"ChatGPT in the Hiring Chair: Evaluating AI-Human Agreement & Divergence in Automated Video Interview Assessment Ratings"

Major Research Project (MRP) for MA Degree

Eliana Brereton

[1080873]

April 2025

Table of Contents

Abstract

As large language models (LLMs) like ChatGPT are increasingly integrated into hiring technologies, critical questions arise about their reliability, fairness, and ethical legitimacy. This study evaluated the extent to which ChatGPT-generated interview ratings aligned with human evaluator judgments from the same dataset, in the context of asynchronous video interviews (AVIs). Using a dataset of 183 transcribed candidate responses from a prior study (Patel et. al., 2025), ChatGPT-4o was prompted to assess responses under two experimental conditions; one with job context included and one without. The model rated transcriptions using a standardized Behaviorally Anchored Rating Scale (BARS) that the human evaluators adhered to as well.

Results showed a moderate-to-strong alignment between ChatGPT and human ratings, with an intraclass correlation of 0.94 and a Pearson correlation of 0.56. However, ChatGPT consistently assigned higher ratings than human raters (Cohen's d = -1.08), and demographic subgroup analyses revealed meaningful disparities in ratings, particularly for Black and middle-aged participants. These divergences were *not* observed in human evaluations, raising concerns about fairness and bias in AI-driven assessments.

This study demonstrates that while ChatGPT exhibits potential as a supplementary evaluation tool, its use must be accompanied by careful calibration, human oversight, and explicit ethical auditing practices. These findings also underscore the importance of evaluating not only the technical performance of LLMs, but also the ethical implications of delegating evaluative authority to AI systems, in high-stakes decision-making such as personnel selection.

Introduction

The integration of Artificial Intelligence (AI) into human resource (HR) practices, particularly personnel selection, represents a significant shift from traditional, human-centered decision-making to automated processes (Li et al., 2021; Tambe et al., 2019; Yam & Skorburg, 2021). Large language models (LLMs) such as ChatGPT, coupled with automated video interview (AVI) technologies, exemplify the increasing reliance on algorithmic evaluation, which promises efficiency, scalability, and potential reductions in human biases (Hickman et al., 2022; Hickman, Tay, & Woo, 2024). However, despite their growing prevalence, the efficacy, fairness, and ethical implications of these technologies remain contested (Dennis & Aizenberg, 2022; Tippins et al., 2021).

This study investigates ChatGPT's reliability in evaluating candidate responses in automated video interviews, comparing AI-generated ratings to human evaluations. Specifically, it examines how well AI ratings align with expert human judgment under varying contextual conditions and explores potential demographic biases across gender, race, and age subgroups. Previous research has identified significant issues regarding the transparency, accountability, and fairness of AI-driven hiring tools (Landers & Behrend, 2023; Selbst & Barocas, 2016), raising concerns about their impact on procedural justice and evaluative legitimacy.

Evaluative authority, which is traditionally vested in human evaluators, is foundational to HR decisions. It implicates moral responsibility, epistemic trust, and procedural legitimacy (Zagzebski, 2012). The potential delegation of this authority to AI models introduces a "responsibility gap," where accountability pathways for biased or erroneous decisions become unclear (Matthias, 2004; Floridi, 2016). Moreover, AI systems exhibit epistemic opacity; the

inherent difficulty in understanding how they derive decisions (commonly referred to as the "black-box" problem), which further complicates questions of accountability and moral authority (Burrell, 2016).

Given these profound ethical and philosophical challenges, the current study addresses critical research gaps concerning text-based AI assessments. Specifically, it explores the degree of alignment between AI and human evaluations, and examines potential biases in ratings given that may disproportionately affect certain demographic groups (Dennis & Aizenberg, 2022; Hickman et al., 2024).

Accordingly, this project addresses three primary guiding research questions:

1. (*RQ1*) Can LLMs reliably evaluate interview responses?
2. (*RQ2*) To what extent do ChatGPT's ratings align with human assessments?
3. (*RQ3*) How does this alignment vary across demographic subgroups?

Several more exploratory questions also guide the discussion: What does reliance on AI-generated ratings imply for fairness and objectivity in hiring? What ethical tensions arise from transferring human evaluative authority to non-human systems, particularly when these models lack true "understanding" / moral reasoning? The exploratory research questions of this study are in place to critically assess the broader implications of delegating evaluative authority to AI systems in high-stakes contexts, such as personnel selection.

