The Unbearable Darkness Of Being (In The Job Market): Privacy and Reliability of LLM-based Candidate Evaluation

Junichi Koizumi
PERSUE Lab
Arizona State University
Tempe, AZ, USA

Lakshya Dhingra
PERSUE Lab
Arizona State University
Tempe, AZ, USA

Rakibul Hasan
PERSUE Lab
Arizona State University
Tempe, AZ, USA

Abstract

Large language models (LLM) are increasingly being used at every step of a hiring process, including resume parsing and candidate rating. This automation relies on LLMs' ability to judge both the relevant qualifications and the "personality" of the candidates; the latter in turn requires LLMs to "infer" personal information of the candidates. However, automated inference of personal data, correct or wrong, poses privacy threats to the applicants; and using inferred information in decision-making without a careful consideration may turn the hiring process further opaque, unreliable, and untrustworthy.

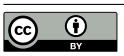
On the other side, job applicants' increasing use of LLMs to prepare or customize application materials further complicates the situation. It may blur the uniqueness of different applicants and confuse the hiring agents by making all applicants look equally good. This study investigates LLMs' capability to infer personality traits from resume—a common step in the hiring process—and how customizing resumes affect the accuracy of trait inference and the rating of a candidates for a specific job. To do that, we collected resumes from 72 undergraduate and graduate students, along with the big five personality trait measures. We then tested three popular LLMs in terms of how accurately and consistently they can predict traits, how they rank applicants, and how trait prediction and ranking change when the resumes are customized using a fourth LLM. Results demonstrate that LLMs perform poorly in inferring traits, and inter-rater reliability scores across LLMs are below chance-level. Likewise, the effect of resume customization on the rating was inconsistent across applicants. Using the harm-centric framing of privacy and the procedural justice theory, we discuss the implications of these findings on job seekers' right to privacy, and on the transparency and reliability of the recruitment process.

CCS Concepts

• **Security and privacy** → **Social aspects of security and privacy**; **Privacy protections**.

Keywords

Privacy, Large Language Models, Hiring, Inference, Privacy harms



LM-SHIELD '26, Bangalore, India

ACM ISBN 979-8-4007-2578-4/2026/06
https://doi.org/10.1145/3803628.3807974

ACM Reference Format:
Junichi Koizumi, Lakshya Dhingra, and Rakibul Hasan. 2026. The Unbearable Darkness Of Being (In The Job Market): Privacy and Reliability of LLM-based Candidate Evaluation. In *Workshop on Privacy in Large Language Models (LLM) and Natural Language Processing (LM-SHIELD '26), June 01–05, 2026, Bangalore, India.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3803628.3807974

1 Introduction

Large language models (LLM) have become ubiquitous in the job candidate evaluation pipeline. They parse resumes and other materials, short-list candidates for subsequent interviews, and analyze interviews to aid the final selection process [11, 12, 14]. This widespread automation, however, has created frustration among both job seekers and employers [9, 13]. Job seekers complain about the 'black-box' nature of candidate evaluation systems that provide no insights into the decision-making process, no explanation for why an applicant was rejected, or no feedback for improvement [18, 20]. Such opaqueness has compelled applicants to take extreme measures, such as using dating apps to connect with potential employers to avoid automated rejection by algorithms [5].

On the other side, LLMs have made it easy to create or polish resumes and other materials, allowing candidates to quickly apply for a large number of jobs. In a recent report, 55% of hiring managers stated that candidates use LLMs to craft resumes and cover letters [65]. This does not only flood the system but also may hinder the hiring agents' ability to find the right candidate if all resumes look similar to each other as they are written or polished using a handful of LLM services. This creates a vicious cycle that makes the hiring process increasingly frustrating for all.

This paper investigates the privacy, reliability, and transparency issues from using LLM in the application and hiring process. A common step in evaluating job candidates by LLMs, both in research and commercial tools, is to infer personality traits from resume and other written materials. This inferred knowledge may be provided to human evaluators or used in subsequent automated steps in the hiring pipeline. But LLMs' capability to infer such intimate personal data creates severe privacy risks, as they can be used for profiling, online targeting, and other secondary uses [29, 62, 66, 73]; indeed, recent reports suggest the prevalence of fake job postings for resume harvesting [3], presumably for large-scale amassing of personal data that fuels the inference economy [63].

Privacy harms may also occur when potential employers use application materials, which they had legitimately obtained, to automatically infer traits and use them in decision-making. Looking through the lens of harm-centric framework of privacy and procedural justice theory [49], using these inferred data for downstream decision-making may cause privacy harms, such as discrimination,

stereotyping, or harassment [28, 29, 62]. Because traits are complex psychological constructs, they may not be accurately or reliability inferred, and decisions made based on them may unfairly penalize data subjects (e.g., a financial harm [28] when a deserving candidate does not get a job). More, the distribution of traits can differ across populations, creating inconsistency in decision-making [37]. These harms may arise even when LLMs (or other AI tools) are used to directly evaluate and rank candidate, without explicitly asking them to infer traits. For example, a recent study reported that ChatGPT extracted and over relied on disability-related information from the resume without being asked to do so [35]. All of these may contribute to making the hiring process opaque and untrustworthy, and the decisions from that process unexpected and disheartening.

Finally, using LLMs to customize resumes may impact the automated hiring process in surprising ways, either positively or negatively. In one hand, customization may better highlight a candidate's unique strengths and fit to a job description, and help mitigate some biases (e.g., bias against non-native English speakers). On the other hand, LLMs generated text may obscure candidate's unique writing style, making all resumes similarly 'polished.' This may confuse the hiring agent, make its decision process further unreliable, and the outcomes may be more random and inconsistent, and unfair to the candidates. Consequences for hiring organizations can be substantial as well: if their selection procedures are arbitrary or unfair, they risk litigation and class action lawsuits. As such, any algorithm deployed in the field of hiring needs rigorous scrutiny.

Thus, it is both critical and urgent to investigate the privacy, reliability, and transparency aspects of LLMs in the context of job candidate evaluation. We contribute toward this by investigating the extent to which LLMs create privacy risks through inference, and their consistency in evaluating traits and rating job candidates. Concretely, this study aims to answer the following questions:

RQ1: How well do LLMs predict personality traits from resume? How consistent they are in that prediction?

RQ2: How does resume customization affect trait prediction?

RQ3: How does resume customization affect candidate rating?

To answer these questions, we collected resumes and ground truth personality traits from 72 students at a large public university in the US. We then created customized versions of these resumes using the DeepSeek model. Another three popular LLMs (GPT, Llama, and GritLM) were employed to predict the big five personality traits [64] and rate candidates using both the original and the customized resumes. We found low inference accuracy from all LLMs for all traits (§4.2). All models systematically either overestimated or underestimated traits, indicating the inherent unreliability of inferring traits from resume. We computed how consistent the model predictions are with respect to the ground truth and across models, and found that, in both cases, the inter-rater reliability scores were below chance agreement (§4.3). Trait inference from the customized versions of the resumes had non-uniform impacts: for some traits, it improved accuracy, for others, it decreased accuracy (§4.4). Finally, customization led to increased rating of the candidates most of the time, but not always; and ratings across models were inconsistent for both original and custom resumes (§4.5).

These results advance our understanding of how the use of LLMs in hiring might undermine job applicants' privacy and add confusion to the already opaque hiring process. We discuss the implications of these findings based on the harm-centric framework of data privacy [28, 34, 62], the procedural justice theory [49], and legal protections granted by various regulations such as GDPR [4] and CCPA [2]. To the best of our knowledge, this is the first study that systematically investigated the privacy and transparency of LLMs in the context of trait prediction and job candidate rating, and how they are impacted by resume customization, also using LLMs. We believe that the findings will inspire further research and inform AI deployment and governance policies.

2 Background and Related Work

This section reviews prior art on the use of LLMs and other types of artificial intelligence for job candidate evaluation, and the associated issues of privacy violations, algorithmic bias and discrimination, and a lack of transparency and trustworthiness.

