen Lin, and Mingzhu Tai. 2024b. Measuring gender and racial biases in large language models. *arXiv preprint arXiv:2403.15281*.

Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The silicone ceiling: Auditing gpt's race and gender biases in hiring. *arXiv preprint arXiv:2405.04412*.

Lee A Becker. 2000. Effect size (es).

SRBH Chaturvedi and RC Shweta. 2015. Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology*, 41(3):20–27.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. 2024a. Are more llm calls all you need? towards scaling laws of compound inference systems. *arXiv preprint arXiv:2403.02419*.

Yi Chen, Hanming Fang, Yi Zhao, and Zibo Zhao. 2024b. Recovering overlooked information in categorical variables with llms: An application to labor market mismatch. Technical report, National Bureau of Economic Research.

Ryandito Diandaru, Lucky Susanto, Zilu Tang, Ayu Purwarianti, and Derry Wijaya. 2024. What linguistic features and languages are important in llm translation? *arXiv preprint arXiv:2402.13917*.

Johann D Gaebler, Sharad Goel, Aziz Huq, and Prasanna Tambe. 2024. Auditing the use of language models to guide hiring decisions. *arXiv preprint arXiv:2404.03086*.

Chengguang Gan, Qinghao Zhang, and Tatsunori Mori. 2024. Application of llm agents in recruitment: A novel framework for resume screening. *arXiv preprint arXiv:2401.08315*.

Preetam Ghosh and Vaishali Sadaphal. 2023. Jobrecogpt–explainable job recommendations using llms. *arXiv preprint arXiv:2309.11805*.

Nithya J Gogtay and Urmila M Thatte. 2017. Principles of correlation analysis. *Journal of the Association of Physicians of India*, 65(3):78–81.

Amit Haim, Alejandro Salinas, and Julian Nyarko. 2024. What's in a name? auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875*.

Christopher G Harris. 2024. Combining human-in-the-loop systems and ai fairness toolkits to reduce age bias in ai job hiring algorithms. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 60–66. IEEE.

Charles Humble. 2023. Top 5 large language models and how to use them effectively. *The New Stack*. https://thenewstack.io/top-5-large-language-models-and-how-to-use-them-effectively/.

Esranur Kaygin. 2024. Comparative analysis of ml (machine learning) and llm (large language models) in resume parsing. *Hirize AI*. https://hirize.hr/blogs/ml-llm-comparison-in-resume-parsing.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

JR Landis and GG Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159—174.

Changmao Li, Elaine Fisher, Rebecca Thomas, Steve Pittard, Vicki Hertzberg, and Jinho D Choi. 2020. Competence-level prediction and resume & job description matching using context-aware transformer models. *arXiv preprint arXiv:2011.02998*.

Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Dena F Mujtaba and Nihar R Mahapatra. 2019. Ethical considerations in ai-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–7. IEEE.

Imani Munyaka, Zahra Ashktorab, Casey Dugan, J. Johnson, and Qian Pan. 2023. Decision making strategies and team efficacy in human-ai teams. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1).

M Rithani, R Venkatakrishnan, et al. 2024. Empirical evaluation of large language models in resume classification. In *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–4. IEEE.

Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. 2024. Llms in e-commerce: a comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056.

Morgan Klaus Scheuerman, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham. 2020. Hci guidelines for gender equity and inclusivity.

Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.

Anna Stepanova, Alexis Weaver, Joanna Lahey, Gerianne Alexander, and Tracy Hammond. 2021. Hiring cs graduates: What we learned from employers. *ACM Transactions on Computing Education (TOCE)*, 22(1):1–20.

Margaret-Anne Storey, Neil A Ernst, Courtney Williams, and Eirini Kalliamvakou. 2020. The who, what, how of software engineering research: a socio-technical framework. *Empirical Software Engineering*, 25:4097–4129.

David Thissen, Lynne Steinberg, and Meg Gerrard. 1986. Beyond group-mean differences: The concept of item bias. *Psychological bulletin*, 99(1):118.

M Torres. 2. million candidates are desperate to work at google. why. *Why? Ladders*.

Wei-Chi Tsai, Nai-Wen Chi, Tun-Chun Huang, and Ai-Ju Hsu. 2011. The effects of applicant résumé contents on recruiters' hiring recommendations: The mediating roles of recruiter fit perceptions. *Applied Psychology*, 60(2):231–254.