By investigating these issues, this study contributes significantly to understanding AI's role in personnel selection contexts, hoping to inform best practices and development of ethical guidelines for implementing AI-driven personnel assessment tools. This study also assists in

evaluating LLMs' capability to standardize hiring assessments, while identifying potential fairness and bias risks. Finally, this study offers practical insights for HR professionals, AI developers, and policymakers on how to utilize AI-driven hiring tools in personnel selection in an ethical manner.

Literature Review

AI in Personnel Selection: Opportunities and Challenges

AI-driven tools such as AVIs, applicant tracking systems, and other automated assessment tools are increasingly being adopted and utilized by HR professionals to streamline hiring and personnel selection processes (Tambe et al., 2019; Yam & Skorburg, 2021). These technologies promise substantial operational advantages, including expanded candidate pools, cost reduction, and the potential minimization of explicit human biases (Li et al., 2021).

Despite this, AI's integration into personnel selection has raised serious concerns about transparency, fairness, and, ironically, the potential *amplification* of societal biases (Barocas & Selbst, 2016, 2018; Dennis & Aizenberg, 2022). Algorithmic opacity particularly undermines the communicative transparency essential for moral accountability, eroding stakeholders' trust, questioning evaluative legitimacy, and raising profound concerns about the displacement of human evaluative authority (Landers & Behrend, 2023).

Automated Video Interviews (AVIs) and AI-Human Alignment

AVIs are capable of utilizing multimodal data, potentially assessing candidates through verbal, paraverbal, and nonverbal cues. Despite demonstrating predictive validity regarding job performance, significant fairness concerns persist (Hickman et al., 2024; Tippins et al., 2021). For instance, AI assessments often struggle to generalize across diverse demographic groups, exacerbating biases against non-native speakers and racial minorities (Zhang et al., 2024; Hickman, Tay, & Woo, 2024). Prior studies suggest moderate-to-strong correlations between

human and AI ratings but highlight consistent divergences particularly along demographic lines (Hickman et al., 2022; Hickman et al., 2024).

Bias and Fairness Considerations in AI Evaluations

Bias in AI systems commonly arises from biased training datasets and can inadvertently perpetuate societal inequities that bring harm to those in minority and/or societally disadvantaged groups (Barocas & Selbst, 2016; Bommasani et al., 2021). Researchers emphasized the critical need for robust auditing frameworks to try to mitigate these biases (Landers & Behrend, 2023). The fairness of AI-driven hiring assessments, particularly those based on language and speech patterns, remains contentious at best, with evidence showing systematic disadvantages for certain demographic groups (Liff et al., 2024; Zhang et al., 2024). This raises concerns about the distribution of responsibility (and accountability) in AI-mediated personnel selection decisions (Matthias, 2004; Floridi, 2016).

Large Language Models (LLMs) in Hiring

Text-based LLMs, such as ChatGPT, present unique advantages by systematically analyzing linguistic content without relying on nonverbal cues (Hickman et al., 2024). However, the reliability and fairness of these tools remains under scrutiny. Recent studies suggest that while LLMs can reliably mirror certain human evaluations, significant disparities arise in terms of demographic fairness (Zhang et al., 2024; Liff et al., 2024).

Critics have cautioned that the use of automation in evaluative contexts may erode the moral and epistemic grounding traditionally associated with human judgment (Burrell, 2016;

Vallor, 2016). While these systems may achieve consistency in surface-level outputs, their lack of moral reasoning and contextual understanding limits their legitimacy as evaluative agents.

This study addresses this critical gap by focusing specifically on textual evaluations by LLMs, providing a comparative 1:1 analysis of human and AI ratings, within a structured interview context.

These empirical and conceptual gaps warrant further investigation into how LLMs operate in high-stakes contexts like hiring, and under what conditions their use can be justified. By addressing these gaps, this study aims to illuminate the conditions in which LLMs like ChatGPT could effectively complement human judgment, and highlight where significant ethical and practical adjustments are necessary to preserve fairness and legitimacy in these automated assessments.