2.1 Use of LLM to assess job candidates and their personality

Large language models (LLM) and other AI-based tools are increasingly being used to evaluate job candidates before hiring. Reportedly, such tools are being used by 51% of companies [36], including most of the Fortune 500 companies. The first step for candidate evaluation is now parsing resumes using language processing tools [59]But language models are increasingly being integrated into all subsequent stages of human resource management, including short listing candidates for an interview, making final selections and optimizing job offers, and after-hire evaluations [41].

Personality trait assessment. Predicting or evaluating candidate personality has been an integral part of the hiring process, as personality traits are thought to be indicative of job performance, management skills, and other characteristics relevant to employment [23, 31, 38, 40]. Among various traits, the big five personality traits—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [64]—has been the most widely used in AI-based hiring tools [38, 59]. Many commercial tools provide big five trait assessment that are integrated into the hiring pipeline. For example, *Humantic AI* [10] and *Crystal* [8], two tools audited in a prior work are reportedly being used by 90% of the Fortune 500 companies, including tech giants like Apple [59].

Accuracy and reliability of trait prediction. Prior works tested various machine learning models on trait prediction and reported mediocre-at-best performance. One work trained a custom model on resumes and free-text interview answers collected during recruiting process to predict applicants' Big Five personality traits; the results showed moderate predictive accuracy [38]. Another study used an LLM to analyze transcripts of video interviews; the models showed predictive performance comparable to conventional personality assessment methods, but exhibited systematic biases across trait dimensions [76]. Hilliard *et al.* compared LLMs with varying parameter values and reported that most models consistently overestimated agreeableness, openness, and emotional stability from interview prompts [42]. Zhu *et al.* tested three state-of-the-art LLMs using zero-shot and chain-of-thought prompting to

predict the big five traits from interview data. They found weak correlations with ground-truth across models (max Pearson's $r = 0.27$), low inter-rater agreement (Cohen's $\kappa < 0.10$), and predictions to be biased toward moderate or high trait levels [77]. Finally, Alene *et al.* conducted a controlled experiment to audit two commercial systems designed to predict personality from resumes and social media profiles: Humantic AI and Crystal [59]. The audit revealed a lack of reliability in predictions that were being affected by seemingly irrelevant changes to the resume, such as using a different template, without any change in the actual content [59].

2.2 Privacy concerns

AI-based recruitment raises fundamental questions about applicants' right to privacy from multiple perspectives. One of the core tenets of privacy is individual freedom in choosing how their data is disclosed, used, and disseminated. But as LLMs, and AI in general, 'infer' data about an individual that they did not intend to disclose, their right to privacy is violated and their agency has been denied [15, 34, 54]. More, the inferred data might be used by the LLM in downstream decision-making, or extracted by others (e.g., by asking the LLM to explain its decision). For example, Glazko *et al.* found that ChatGPT was biased against individuals whose resumes contained roles or awards that suggested a disability status [35]. This ability to infer new information, and its extraction and (secondary) use, may lead to further harms like discrimination, harassment, or reputation loss [28, 34, 62]

Besides denying control over data, inferences violate privacy from the perspective of impression management: the process by which individuals influence how others perceive them [48]. People's desire to manage a good impression underlies privacy-seeking behaviors [46, 70]. Job seekers might selectively disclose information based on how they want to be perceived by the recruiter (and future colleagues). Consider the professional and personal consequences if, after being hired, a person learns that their colleagues have a pre-conceived assumption about them of being aggressive or rude (which they came to know through LLM-based personality inference).

Other, more direct, privacy risks may arise when LLMs parses resumes. In the USA, there is a multi-billion dollar data broker industry. In recent years, there has been also a growth in what legal scholars call an "inference economy," where predictions about individuals become commodified and traded [63]. As bits and pieces of information about individuals travel through this system and merged together to profile the data subjects, potential for privacy harms increases. This situation has prompted arguments for a "*right to reasonable inference*," to establish boundaries on what can legitimately be predicted about individuals without their knowledge or consent [71]. Recent regulations, such as the California Consumer Privacy Act (CCPA [2]) and the General Data Protection Regularion (GDPR [4]), provide inferred data the same level of protection as directly collected personal information.

It is crucial to note that, the correctness of the inference is often irrelevant to privacy harms; even if AI systems incorrectly infer something, the privacy harms remain the same (if not worsen). Wrongly assuming someone's demographics, personality, or disability status still denies them their right to control personal data, still creates a false impression, and still allows an advertiser to (wrongly) target them based on the profile created; in fact, in the hiring context, wrong inferences may cause more harms in terms of inter-personal relationship building and expectation management. We further note that, while personality (and other) assessments might be made during manual resume screening (or a face-to-face interview), when AI automates this manual processes, it also removes the opportunity to contest or correct wrong assumptions, does it at a much larger scale [15], and allows indefinite retention and secondary use [61].

2.3 Bias and fairness

We review past research on bias and fairness as they are intricately linked to privacy. Recent review papers reported how biases in hiring agents harm both organizations and job seekers [27, 53]. Many studies found discriminatory behaviors across race [74], gender, and race-gender intersections [17, 19], age [39], disability status [35], pregnancy status, and political affiliation [69]. One experiment compared the level of bias across different factors; it found that that applicants with hearing disabilities face greater discrimination than those facing ethnicity, gender, or residence-based discrimination, particularly in public-facing occupations and public sector roles [51]. Such discriminatory behaviors extend to multi-modal data (e.g., simultaneously using resumes and video interviews) as well [21, 25].

2.4 Procedural justice of algorithmic hiring

Procedural justice requires not only that a decision is accurate, but also that the process of making that decision is reliable, consistent, and transparent, as well as trustworthy to the affected population [34, 49]. Prior research has investigated this aspect of automated hiring. One research surveyed 448 computer science students about perceptions of the procedural fairness of automated hiring systems. It reported that job seekers' perceived AI-based hiring decisions as less fair than human judgment, they distrusted automation in this context and expressed a lack of willingness to participate in such a process [20]. In another study, Pyle *et al.* uncovered job seekers' perceived harms around privacy and discrimination from asynchronous interview systems that use the recorded interviews to infer emotion, which might be used to evaluate the candidates. Analyzing online forum posts, He *et al.* documented fairness concerns among job seekers and uncovered four key issues: discrimination against sensitive attributes, interaction bias, unfair interpretations of qualifications, and power imbalances [41]. Another study surveyed US college students about their perceived privacy harms from personality trait prediction by hiring agents; participants expressed concerns about being discriminated, stereotyped, and harassed if future employers learn about such details [34].

Research gap. While many prior works examined LLM's capability to predict traits, they used interview transcripts, chat logs, or social media data. In practice, job candidates first get evaluated based on their resumes, which is getting increasingly automated with LLMs. Further, prior research focused on accuracy, but the impact of trait inference on privacy under different privacy frameworks has not been systematically evaluated. Finally, we did not find any

research on how resume customization, which has become ubiquitous, impacts trait prediction and candidate ranking. Shedding light on these aspects are both critical and urgent to understand the transparency and reliability of LLM-based hiring agents.

3 Methodology

3.1 Survey questionnaire and data collection procedure

We conducted an online study designed following the "data donation" paradigm [56]. Study participants were asked to donate their resume to help discovering privacy implications from automated resume evaluation tools. Note that, uploading resume was optional. Next, participants were asked what job application materials they have ever customized using artificial intelligence. Afterwards, participants were asked to enter three job titles that they were interested in.

The following question block asked a series of 30 questions to measure personality traits. This part used one of the most popular psychometric scales BFI-2-S [64] that measures the big five personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each question instructed the participants to indicate the extent to which they agree or disagree with statements like "I am someone who..." using a five-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree." Our selection of these traits was informed by the extensive literature demonstrating the wide use of these traits in job candidate evaluation (both by humans and AI-based tools [59], and their strong associations with employment-relevant outcomes such as problem-solving ability, creativity [44, 60]. The survey concluded with demographic questions about age, race, gender, education level, and study major.