Swanand Vaishampayan, Sahar Farzanehpour, and Chris Brown. 2023. Procedural justice and fairness in automated resume parsers for tech hiring: Insights from candidate perspectives. In *2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 103–108. IEEE.

Akshaj Kumar Veldanda, Fabian Grob, Shailja Thakur, Hammond Pearce, Benjamin Tan, Ramesh Karri, and Siddharth Garg. 2023. Are emily and greg still more employable than lakisha and jamal? investigating algorithmic hiring bias in the era of chatgpt. *arXiv preprint arXiv:2310.05135*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Jules White, Sam Hays, Quchen Fu, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. *arXiv preprint arXiv:2303.07839*.

A Dataset

A.1 Demographic Information

The racial and gender breakdown of applicants according to each job category used in our evaluation is provided in Table 7.

B Study Design

B.1 Expert Interview Questions and Demographics

For fixing rating metrics used in our study, we interviewed 3 industry professionals with previous experience of recruiting candidates for professional roles in Software Engineering, Construction Management, and Sales. The specific subset of asked questions during the interviews and more information about backgrounds of industry experts are provided in Figure 5 and Table 8.

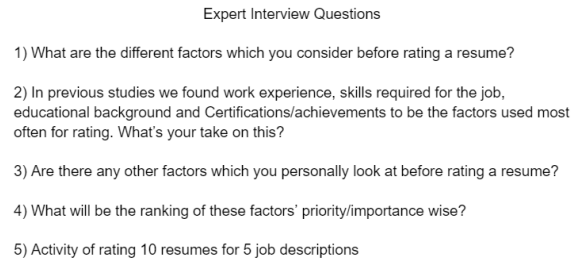

Figure 5: Expert Interview Questions

B.2 Human Rater Demographics

B.2.1 Raters

The human ratings lasted for 3 months and took place from February to April 2024. We made an effort to recruit human raters with our sample consisting of raters from Asian, Black/African-American, White and Hispanic/Latino racial backgrounds as well as four female and four male raters. Every resume was rated by four raters with differing racial and gender backgrounds. This imparted diverse representation to all racial and gender backgrounds for our human ratings.

B.3 LLM Rating Prompt

We used GPT-4 turbo model for rating resumes across constructs of work experience, skills, education and certifications on a scale of 1-5. To form our prompt, we started with an initial simple prompt and updated the prompt incrementally to satisfy all of the requirements. We used the default maximum input and output sizes of GPT-4 turbo model (i.e., 128k tokens for input and 4096 tokens for output). None of our prompts were above the input token limit. The temperature value was set to 0.5, as it gave the most consistent results. The role of GPT-4 was set to "professional HR that rates resumes" while making the API call. Before finalizing our prompt, we tested the prompt on 50 samples outside of our dataset and checked the results for hallucinations. We finalized the prompt after verifying that the GPT-4 generated ratings based on work experience, skills, educational qualifications and certifications adhered to expected formatting with no hallucinations. The final zero-shot prompt had the following structure:

Listing 1: Prompt for Feedback Generation

Table 7: Demographic and Gender Distribution by Job Category

Category	Project Manager	Accountant	Sales	Engineer
Race/Ethnicity				
Asian	16	29	6	64
African-American	11	47	17	14
Hispanic/Latino	24	28	41	30
Two or more races/Multiracial	10	24	25	10
White	135	65	90	46
Gender				
Male	166	57	55	115
Female	30	139	124	50

Table 8: Pilot Study Expert Information

Expert	R1	R2	R3
Current Job Title	Project Estimator	SDE-III	Sales Executive
Experience	4 years	7 years	4 years
Domain	Construction Management	Software Engineering (Full Stack Development)	Pharmacy
Gender	Male	Male	Female
Race/Ethnicity	White	Asian	Asian
Age	26	29	28

Instruction: Generate a score on the scale 1–5 for each work experience match, skills match, educational background match and certifications/extracurricular match based on the job description summary and resume. Additionally provide the reasons for the generated rating. Be strict in rating. The format of the output should be like following: \n\n Rating: \n Work Experience Match: \n Skills Match: \n Educational Background Match\n Certifications\Extracurricular Match: \n\n Reasons for rating:\n\n { job_summary}\n\n {resume_content}

An example zero-shot prompt output is shown in Figure 6, where GPT-4 rates a resume for the "Sales" job description.