Methodology

Study Design

This study employed a cross-sectional, observational design, to compare AI-generated interview ratings with human ratings in the context of asynchronous video interviews (AVIs). The primary aim was to evaluate the *extent of agreement* between ChatGPT-generated ratings, and those provided by trained human evaluators. The study included a between-subjects comparison of AI-generated and human-generated scores, across two experimental conditions that varied the amount of contextual information provided to the AI model (ChatGPT-4o).

Participants and Data Sources

Participant interview data was drawn from a pre-existing dataset of 183 participants collected as part of a prior study by Patel et al. (2025). In Patel et al.'s 2025 study, each participant responded to three structured interview questions, under controlled AVI conditions. Audio recordings were transcribed using Whisper, an open-source automatic speech recognition system that is validated for use in psychological research (Spiller et al., 2023). The transcripts were manually reviewed and cleaned for accuracy before use in both human and AI evaluation pipelines (Patel et al., 2025).

Participants were demographically diverse with respect to gender, race, and age. Ratings were produced by both trained human research assistants, and OpenAI's ChatGPT model (GPT-4o). The human evaluators scored responses using a standardized five-point Behaviorally Anchored Rating Scale (BARS) developed by Patel et al. (2025). ChatGPT also was instructed

(via system message) to apply the same BARS criteria, generating numerical ratings for each candidate question response.

Rating Procedure

This study involved two experimental conditions that varied the amount of contextual information provided to ChatGPT before it rated candidate interview responses. In both conditions, the AI model received the transcribed answers to three structured interview questions per participant, and was prompted to score each response using a standardized five-point Behaviorally Anchored Rating Scale (BARS).

In the *Full Context Condition*, the prompt included the original job description, allowing the model to consider the role-specific expectations while rating each response. In the *Reduced Context Condition*, the prompt included only the BARS criteria without any job-specific information, leaving the model to assess responses based solely on generalized performance standards. Each participant's three responses were rated independently, in both conditions, producing three individual ratings per condition. These were later averaged to create a single composite interview performance score (from ChatGPT) for each participant.

All prompts were executed using the OpenAI API with the GPT-4o model, using Python. To ensure consistency, the wording of the prompt and instructions remained constant across all trials within each condition, with only the presence or absence of job context varying. The system message was taken mostly from Patel et al.,'s 2025 supplemental procedures, with only redundant information removed/changed (anything to do with video or audio information). Both experimental runs followed the same procedural steps, allowing for a controlled comparison of how contextual information influenced the model's evaluations.

Dataset Construction

Following rating completion, the data was organized to facilitate direct comparison between ChatGPT and the human evaluators. AI-generated scores for each participant were averaged across their three interview responses to form a composite performance score for each participant. Human-generated ratings were retrieved from the pre-existing dataset collected in the original Patel et al., (2025) study and similarly organized to reflect each participant's overall interview performance.

Participant identifying ID numbers were standardized to ensure accurate matching between AI- and human-scored records. Both datasets were merged into a unified file using this common identifier, with any missing values clearly labeled for transparency. Participants with incomplete ratings from either the AI, or any human rater, were explicitly excluded from any paired analyses (which were conducted via R and Python).

This process resulted in a clean, aligned dataset, with ChatGPT-generated scores and with human-assigned scores, ready for statistical analysis. All data processing was conducted using Python, and statistical tests were then performed in R.

Statistical Analyses

RQ1: Can ChatGPT Reliably Evaluate Candidate Interview Transcripts?

To assess the internal consistency and reliability of ChatGPT's scoring, a verification mechanism was implemented within the rating pipeline. After providing an initial score, the model was prompted in its system to re-read the candidate response for a *second* time, along with its own initially given rating, to either confirm ("agree") or revise ("disagree") its original

judgment. Instances of self-disagreement were logged into a separate *.txt* file for manual review. This self-check procedure served as a quality control mechanism, conceptually analogous to a secondary human rating, or adjudication step.

RQ2: To What Extent Do ChatGPT's Ratings Align With Human Evaluations?

To address the research question, "To what extent do ChatGPT's ratings align with human evaluator ratings of the same data?", I performed a series of analyses on the aligned dataset.