The study was reviewed and approved by our institutional ethics board. Participants were provided a consent form prior to answering questions. The survey was designed on Qualtrics [7] and advertised throughout our campus (a large public university in the USA) by distributing a short description and the survey URL. The description included information about compensation: we will randomly select 20% of the participants, each will receive a USD 25 gift card. Data collection was done in November 2025.

3.2 Data analysis

Figure 1 summarizes the pre-processing and analysis steps to answer the research questions. We describe them below.

3.2.1 Resume pre-processing. Participants uploaded resumes in either PDF or word format. First, PDF documents were converted to DOCX format using Adobe Acrobat. Next, we used a python library to parse these documents and extract text content from all sections and paragraphs, removing formatting contents such as line dividers. Finally, the first author manually removed personal information, such as names, email and physical addresses, and gender (when present), as such information can potentially bias the LLM models in the subsequent analysis steps.

3.2.2 Resume customization using LLM models. To measure the impact of customization on trait prediction and candidate ranking, we customized each resume for specific job advertisements. Advertisements were collected from LinkedIn [6], one of the most popular professional networking platforms. For this step, two of the authors searched for advertisements using the first job title provided by the participants. Both authors then collaboratively reviewed the individually collected advertisements and selected one per resume, optimizing for similarity with the job title mentioned by the corresponding participant and uniformity in job description lengths as much as possible. We list the top job titles listed by our study participants in Appendix A.1; this list shows the diversity and representativeness of the employment roles used in this study.

Next, each resume was customized to fit the description of the first job title mentioned by the resume author. We used the DeepSeek-R1-Qwen-14B [32] model for this purpose. We note that, in this and all other steps involving LLMs, we locally deployed publicly available models. We did not use any external services to avoid sharing resumes which might have privacy and ethical implications. This choice, however, also prevented us from using the latest or the largest models due to resource constraints. But the impact of model choice on results is expected to be minimal. Past benchmarking research showed that the DeepSeek 14B model posses adequate linguistic capacity for text adaptation tasks [22] and that it achieved competitive performance on text summarization and adaptation tasks compared to larger models [55]. All experiments were conducted on an NVIDIA A100-SXM4-80GB GPU with 80GB VRAM, 8 CPU cores, and 256GB RAM. The DeepSeek was configured to have temperature set to 1 (so it would not be biased to deviate from the training data distribution to generate tokens), token count to 2048 to avoid resume truncation, and handle variable-length inputs.

To design the LLM prompt for customization, we first explored online sources to identify popular prompts used by job candidates in practice (e.g., [68]). After trialing multiple versions and manual review of the output, we settled with the following prompt adapted from [68]:

```
-------------Prompt for resume customization-----------
I'm applying for a [Job Title] position in [Industry],
and I want my resume [Redacted Resume] to stand out.
Based on the job description: [Job Description].
Please rewrite the professional summary (or add one if
it does not exist) that aligns with the job description
and qualifications listed in the resume, and make it
concise and engaging. Also revise or reorder content in
the whole resume so that it reflects my strengths,
relevant skills, and years of experience in a
compelling way.
-----------------------------------------------------
```

Two authors manually reviewed the generated resumes and found that the model generated properly formatted and contextually appropriate resumes that aligned job descriptions with candidate experience, and meeting our requirements for semantic coherence and stylistic consistency.

3.2.3 Predicting personality traits using LLMs. We used another three popular models for trait prediction: gpt-oss-20B, GritLM-7B, and Llama-3.1-8B-Instruct [58, 67]. To ensure stochastic response generation while maintaining the quality of the output, all four models were configured with the identical parameters except for temperature, which was set to 0.7. This slightly lower value conditions the models to avoid diversity or randomness [57] as the goal

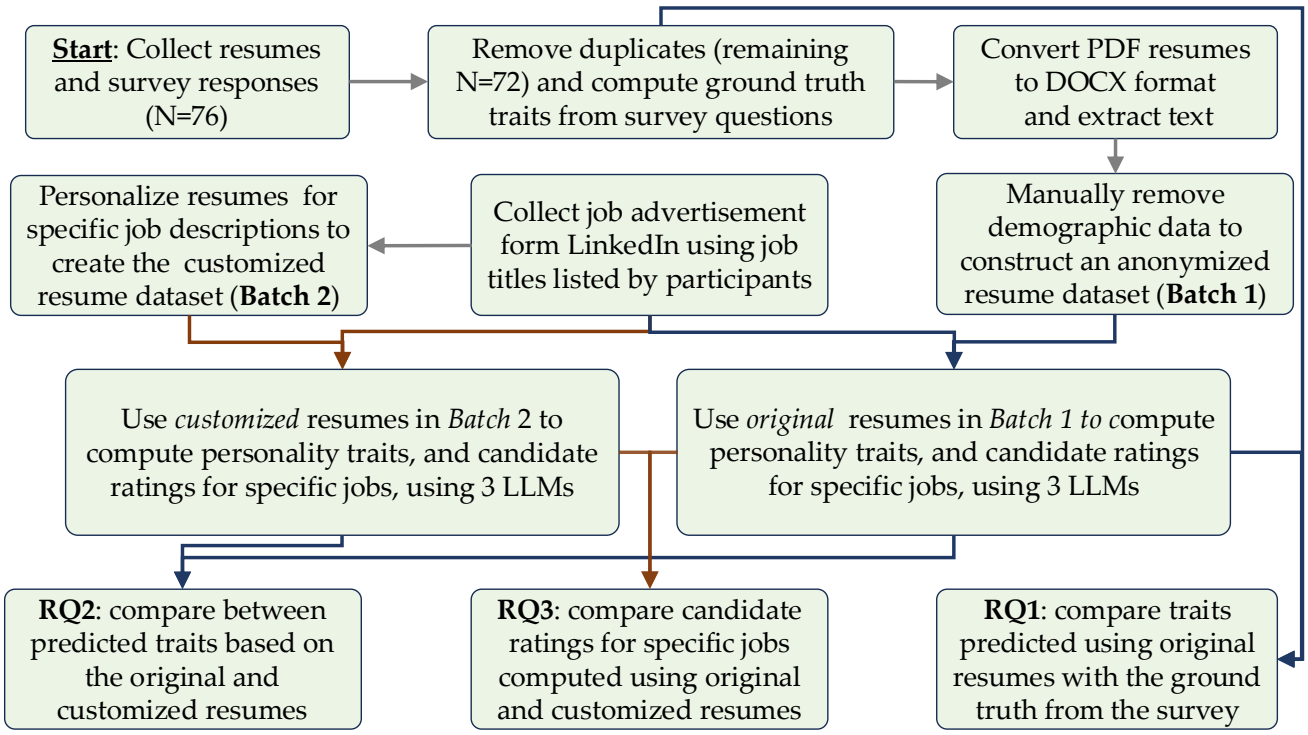


Figure 1: Data collection and analysis pipeline.

here is to predict set categories. Specifically, using a resume, the models were asked to predict if a participant had "high" or "low" level of each of the five traits (binary categorization). Due to the GPU memory constraints, models were loaded using 8 bit quantization using the BitsAndBytesConfig library [1]. Quantization does not cause noticeable performance degradation [33] but allowed us to use models in a single GPU.