C Effect of Prompt Engineering

We analyzed the effect of prompt engineering techniques such as Chain of Thoughts (CoT) on GPT-4 ratings and observed its performance. We found prompt engineering improved the performance of GPT-4 ratings across the constructs of work experience, skills, education and certifications. For analyzing the performance of prompt techniques, we plotted the distribution graphs of differences between GPT ratings using different prompts and human ratings across work experience, skills, education and certification. Difference of zero in graphs signifies the exact match in ratings. Positive difference signifies GPT ratings to be greater, and negative differences signifies human ratings to be greater. Figures 7, 8, and 9 show the distributions of all the three GPT prompt ratings along with the zero-shot GPT ratings for Work experience, Skills and Education.

Additionally, we calculated the Pearson's correlation and Fleiss' kappa to inspect human and GPT-4 agreement for each prompt technique across the resume rating constructs. The detailed results for correlation and Fleiss' kappa are provided in Tables 9 and 10.

Rating:
- Work Experience Match: 5
- Skills Match: 4
- Educational Background Match: 3
- Certifications/Extracurricular Match: 5

Reasons for rating:

Work Experience Match:
Steven's extensive experience as a Real Estate Broker and Salesperson, directly aligns with the job requirement of being a licensed Realtor. His history of identifying residential home sale and leasing needs,...

Skills Match:
Steven has demonstrated strong skills in closing sales, client relationship management, market research, and negotiation all of which are crucial for the job. His experience in developing marketing strategies and using cloud software for business optimization shows his capability to close inbound leads and optimize transaction profitability....

Educational Background Match:
Steven holds an Associate in Science in Business Administration for Transfer, which provides him with a foundational understanding of business principles relevant to real estate investment and brokerage....
Certifications/Extracurricular Match:
Steven's certifications perfectly align with the job requirements. He is a licensed California Real Estate Salesperson, which is explicitly required for the job....

Figure 6: Example Truncated GPT-4 Rating for Sales job title

Table 9: Correlation for Prompt Engineering Techniques Compared to Human Raters

Category	Zero-shot GPT-4	Task Based	Task Based COT	Task Based COT with Example
Work Experience	0.1104	0.2643	0.2562	0.3625
Skills	0.0697	0.2174	0.2367	0.3208
Education	0.2317	0.4792	0.4961	0.6109
Certification	0.1497	0.3677	0.4310	0.6588

Table 10: Fleiss' Kappa for Prompt Engineering Techniques Compared to Human Raters

Category	Zero-shot GPT-4	Task Based	Task Based COT	Task Based COT with Example
Work Experience	0.0718	0.0913	0.0338	0.128
Skills	-0.068	-0.0478	-0.0442	0.028
Education	0.072	0.217	0.291	0.231
Certification	-0.113	0.268	0.378	0.444

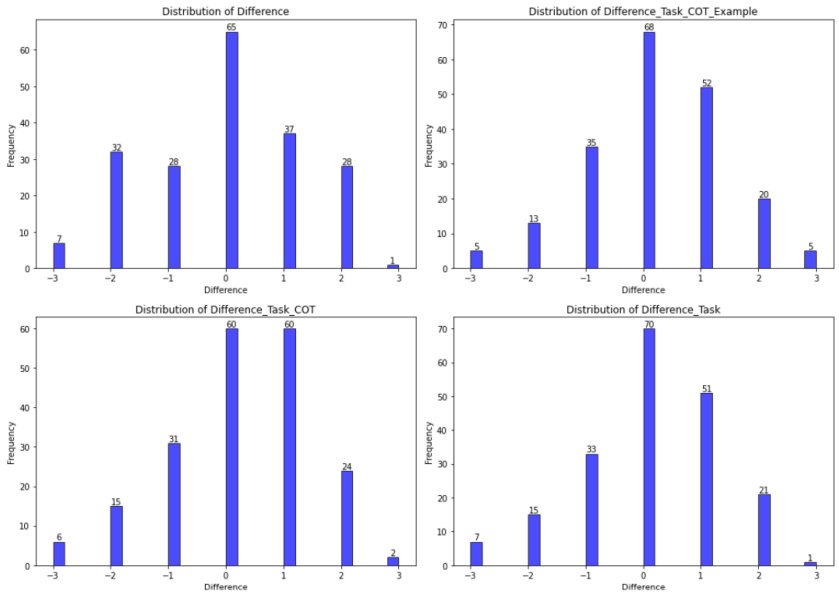

Figure 7: Distribution of Differences between GPT-4 and Human Ratings for Work Experience. Top left is zero-shot GPT rating, Top Right is Task based CoT With Example, Bottom Left is Task Based CoT and Bottom Right is Task Based Prompting