1. **Paired-samples t-tests** were conducted to determine whether there was a systematic difference in the mean scores assigned by ChatGPT and human evaluators.
2. **Intraclass Correlation Coefficients (ICC)** were calculated using a two-way mixed-effects model (absolute agreement, average measures; ICC1k, ICC3k) to assess overall rating consistency between AI and human evaluators, following best practices for inter-rater reliability in psychological research (Koo & Li, 2016; Harwood, 2024).
3. **Pearson correlation coefficients** were computed to quantify the strength of the linear relationship between AI and human ratings. Pairwise deletion was used to handle missing data.

To supplement these analyses, individual correlations were also calculated, using R, between ChatGPT and each of the human raters. This allowed for inspection of variation in alignment across individual human evaluators. Results were visualized (via R's corrplot package) using an averaged pairwise correlation matrix, and a scatterplot to illustrate patterns of convergence and divergence.

RQ3: How does AI-human agreement vary across demographic subgroups (gender, race, age)?

To assess whether the alignment between ChatGPT 4o and human ratings varied across demographic subgroups, I implemented a correlation comparison approach similar to Hickman et al. (2024).

First, I dummy-coded self-reported participant demographic variables (from Patel et. al.,'s 2025 study) representing gender (e.g., gender_female, gender_agender, gender_fluid), race/ethnicity (e.g., ethnicity_black, ethnicity_southasian, ethnicity_latin, etc.), and age group (e.g., age_under_25, age_35_44, age_45_54, age_55_plus). Each dummy variable was then binary-coded (1 = belongs to subgroup; 0 = does not belong). For instance, for the variable gender_female, participants who self-identified as female were assigned a '1,' whereas all other participants were assigned a '0.'

Secondly, for each demographic subgroup, I calculated two Pearson correlation coefficients: one between subgroup membership and the average ChatGPT-assigned interview rating, and one between subgroup membership and the average human-assigned interview rating.

Third, I compared these correlations using the cocor R package (Diedenhofen & Musch, 2015), which supports statistical comparison of dependent, overlapping correlations using multiple methods. The primary method of interest was Fisher's *r-to-z*-test for overlapping correlations, specifically the Meng, Rosenthal, and Rubin (1992) procedure, which applies Fisher's *z*-transformation to assess the statistical significance of the difference between two dependent correlations. Additional methods were used to support and contextualize findings, including Williams's t-test (1959), Hotelling's t-test (1940), & Zou's (2007) method for

computing confidence intervals around the difference in correlations ($\Delta r = r$ChatGPT - rHuman).

In addition to the correlation analyses, Welch's t-tests were employed to assess whether average interview ratings differed significantly between any demographic subgroup and its complement when the correlation results indicated a meaningful difference between ChatGPT-assigned and human-assigned ratings. For each subgroup identified in this manner, two separate t-tests were performed: one comparing ChatGPT-based ratings, and one comparing human-based ratings. This approach allowed for the examination of whether systematic differences observed in correlations were reflected in mean-level discrepancies as well.

Results

RQ1: Can ChatGPT Reliably Evaluate Candidate Interview Transcripts?

To assess the reliability of ChatGPT-generated interview ratings, I examined the rate of model self-disagreement between its given evaluations and an acceptable human reference standard. See **Figure 1** for a visualization of model agreement vs. disagreement rates.

In the *Full Context Condition* (where ChatGPT had access to the job description in its system prompt), 41 out of 549 ratings (7.47%) were flagged as disagreement cases. In the *Reduced Context Condition* (without job description included in the system prompt), the number of flagged disagreement ratings increased to 55 out of 549, yielding a disagreement rate of 10.02%. With both of these conditions combined, the overall disagreement rate fell at 8.74%. These disagreement rates fall within the acceptable range that is typically observed (and accepted) among human raters, in structured evaluation settings such as personnel selection and psychological research (McHugh, 2012).

Prior research on inter-rater reliability highlights that even trained human evaluators rarely achieve *perfect* agreement, with acceptable thresholds often considered at around 80% agreement, or Cohen's kappa values indicating substantial reliability (McHugh, 2012). The relatively low model disagreement rate in this study reinforces the reliability and credibility of this automated system. The flagged disagreement cases served as a quality-assurance measure, analogous to the review process used in human inter-rater reliability studies (McHugh, 2012).