3.3 Prompt engineering for trait prediction

For trait prediction, the prompt design process was more involved compared to resume customization. We employed zero-shot Chain-of-Thought (CoT) prompting; this choice was informed by prior research demonstrating its effectiveness in enhancing large language models' performance on text based personality recognition tasks [43]. We started with a simple prompt: *For each resume, analyze the* `resume_text` *and determine the person's levels of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Only return "Low" or "High" for each trait. Output as a table with one row per person.*

One author manually reviewed the generated responses from each model and identified four issues:(1) all models occasionally produced responses containing irrelevant details (2) formatting inconsistencies were prevalent, with outputs failed to maintain a specific output header schema, (3) GritLM omitted at least one personality trait assessment in the output most of the times, and (4) GPT included reasoning that were contradictory for the same resume. To address these issues, we refined the initial prompt in multiple passes followed by manual reviews. First, we incorporated explicit constraints, including header schema and strict grounding instructions forcing the model to base the analysis on the provided resume content. This approach aligns with the findings by Banghao et al., who demonstrated that contextual constraints improved output relevance drastically in large language models [26]. Second, we adopted a one-shot-prompting strategy by providing an example resume. The example was manually customized using one of the author's (a computer science student) resume to fit for a software engineering role. Providing an example established a clearer output format and reasoning, thereby reducing ambiguity in models [26]. Finally, we reminded the models to force an output as sometimes they skipped producing an output for a certain trait. The final prompt is provided below.

```
----------Prompt for trait prediction-----------
Task: Analyze the {resume_text} and determine the
person's levels of Big Five Personality traits:
Openness, Conscientiousness,Extraversion,
Agreeableness, and Neuroticism.

Instructions:
- Only output a table with one resume
- For each trait generate a 'high' or 'low' score
- Do NOT make multiple tables for the same person
- Do NOT make empty tables with no outputs
- DO NOT output with all highs or lows or duplicates
```

```
Reasoning: high
THINK ONE STEP BY STEP

Example output format:
| Name | Openness | Conscientiousness | Extraversion|
|------|----------|-------------------|--------------|
| name | [high or low] | [high or low] | [high or low] |
| Agreeableness | Neuroticism |
|---------------|-------------|
| [high or low] |[high or low]|

Based on the resume provided, infer the most likely
personality scores even if evidence is incomplete.
Do NOT refuse to answer.
-----------------------------------------------------
```

3.3.1 Rating job candidates. To evaluate if and how resume customization impacts applicant ranking, we asked the same three LLM models to rate each participant in a scale from one to five (higher ranking is better). Candidates were rated both using the original and the customized resume (separately); both times the rating was done for the job the resume was customized for. The prompt for candidate rating is shown below.

```
--------------Prompt for candidate rating------------
 You are a recruiter. You are tasked with evaluating each
 candidates Resume content and decide whether they are
 a good fit for the job position or not.

 Job Position:{job}
 Resume: {resume_text}

 Instructions:
 - Rate the resume from 1 to 5 (1 = lowest,
   5 = highest)
 - Base your judgment on qualifications, experience,
 and job fit
 - MAKE SURE to consider the provided everything
 written in the resume text for evaluation
 - Think step by step about each resume's strengths
 and weaknesses based solely on the resume text
 - Output the rating in the following format EXACTLY:

 Rating: [number]
-------------------------------------------------
```

3.4 Computing and comparing traits and rankings

To answer *RQ1*, we compared the ground truth traits with those predicted by LLMs using the original resumes after redacting personal data (*Batch1* in Figure 1). The ground truth for the five traits were computed from the responses to the big five inventory. For LLM prediction, the problem was converted to a binary classification; specifically, LLMs were instructed to predict if a resume indicated a *high* or *low* level of each trait. This was done to reduce nuance and simplify the task for LLMs. To binarize the ground truth trait scores (ranging from 1 to 5), we split them based on the sample median: each score higher than the median was labeled as "High" (and "Low" otherwise). The LLM-predicted binary classes were then compared with the ground truth labels.

For *RQ2* (effect of customization on trait prediction), the above process was repeated, but this time with customized resumes (Batch2 in Figure 1). We further compared the predicted traits from customized resumes to that with original resumes.

To assess how resume customization impact the ranking of job candidates (*RQ3*), we compared the two ratings, one produced using the original resume and another the customized version. Each time an LLM was used, it was initialized from scratch to prevent contamination from previous executions.

4 Results

4.1 Participants and their personality traits

We received 102 survey responses. Among them, 76 included a resume, but four of them were duplicate responses. Our subsequent analyses will only include the 72 unique responses. Detailed breakdown of the participants are listed in Table 1. The random selection for gift card was done on the full set of participants, regardless of if they uploaded a resume or not.

Table 2 presents the mean, median, and standard deviation of participants' trait scores, along with Cronbach's alpha that indicates the reliability of these scores. Cronbach's alpha was higher than 0.7 for each of the five traits, indicating high reliability [24]. Prior works using variants of the scale reported comparable scores (e.g., [45, 50]), suggesting representativeness of our sample. Thus, we treat these scores as the "ground truth" trait scores for the subsequent analyses.

4.2 Trait prediction accuracy

As explained in §3.4, to simplify the trait prediction task for LLMs, we converted it to a binary classification problem. Three LLMs were provided the original resume of each participant and prompted to predict if that participant had a "Low" or "High" level of each trait. Table 3 presents the prediction accuracy compared to the ground truth. For most traits, all three models predicted correctly between 40 to 60% of the time. Note that, since we divided the two classes based on median value, we had equal number of samples in both classes. Thus, all models performed close to the random guess level, sometimes below that. Across the five traits, GritLM's performance was the most consistent, while GPT and Llama exhibited more volatility. Thus, LLM performance is neither sufficiently accurate nor consistent across traits, even after simplifying trait prediction as a binary classification task.

To get a more nuanced understanding of when LLMs make correct or wrong predictions, we computed confusion matrices for each trait and model (Table 4). Note that, TP (True Positive) refers to the cases where both the ground truth category and LLM prediction were "High," TN (True Negative) refers to cases where both were "Low." Overall, all models consistently overestimated (i.e., predicted "High" when the ground truth was "Low") openness, conscientiousness, and agreeableness. Extraversion was also overestimated by all models, but not to the same degree. In contrast, all models underestimated neuroticism, predicting "Low" when the ground truth was "High." These results are intuitive: resumes are more likely to express traits that are favorably judged, such as openness to

Table 1: Participant Demographics (N = 72)

Characteristic	N (%)
Gender	
Male	45 (62.5%)
Female	27 (37.5%)
Age	
18–24	28 (38.9%)
25–34	15 (20.8%)
35–44	13 (18.1%)
45–54	13 (18.1%)
55–64	2 (2.8%)
Race / Ethnicity	
White	33 (45.8%)
South Asian	11 (15.3%)
Hispanic, Latino, or Spanish Origin	7 (9.7%)
East Asian	6 (8.3%)
Mixed	4 (5.6%)
South East Asian	4 (5.6%)
Black or African American	3 (4.2%)
Middle Eastern, Arab	1 (1.4%)
Native Hawaiian or Other Pacific Islander	1 (1.4%)
Other (White Sephardic Jewish)	1 (1.4%)
Other (White Ashkenazi Jewish and White European ancestry)	1 (1.4%)
Missing	1 (1.4%)
Education	
Undergraduate degree	41 (56.9%)
Postgraduate degree (MA/MSc/MPhil/other)	17 (23.6%)
High school diploma / A-levels	11 (15.3%)
Other	2 (2.8%)
Associate Degree (AS)	1 (1.4%)
Field of Study	
STEM	48 (66.7%)
Non-STEM	24 (33.3%)
Prior AI Use for Job Applications	
Full resume	28
Cover letter	27
Summary / personal statement	20
Never used AI for this purpose	15

new experiences and challenges, conscientiousness that correlates with well-thought and organized workflows and competence, and agreeableness that correlates with friendly and cooperative nature. Further, these traits were shown to positively correlate with job performance [23, 31, 40]. On the other hand, neuroticism hints at emotional instability and a lack of confidence; thus professional resumes are unlikely to contain any signals to this trait. In summary, trait prediction from resume will most often be systematically incorrect: overestimating some traits while underestimating others. This observation aligns with Hilliard *et al.*, who investigated personality trait "exhibited" by LLM models themselves, and reported systematic over- or under-estimation [42].

Table 2: Reliability and Statistics of 30-item Big Five Personality Assessment

Trait	Mean	Median	SD	Cronbach's α
Openness	3.78	3.67	0.74	0.709
Conscientiousness	3.81	3.83	0.80	0.802
Extraversion	3.17	3.17	0.77	0.738
Agreeableness	3.88	4.00	0.79	0.823
Neuroticism	2.67	2.67	0.92	0.843

Takeaways:
Overall, trait prediction accuracy was around the chance level. All models overestimated Openness, Conscientiousness, and Agreeableness; underestimated Neuroticism; and mixed up low and high levels of Extraversion.