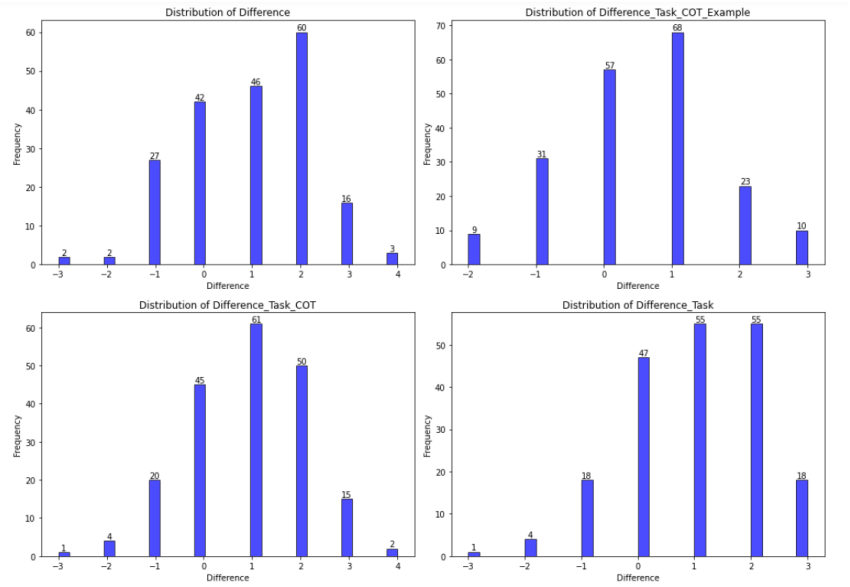

Figure 8: Distribution of Differences between GPT-4 and Human Ratings for Skills. Top left is zero-shot GPT rating, Top Right is Task based CoT With Example, Bottom Left is Task Based CoT and Bottom Right is Task Based Prompting

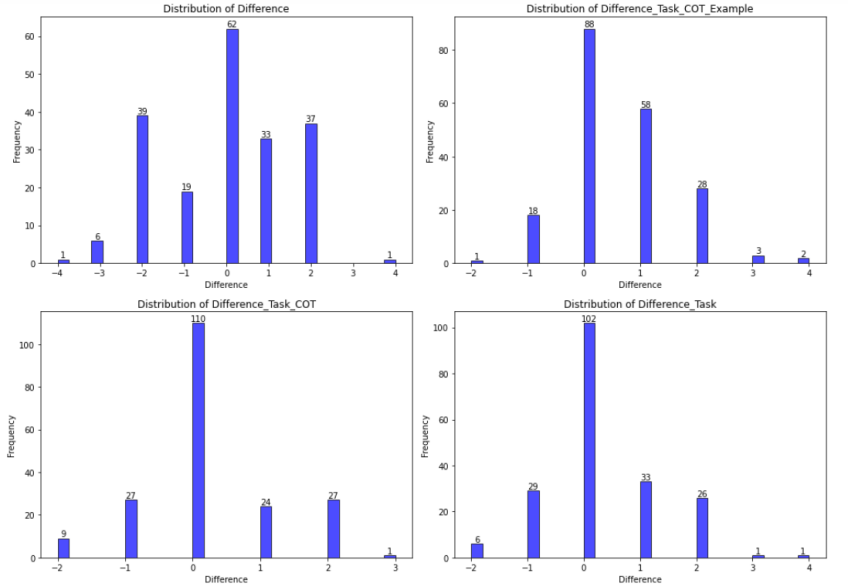

Figure 9: Distribution of Differences between GPT-4 and Human Ratings for Education. Top left is zero-shot GPT rating, Top Right is Task based CoT With Example, Bottom Left is Task Based CoT and Bottom Right is Task Based Prompting