Taken together, these results suggest that ChatGPT *could* function as a reliable evaluator of interview responses, with its performance aligning closely with established human inter-rater

reliability variability. The relatively low disagreement rates indicate this was a stable assessment process. The differences across conditions highlight the role of contextual information in optimizing evaluation accuracy.

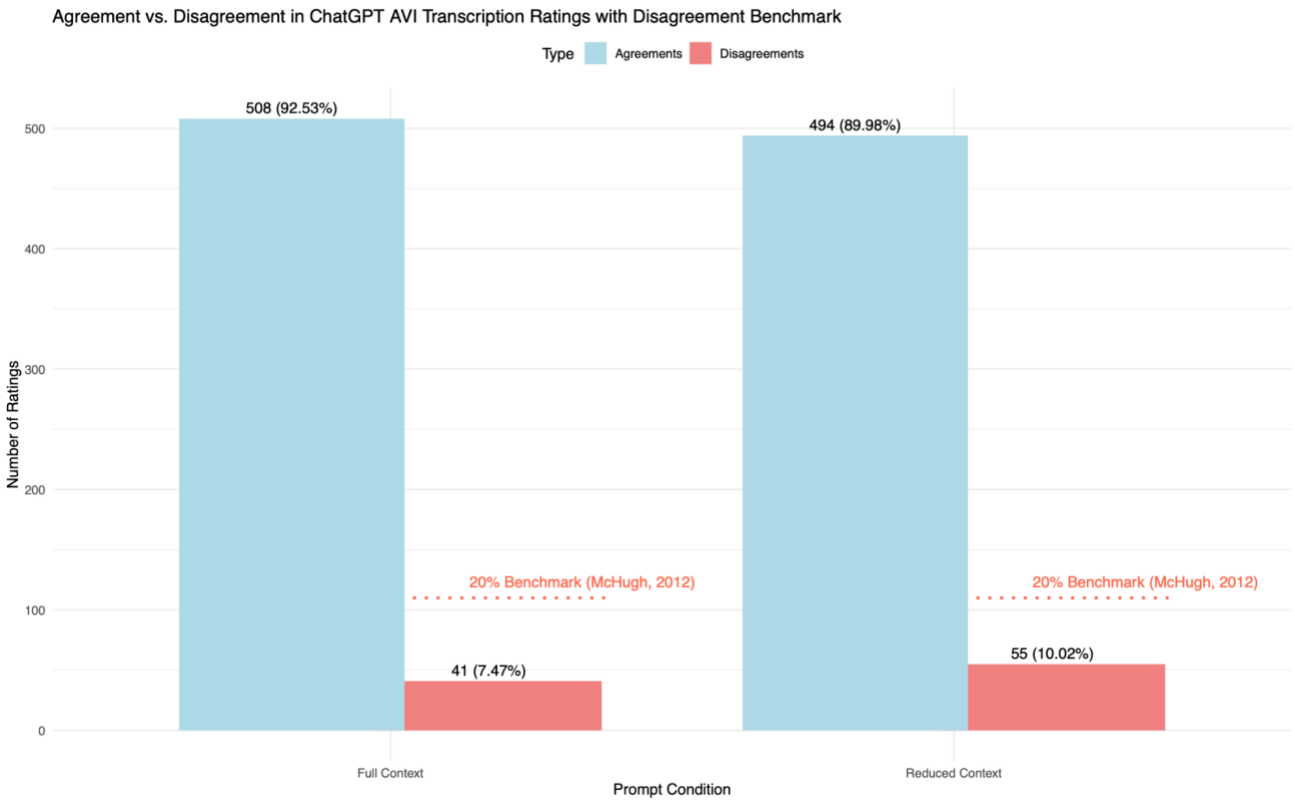

Figure 1: Model Self-Agreement and Self-Disagreement Visualization

Agreement vs. disagreement rates in ChatGPT-generated interview ratings across two prompt conditions (Full Context vs. Reduced Context). Disagreement rates were 7.47% with full context, and 10.02% with reduced context.

RQ2: To What Extent Do ChatGPT's Ratings Align With Human Evaluations?