4.3 Trait prediction reliability

This section presents results from inter-rater reliability analysis. First, we computed how reliable LLM predictions are with respect to the ground truth. That is, we treat the ground truth computed from the psychometric scale and the LLM predicted categories as coming from different raters or annotators. Inter-rater reliability metrics take care of chance agreements (i.e., two annotators using the same class by chance) and thus is a better measure than accuracy to understand how reliable a model is. Since there are two raters, we use the Cohen's Kappa metric to assess reliability. Table 5 shows the results: in most cases, the reliability score is negative or near zero, indicating worse than chance agreement [47]. For extraversion, GPT achieves a score of 0.25, which is still below the threshold of acceptable reliability [47].

Next, we compute reliability among the models. That is, we treat each model as a rater and investigate how reliably they converge to the same decision. Since we have three raters (models), we use Fleiss' Kappa, which is an extension of Cohen's Kappa for more than two raters [16]. Table 6 shows the results. Extraversion reached the highest score, similar to computing reliability with respect to the ground truth. But the score is still close to zero, indicating no better than chance agreement. The other four traits had negative values, indicating worse than chance agreement [47]. These results suggest that the same resume may be interpreted in substantially different ways by different LLM-driven hiring agents.

Takeaways:
Trait assessments from different LLMs do not agree with each other, questioning the reliability of such assessments.

4.4 Impact of customizing resume on trait prediction

This section details how customizing the resumes for a specific job impacted trait predictions. As explained earlier, we used DeepSeek for customization, and the same three models (GPT, GritLM, and Llama) were used to predict traits from the customized resumes.

Table 3: Accuracy of (binary: "Low" or "High") trait prediction from original resumes.

Model	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
gpt-oss-20B	58%	53%	63%	51%	47%
GritLM-7B	57%	54%	50%	50%	54%
Llama 3.1-8B	61%	49%	46%	49%	44%

Table 4: Confusion matrices of trait prediction from original resumes compared to ground truth

Model	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
GPT-OSS-20B	TN: 0 FP: 27 FN: 3 TP: 42	TN: 0 FP: 34 FN: 0 TP: 38	TN: 18 FP: 18 FN: 9 TP: 27	TN: 3 FP: 34 FN: 1 TP: 34	TN: 34 FP: 0 FN: 38 TP: 0
GritLM-7B	TN: 5 FP: 23 FN: 8 TP: 36	TN: 2 FP: 31 FN: 2 TP: 37	TN: 9 FP: 28 FN: 8 TP: 27	TN: 0 FP: 36 FN: 0 TP: 36	TN: 26 FP: 9 FN: 24 TP: 13
Llama3.1-8B	TN: 1 FP: 27 FN: 1 TP: 43	TN: 4 FP: 30 FN: 7 TP: 31	TN: 5 FP: 32 FN: 7 TP: 28	TN: 1 FP: 36 FN: 1 TP: 34	TN: 32 FP: 2 FN: 38 TP: 0

Table 5: Cohen's Kappa(κ) between ground truth trait scores and scores predicted using original resumes.

Model	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
gpt-oss-20B	-0.081	0.000	0.250	0.051	0.000
GritLM-7B	-0.024	0.000	-0.017	0.000	0.103
Llama 3.1-8B	0.016	-0.069	-0.064	-0.002	-0.056

Table 6: Consistency among the three LLM models in predicting traits from original resumes.

Personality trait	Fleiss' Kappa (κ)
Openness	−0.015
Conscientiousness	−0.051
Extraversion	0.004
Agreeableness	−0.022
Neuroticism	−0.100

4.4.1 Comparison with ground truth. First, we compared traits predicted from customized resumes with the ground truth traits. Table 7 shows the accuracy scores, with accuracy from original resumes are number within parentheses for comparison. When customized resumes were used, Llama performed consistently better compared to when the original resumes were used. Results for GPT and GritLM were mixed with no obvious pattern. For GPT, custom resumes heavily degraded performance for extraversion, but slightly improved performance for other traits. For GritLM, changes in performance in either direction were small for across traits. In summary, customizing resumes impacted LLM prediction but the impacts are non-uniform across traits and models.

4.4.2 Comparison with previous predictions. Next, we investigated if trait predictions are consistent across resume versions, that is, if the same model predicts the same trait level based on both the original and customized resume. Table 8 shows the results in terms of accuracy where a prediction is "correct" if it remained unchanged across resume versions. As the table shows, customization changed prediction for some traits but not others. In particular, all models drastically changed predictions for extraversion when the customized version was used. Also noteworthy: for neuroticism, GritLM's prediction for customized resume matched fewer than half of the times with prediction for the original resume.

For a finer diagnostic, we again computed confusion matrices for predictions from customized resumes (Table 9). TP (true positive) refers to cases when a model predicted "High" for both the original and customized resumes, and TN (true negative) denotes "Low" predictions for both versions. FP (false positive) refers to cases when original prediction was "Low" and customization flipped it, and FN (false negative) indicates flipping "High" predictions.

For conscientiosness and agreeableness, predictions are overwhelmingly in the TP cell for all models. Recall that all models overestimated these traits (see Table 4), and this trend continues for customized versions. Openness was overestimated by all models as well (Table 4); this behavior remains consistent for GPT and Llama, but GritLM yielded moderately large number of FP and FN, indicating that it switched predictions (from "High" to "Low" or vice versa) for some resumes across versions. This pattern is stronger for extraversion across all models: we see slight increase of FP but a larger increase of FN, again indicating changes in predictions,

Table 7: Accuracy of trait prediction from customized resumes compared to ground truth.

Model	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
gpt-oss-20B	59% (58%)	54% (53%)	44% (63%)	56% (51%)	48% (47%)
GritLM-7B	54% (57%)	54% (54%)	47% (50%)	51% (50%)	45% (54%)
Llama 3.1-8B	62% (61%)	56% (49%)	56% (46%)	52% (49%)	48% (44%)

Table 8: Accuracy of trait prediction using customized resume compared to prediction from original resume.

Model	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
gpt-oss-20B	93%	100%	69%	87%	100%
GritLM-7B	71%	94%	64%	100%	48%
Llama 3.1-8B	90%	81%	61%	96%	97%

with most going from "High" (for original resumes) to "Low" (for customized resumes). Perhaps the most interesting results were for neuroticism. Recall that, unlike the other four traits, neuroticism was underestimated by all models, but more so by GPT and Llama (see Table 4); results for customized resumes remained faithful to prior estimates. GritLM also underestimated neuroticism before, but not to the same degree as the other two models. But this tendency increased after customizing resumes, as indicated by a larger number of false positives in Table 9 than before, and a reduction in false negatives. That means, GritLM switched from "Low" to "High" for customized resumes. This is counterintuitive. One would expect that polishing a resume will make the writing style appear more confident. However, after customization, GritLM more frequently predicted high values of neuroticism, which indicates a lack of confidence and ability to manage emotion.

Takeaways:
Customizing resumes affected trait inference; but the effect was non-uniform across models, raising further concerns about reliability.

4.5 Candidate rating

Here, we present if LLMs rated the same participant differently based on the customized resume compared to the original resume. How participant rating changed after customization is visualized in Figure 2. The numbers in each diagram indicates the rating of the candidate in terms of their fit to a specific job (1 indicates the poorest fit, 5 indicates the best fit).

The upward transition in all plots indicate that all models increased rating for the customized version compared to the original resume. In particular, GritLM and Llama rated a large number of original resumes in 1, 2, and 3; but rated almost all resumes 4 or 5 after customization. For a handful of resumes the rating dropped after customization, but the overall trend indicates that customization for a specific job may potentially increase the likelihood to pass the screening phase. However, if all resumes are customized, then the net effect might be more confusion!

Next, we investigated if the ratings were consistent across models. Since the rating has more than two categories (five levels in this case), we compute Krippendorff's alpha. For ratings using the original resumes, the alpha value was 0.112, which is considered to be low level of reliability [30]. Interestingly, the inter-LLM agreement in ratings *decreased* after resume customization ($\alpha = 0.024$). Thus, while customization generally improved ratings, its effect was non-uniform across the models.

Takeaways:
Customizing resumes improves candidate rating in most cases, but not all. The improvement is heterogeneous across models.

5 Discussions

This study sheds light on how LLMs, as part of a hiring pipeline, impact the privacy of job applicants, and the transparency of the hiring process. We find that, LLMs could not predict a personality trait better than a random guess, even after simplifying the problem to a binary classification task. This inaccuracy is likely in part because people tend to highlight desirable characters (e.g., openness) while obscuring undesirable ones (e.g., neuroticism) in their resumes. This result coincides with the findings from Hilliard *et al.*, who examined trait distributions among LLM themselves and reported that LLMs selectively exhibit high or low levels of traits [42]. Thus, a direct implication of this result is that automated trait prediction, either using general purpose LLMs or custom machine learning models, from resumes and other job application materials may always be systematically biased. Thus, LLM-inference risks misrepresenting job applicants, likely in direct conflict of privacy laws such as CCPA [2] and GDPR [4]. As privacy scholars Solove and Citron argued [28, 62], privacy violations must include harms caused by the use of personal data, not just the leak or collection of data. Under this conceptualization, both correct and incorrect trait inference may lead to privacy harms. In the former case, it may allow anyone to amass resumes and profile people for later targeting; such profiling may further lead to manipulation, stereotyping, and harassment [34]. Incorrect inferences might lead to more direct negative consequences, like economic harms [28] when deserving candidates are denied opportunities because of misrepresentation. One could argue that personality assessment might be a part of manual resume evaluation and face-to-face interviews. However,

Table 9: Confusion matrices of Anonymized personality prediction VS Customized personality prediction.

Model	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
GPT-OSS-20B	TN: 0 FP: 2 FN: 3 TP: 65	TN: 0 FP: 0 FN: 0 TP: 70	TN: 20 FP: 6 FN: 16 TP: 28	TN: 0 FP: 3 FN: 6 TP: 61	TN: 70 FP: 0 FN: 0 TP: 0
GritLM-7B	TN: 3 FP: 10 FN: 10 TP: 47	TN: 0 FP: 4 FN: 0 TP: 66	TN: 7 FP: 10 FN: 15 TP: 38	TN: 0 FP: 0 FN: 0 TP: 70	TN: 28 FP: 22 FN: 16 TP: 4
Llama3.1-8B	TN: 1 FP: 1 FN: 6 TP: 62	TN: 2 FP: 9 FN: 4 TP: 55	TN: 6 FP: 6 FN: 21 TP: 37	TN: 0 FP: 2 FN: 1 TP: 67	TN: 68 FP: 0 FN: 2 TP: 0

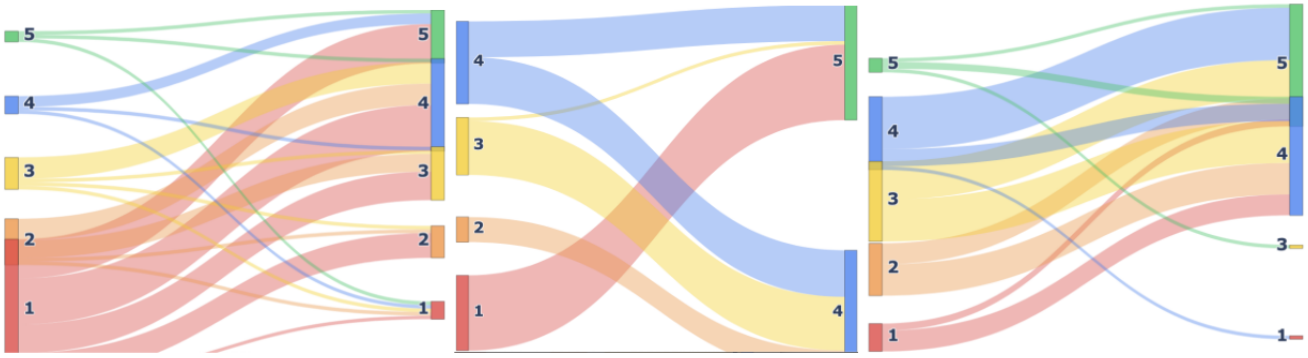


Figure 2: Rating transitions from original resumes to customized resumes. Left: GPT, Middle: GritLM, Right:Llama

automated inference drastically increases the scale of profiling, and facilitates their indefinite retention and secondary use [61].

We also found that different models show varying levels of capability to infer different traits, and they are not consistent with each other. Similar inconsistencies were observed for resume rating, both for the original and customized versions. While part of it could be due to the slightly different training procedures and hyperparameter settings of these models, we note that all three models use the decoder-only transformer architecture as their backbone, and were trained on online data sources. Thus, this inconsistency likely highlight the inherent nondeterministic nature of these models. Further, as prior works show, these models are very sensitive to their hyperparameter values [42], and changing them can drastically change model predictions. We also found that customizing resumes had non-uniform impact: it changed predictions for some traits but not others. In particular, it had a surprising effect on GritLM: it reduced the underestimation for neuroticism, which is counterintuitive as customization typically leads to a more confident style of writing. In sum, including LLMs likely added to the opaqueness and unreliability of algorithmic hiring pipelines. This conclusion aligns with research investigating the reliability of commercial trait prediction services [59].

Finally, we found that customizing resumes increased rating most of the times, as expected; however, surprisingly, they sometimes led to a decrease as well. From the perspective of transparency and reliability, this seems to be a double edged sword. Particularly, many resumes that were ranked the lowest (1) were ranked the highest (5) after customization. This is a big increase, caused without any material change in the core skill set stated in the resume. Thus, LLM models may expedite identifying candidates who "appear" to be a good fit but may not actually poses the necessary skills, while more skilled candidates might get overlooked. On the other side, a lower rank due to customization will only add to the confusion.

5.1 Limitations

Due to resource constraints, we had to use smaller than the state-of-the-art models; larger models might show a higher prediction accuracy. However, both comprehensive benchmarking studies [52, 75] and large-scale surveys [72] show that medium and even small size models perform on par with large models across various domains. We also note that, tweaking the prompts, or creating separate, customized prompts for each model might improve accuracy. But it also demonstrate the brittle nature of these models, where a slightly different prompt or an updated version of the model might drastically change behaviors. Overall, it speaks to the inherent unreliability of predicting complex psychological constructs from resume text. Finally, the accuracy improvement not solve the core privacy problem, which is profiling people which might lead to discrimination and stereotyping.

Another limitation is that our participant pool is not representative of the entire job-seeking population. However, this strengthens, rather than diminishing, the value and implications of our results. This is because a more heterogeneous population will also vary in

the writing style and formatting, and as prior research suggests [59], more variations will likely increase the inconsistency.

6 Conclusions

This study investigated how LLM-based trait inference and job applicant ranking impacts applicants' right to privacy under a harm-centric privacy framework. It further investigated how this practice impacts the reliability and transparency of the hiring process. Findings portray a grim picture: LLMs may contribute to the already opaque recruitment procedure, and harm applicants in multiple ways. This research thus calls for a more cautious approach in deploying automation in hiring.

Ethical Declaration and Consideration

The manuscript was written completely by the authors. To protect participants' privacy, we used locally-hosted models instead of uploading resumes to external services.

Acknowledgments

This research was partially supported by the National Science Foundation under Grant #2350036 and the Air Force Office of Scientific Research (AFOSR) under Award FA9550-24-1-0227. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Air Force Office of Scientific Research or the National Science Foundation.