To assess how closely ChatGPT-generated ratings aligned with previously established human ratings in Automated Video Interview (AVI) performance assessments (Patel et al., 2025), several statistical analyses were conducted, including a paired-samples t-test, effect size calculations, intraclass correlation coefficient (ICC) analysis, and Pearson correlation analysis.

Comparison of Human and ChatGPT Ratings

A paired-samples t-test was conducted to compare human ratings and ChatGPT-generated ratings of assessed interview performance. The test was two-tailed, assessing whether ChatGPT's ratings significantly differed from human raters without assuming a specific direction.

Results revealed that ChatGPT's ratings (M = 3.14, SD = 0.92) were significantly higher than human ratings (M = 2.27, SD = 0.69), t(166) = -14.98, p < .001 (two-tailed). These ratings were based on a five-point Behaviourally Anchored Rating Scale (BARS), where higher values indicate "stronger" assessed interview performance. The mean difference was M_{diff} = -0.88 points (95% CI [-1.00, -0.77]), indicating that ChatGPT systematically assigned higher scores than human raters.

To quantify the effect size, Cohen's d was calculated to assess the magnitude of the difference between human and ChatGPT ratings. The analysis yielded Cohen's d = -1.08, 95% CI [-1.26, -0.90], indicating a large effect size based on Cohen's (1988) benchmarks. The negative sign of Cohen's d arises from the computation method (human minus ChatGPT means), mathematically representing the lower human ratings relative to ChatGPT scores.

These results suggest that the difference between ChatGPT and human ratings is substantial and unlikely to be due to chance. The negative value of Cohen's *d* reflects the direction of the difference, confirming that ChatGPT ratings were consistently higher than human ratings.

A Pearson correlation analysis was also conducted to assess the relationship between human and ChatGPT interview ratings. The results indicated a moderate-to-strong positive correlation, $r = 0.56$, 95% CI [0.45, 0.66], $p < .001$. Following Bosco et al. (2015), this effect size falls within the moderate-to-strong range, suggesting a consistent relationship between human and AI-generated ratings. However, the correlation does not imply perfect agreement, reinforcing the need for careful calibration when incorporating AI-generated ratings into selection procedures. These results underscore the potential value of hybrid rating systems that combine human judgment with calibrated AI outputs to maximize both efficiency and fairness in high-stakes decision-making.

See ***Figure 2*** for a visualization of the distribution of ChatGPT and human ratings.