References

[1] [n. d.]. Bitsandbytes. https://huggingface.co/docs/transformers/en/quantization/bitsandbytes
[2] [n. d.]. California Consumer Privacy Act (CCPA) | State of California - Department of Justice - Office of the Attorney General. https://oag.ca.gov/privacy/ccpa
[3] [n. d.]. CV Harvesting and Fake Job Posting - What You Need to Know. https://inspiredcvs.co.uk/blog-articles/78-cv-harvesting-and-fake-job-posting-what-you-need-to-know
[4] [n. d.]. General Data Protection Regulation (GDPR) – Legal Text. https://gdpr-info.eu/
[5] [n. d.]. Job Seekers Using Dating Apps for Work? Here's Why It's Risky. https://www.forbes.com/sites/kimelsesser/2025/11/17/job-seekers-are-using-dating-apps-to-get-ahead-heres-why-thats-risky/
[6] [n. d.]. LinkedIn. https://www.linkedin.com/feed/
[7] [n. d.]. Qualtrics. www.qualtrics.com
[8] 2025. Crystal. https://www.crystalknows.com/
[9] 2025. The Downside of Relying on AI in Recruiting. https://www.trykondo.com/blog/ai-recruiting-downsides
[10] 2025. Humantic AI. https://humantic.ai
[11] 2025. Meet Hiring Assistant. https://business.linkedin.com/hire/hiring-assistant
[12] 2025. The most powerful AI candidate search. https://juicebox.ai/peoplegpt
[13] 2025. The résumé is dying, and AI is holding the smoking gun. https://arstechnica.com/ai/2025/06/the-resume-is-dying-and-ai-is-holding-the-smoking-gun/
[14] 2025. Your ATS is Burying your Best Hires. https://www.testgorilla.com/ai-resume-scoring/
[15] Evgeni Aizenberg, Matthew J. Dennis, and Jeroen van den Hoven. 2025. Examining the assumptions of AI hiring assessments and their impact on job seekers' autonomy over self-representation. *AI & SOCIETY* 40, 2 (Feb. 2025), 919–927. doi:10.1007/s00146-023-01783-1
[16] Iman Muftah Albakkosh. 2024. Using Fleiss' Kappa Coefficient to Measure the Intra and Inter-Rater Reliability of Three AI Software Programs in the Assessment of EFL Learners' Story Writing. *International Journal of Educational Sciences and Arts* 3, 1 (2024), 69–96. doi:10.59992/IJESA.2024.v3n1p4
[17] Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024. Do Large Language Models Discriminate in Hiring Decisions on the Basis of Race, Ethnicity, and Gender? (2024). doi:/10.48550/arXiv.2406.10486
[18] Lena Armstrong, Jayne Everson, and Amy J. Ko. 2023. Navigating a Black Box: Students' Experiences and Perceptions of Automated Hiring. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 1 (ICER '23, Vol. 1)*. Association for Computing Machinery, New York, NY, USA, 148–158. doi:10.1145/3568813.3600123
[19] Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring. (2024). doi:10.48550/arXiv.2405.04412
[20] Lena Armstrong and Danaé Metaxa. 2025. Navigating Automated Hiring: Perceptions, Strategy Use, and Outcomes Among Young Job Seekers. *Proc. ACM Hum.-Comput. Interact.* 9, 2 (May 2025), CSCW140:1–CSCW140:26. doi:10.1145/3711038
[21] Alejandro Peña Ignacio Serna Aythami Morales, Julian Fierrez. 2020. Bias in Multimodal AI: Testbed for Fair Automatic Recruitment. *IEEE CVPR Workshop on Fair, Data Efficient and Trusted Computer Vision*, (2020). doi:10.48550/arXiv.2004.07173
[22] Boris Bankov, Silvia Parusheva, Olga Marinova, Petya Strashimirova, and Denitsa Petkova. 2024. *Academic Profile Management: Benchmarking DeepSeek-R1 for Publication and Citation Data*. University of Economics – Varna. https://doi.org/10.1007/978-3-032-05607-8_11
[23] Gerhard Blickle, James A. Meurs, Andreas Wihler, Christian Ewen, Andrea Plies, and Susann Günther. 2013. The interactive effects of conscientiousness, openness to experience, and political skill on job performance in complex jobs: The importance of context. *Journal of Organizational Behavior* 34, 8 (2013), 1145–1164. doi:10.1002/job.1843 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/job.1843.
[24] Douglas G Bonett and Thomas A Wright. 2015. Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior* 36, 1 (2015), 3–15. doi:10.1002/job.1960
[25] Brandon M. Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K. D'Mello. 2021. Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews. *Proceedings of the 2021 International Conference on Multimodal Interaction* (2021). doi:10.1145/3462244.3479897
[26] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2025. Unleashing the potential of prompt engineering for large language models. (2025). doi:10.48550/arXiv.2310.14735
[27] Zhisheng Chen. 2023. Ethics and discrimination in artificial intelligence-enabled recruitment practices. (2023). doi:10.1057/s41599-023-02079-x
[28] Danielle Keats Citron and Daniel J Solove. 2022. Privacy harms. *BUL Rev.* 102 (2022), 793.
[29] Kate Crawford and Jason Schultz. 2014. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.* 55 (2014), 93.
[30] Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (Sept. 1951), 297–334. doi:10.1007/BF02310555
[31] Petru Lucian Curşeu, Remus Ilies, Delia Vîrgă, Laurenţiu Maricuţoiu, and Florin A. Sava. 2019. Personality characteristics that are valued in teams: Not always "more is better"? *International Journal of Psychology* 54, 5 (2019), 638–649. doi:10.1002/ijop.12511 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijop.12511.
[32] DeepSeek-AI , Daya Guo, Dejian Yang, and others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. (2025). doi:10.48550/arXiv.2501.12948
[33] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. (2022). doi:10.48550/arXiv.2208.07339
[34] Sri Harsha Gajavalli, Junichi Koizumi, and Rakibul Hasan. 2026. What's Privacy Good for? Measuring Privacy as a Shield from Harms due to Personal Data Use. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. doi:10.1145/3772318.3791102
[35] Kate Glazko, Yusuf Mohammed, Ben Kosa, Venkatesh Potluri, and Jennifer Mankoff. 2024. Identifying and Improving Disability Bias in GPT-Based Resume Screening. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 687–700. doi:10.1145/3630106.3658933
[36] Mihai Golovatenco. 2025. 50+ AI in Job Interview Statistics For 2026. (2025). https://www.index.dev/blog/ai-job-interview-statistics
[37] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32. Number: 1.
[38] Eric Grunenberg, Heinrich Peters, Matt J. Francis, Mitja D. Back, and Sandra C. Matz. 2024. Machine learning in recruiting: predicting personality from CVs and short text responses. (2024). doi:10.3389/frsps.2023.1290295
[39] Christopher Harris. 2023. Mitigating Age Biases in Resume Screening AI Models. (2023). doi:10.32473/flairs.36.133236
[40] Shazia Hassan, Naveed Akhtar, and Ayse Kucuk Yilmaz. [n. d.]. Impact of the Conscientiousness as Personality Trait on both Job and Organizational Performance. 1 ([n. d.]).
[41] Changyang He, Yue Deng, Alessandro Fabris, Bo Li, and Asia Biega. 2025. Developing a Fair Online Recruitment Framework Based on Job-seekers' Fairness Concerns. doi:10.48550/arXiv.2501.14110 arXiv:2501.14110 [cs].

[42] Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. 2024. Eliciting Personality Traits in Large Language Models. (2024). doi:10.48550/arXiv.2402.08341
[43] Yu Ji, Wen Wu, Hong Zheng, Yi Hu, Xi Chen, and Liang He. 2023. Is ChatGPT aGood Personality Recognizer?A Preliminary Study. (2023). doi:10.48550/arXiv.2307.03952
[44] Michal Jirásek and František Sudzina. 2020. Big Five Personality Traits and Creativity. (2020). doi:10.12776/QIP.V24I3.1509
[45] James C. Kaufman, Tessy T. Pumaccahua, and Ryan E. Holt. 2013. Personality and creativity in realistic, investigative, artistic, social, and enterprising college majors. *Personality and Individual Differences* 54, 8 (June 2013), 913–917. doi:10.1016/j.paid.2013.01.013
[46] Alfred Kobsa, Sameer Patil, and Bertolt Meyer. 2012. Privacy in instant messaging: an impression management model. *Behaviour & Information Technology* 31, 4 (April 2012), 355–370. doi:10.1080/01449291003611326 _eprint: https://doi.org/10.1080/01449291003611326.
[47] Tarald O Kvålseth. 2015. Measurement of Interobserver Disagreement: Correction of Cohen's Kappa for Negative Values. *Journal of Probability and Statistics* 2015 (2015), 1–8. doi:10.1155/2015/751803
[48] Mark R Leary and Robin M Kowalski. 1990. Impression management: A literature review and two-component model. *Psychological bulletin* 107, 1 (1990), 34.
[49] Gerald S Leventhal. 1980. What should be done with equity theory? New approaches to the study of fairness in social relationships. In *Social exchange: Advances in theory and research*. Springer, 27–55.
[50] John W. Lounsbury, Ryan M. Smith, Jacob J. Levy, Frederick T. Leong, and Lucy W. Gibson. 2009. Personality Characteristics of Business Majors as Defined by the Big Five and Narrow Personality Traits. *Journal of Education for Business* 84, 4 (March 2009), 200–205. doi:10.3200/JOEB.84.4.200-205 _eprint: https://doi.org/10.3200/JOEB.84.4.200-205.
[51] Yannick l'Horty, Naomie Mahmoudi, Pascale Petit, and François-Charles Wolff. 2022. Is disability more discriminatory in hiring than ethnicity, address or gender? Evidence from a multi-criteria correspondence experiment. (2022). doi:10.1016/j.socscimed.2022.114990
[52] Mohammad Meymani, Hamed Jelodar, Parisa Hamedi, Roozbeh Razavi-Far, and Ali A. Ghorbani. 2025. Can Small GenAI Language Models Rival Large Language Models in Understanding Application Behavior? doi:10.48550/arXiv.2511.12576 arXiv:2511.12576 [cs].
[53] Dena F. Mujtaba and Nihar R. Mahapatra. 2025. Fairness in AI-Driven Recruitment: Challenges, Metrics, Methods, and Future Directions. (May 2025). doi:10.48550/arXiv.2405.19699
[54] Rainer Mühlhoff. 2023. Predictive privacy: Collective data protection in the context of artificial intelligence and big data. *Big Data & Society* 10, 1 (Jan. 2023), 20539517231166886. doi:10.1177/20539517231166886
[55] Amin Naemi and Ali Sahafi. 2025. Benchmarking Large Language Models for MIMIC-IV Clinical Note Summarization. *Journal of Healthcare Informatics Research* (2025). doi:10.1007/s41666-025-00221-9
[56] Alejandra Gómez Ortega. 2025. Sensitive data donation in practice: unforeseen challenges and lessons learned. (2025). doi:10.3389/fhumd.2025.1404855
[57] Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. [n. d.]. Is Temperature the Creativity Parameter of Large Language Models? ([n. d.]).
[58] Marco Piastra and Patrizia Catellani. 2025. On the emergent capabilities of ChatGPT 4 to estimate personality traits. (2025). doi:10.3389/frai.2025.1484260
[59] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. 2022. An external stability audit framework to test the validity of personality prediction in AI hiring. *Data Mining and Knowledge Discovery* 36, 6 (Nov. 2022), 2153–2193. doi:10.1007/s10618-022-00861-0
[60] Sebastiaan Rothmann and Elize Coetzer. 2003. The Big Five Personality Dimensions and Job Performance. (2003). doi:10.4102/sajip.v29i1.88
[61] Daniel J. Solove. 2005. A Taxonomy of Privacy. *University of Pennsylvania Law Review* 154, 3 (2005), 477–564. https://heinonline.org/HOL/P?h=hein.journals/pnlr154&i=491
[62] Daniel J. Solove. 2024. Data Is What Data Does: Regulating Based on Harm and Risk Instead of Sensitive Data. doi:10.2139/ssrn.4322198
[63] Alicia Solow-Niederman. 2022. Information Privacy and the Inference Economy. (2022). https://scholarlycommons.law.northwestern.edu/nulr/vol117/iss2/1/
[64] Christopher J. Soto and Oliver P. John. 2017. Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality* 68 (June 2017), 69–81. doi:10.1016/j.jrp.2017.02.004
[65] Keith Spencer. 2025. AI Is Flooding Hiring—But 62% of Employers Reject Resumes That Lack a Personal Touch. https://www.resume-now.com/job-resources/careers/ai-applicant-report.
[66] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Beyond memorization: violating privacy via inference with large language models. (2024).
[67] Yixuan Tang, Yi Yang, and Ahmed Abbasi. 2025. PersonaFuse: A Personality Activation-Driven Framework for Enhancing Human-LLM Interactions. (2025). https://arxiv.org/html/2509.07370v1
[68] u/PromptGeniusUser. 2025. 10 Prompts I Used to Fix My Resume. Reddit, r/ChatGPTPromptGenius. https://www.reddit.com/r/ChatGPTPromptGenius/comments/1ks1mfm/10_prompts_i_used_to_fix_my_resume/ Accessed: 2025-11-03.
[69] 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. (2023). doi:10.48550/arXiv.2310.05135
[70] Jessica Vitak. 2015. Balancing privacy concerns and impression management strategies on Facebook. In *Symposium on usable privacy and security (SOUPS)*. 22–24.
[71] Sandra Wachter and Brent Mittelstadt. 2019. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. (2019), 494.
[72] Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, TzuHao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. 2025. A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness. *ACM Trans. Intell. Syst. Technol.* 16, 6 (Nov. 2025), 145:1–145:87. doi:10.1145/3768165
[73] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. doi:10.1145/3531146.3533088
[74] Kyra Wilson and Aylin Caliskan. 2024. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 1578–1590. doi:10.1609/aies.v7i1.31748 Number: 1.
[75] Mohammad Amin Zadenoori, Vincenzo De Martino, Jacek Dabrowski, Xavier Franch, and Alessio Ferrari. 2025. Does Model Size Matter? A Comparison of Small and Large Language Models for Requirements Classification. doi:10.48550/arXiv.2510.21443 arXiv:2510.21443 [cs].
[76] Tianyi Zhang, Antonis Koutsoumpis, Janneke K. Oostrom, Djurre Holtrop, Sina Ghassemi, and Reinout E. de Vries. 2024. Can Large Language Models Assess Personality From Asynchronous Video Interviews? A Comprehensive Evaluation of Validity, Reliability, Fairness, and Rating Patterns. (2024). https://ieeexplore.ieee.org/document/10463124/references#references
[77] Jianfeng Zhu, Ruoming Jin, and Karin G. Coifman. 2025. Can LLMs Infer Personality from Real World Conversations? (2025). doi:10.48550/arXiv.2507.14355

A Appendix

A.1 List of Job Titles

Office manager, financial aid counselor, Buyer, Principal engineer, Manager, analytics, Speech Therapist, Software Engineer, Data Analyst, Project Manager, Machine Learning Engineer, Writer, Personal Injury Attorney, Customer Service Specialist, R&D Engineer, Attorney, Real estate attorney, Project Coordinator, Product Manager, Full Stack Developer, Business Development Manager, Customer Success Manager, Workstation Specialist Advanced, Manager, Political Analyst, Social Media Creator, Information Technology, Marketing Specialist, Manager, Sales Coordinator, Pathologist, Systems Engineer, Political Analyst, QA Specialist, Banker, Paralegal, Work From Home, Senior Manager, Professor, Technology Manager, Commercial Lending System Administrator, Data engineer, Systems Architect, Firmware Engineer, Childcare, Data Analyst, Health Administration, Audio Director, Tech Specialist, Senior Software Developer, Actuary, User Researcher, Financial planning, Consultant, Mechanical Engineer, Key Account, Accounting Manager, Financial Planning Officer, Nursing Faculty, QA Tester, Systems Engineer