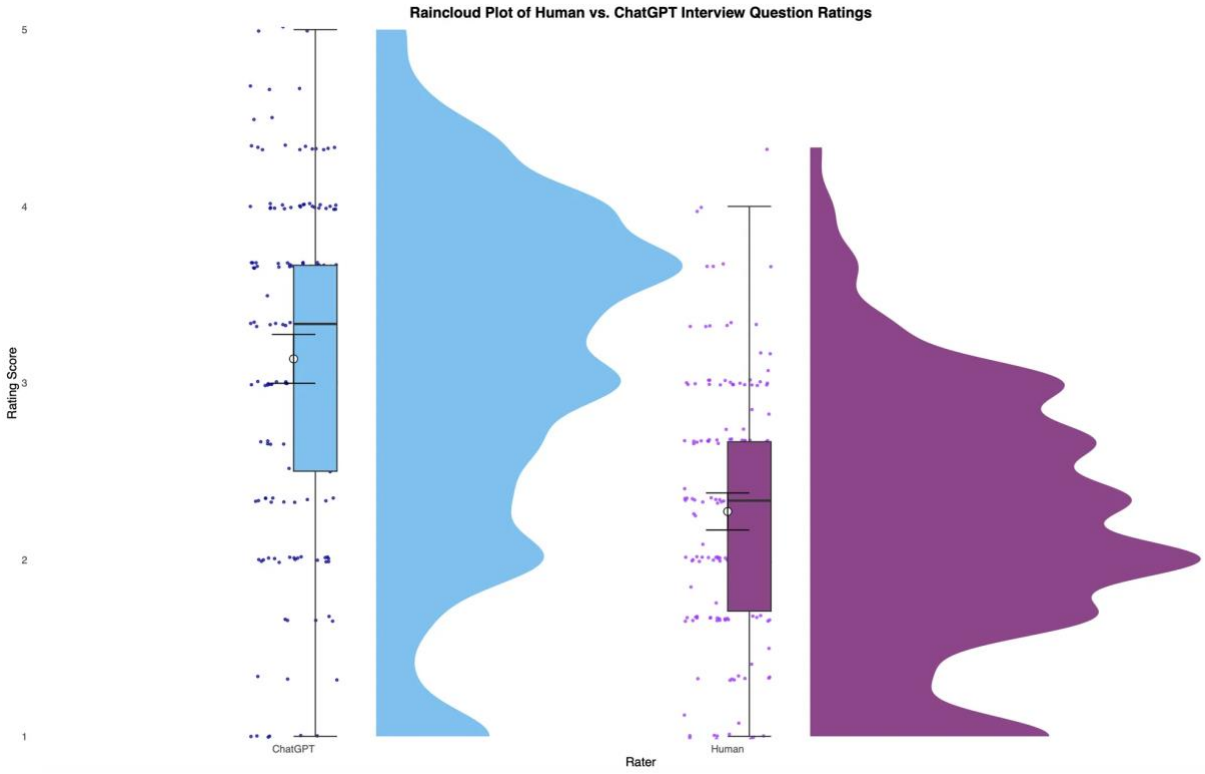

Figure 2: ChatGPT & Human Rating Spread & Distribution

Raincloud/boxplot data visualization displaying & comparing the distribution of ChatGPT and human ratings of assessed interview performance on an anchored BARS scale (1-5). ChatGPT ratings (left, blue) were generally higher than human ratings (right, purple), with a greater spread in their distribution.

The box plots illustrate the median and interquartile range, with the mean of each rater being represented as a white circle. The raincloud density plots (right) represent the distribution of scores for each rater, comparatively. The individual points indicate raw data (individual ratings) from each rater.

Inter-Rater Reliability Between ChatGPT and Human Evaluators

To assess the reliability of ChatGPT's ratings compared to human raters, an Intraclass Correlation Coefficient (ICC) analysis was also conducted. This followed the methodology outlined by Patel et al. (2025), where four trained human raters evaluated a subset of participant interview responses to establish inter-rater reliability. In the original study, inter-rater reliability was assessed using 35 randomly selected participants from the "without probes" condition, resulting in ICC(1,4) = 0.96, which is considered excellent (Koo & Li, 2016; Liljequist, Elfving, & Skavberg Roaldsen, 2019).

To extend this analysis to the current study, ChatGPT was added as a fifth rater to the previously-done analysis, and ICC values were recalculated to assess its agreement with the original human raters. Following best practices for reliability assessment (Koo & Li, 2016), a two-way mixed-effects model with absolute agreement and average measures (ICC3k) was used, as it accounts for the fixed set of raters and evaluates agreement on absolute values rather than just consistency.

The results indicated excellent inter-rater reliability, ICC(3k) = 0.94, $p < .001$, 95% CI [0.90, 0.97], demonstrating strong agreement between human evaluators and ChatGPT. Additionally, ICC(1k) = 0.85, $p < .001$, 95% CI [0.75, 0.92], suggests that even when considering a single rater's agreement within the group, the reliability remains high. While these findings demonstrate high inter-rater reliability, the consistent score inflation by ChatGPT highlights the need to further evaluate the validity of these ratings in future work.

While ChatGPT-generated ratings did significantly differ in magnitude from human evaluations, they nonetheless demonstrate moderate-to-good reliability, and a strong, positive linear relationship with ratings that were assigned by human evaluators to the same participants.

These findings support ChatGPT's potential as a supplemental evaluator in asynchronous video interview scoring; provided systematic calibration efforts are implemented, and manual review is still prioritized.

Correlational Analysis Between Raters and ChatGPT

I also examined the relationship between ChatGPT ratings and individual human ratings from four raters (*Rater1, Rater2, Rater3, and Rater4*). According to guidelines by Bosco et al. (2015), there were moderate-to-strong positive relationships between ChatGPT ratings and individual human ratings, such that as ChatGPT ratings increased, so did human-generated ratings (rs = .40 to .69, ps < .001).

For example, the correlation between ChatGPT and Rater2's ratings was (r = .69, p < .001)**,** while the lowest correlation was with Rater1 (r = .40, p = .003). These values indicate moderate-to-strong alignment between ChatGPT and individual human raters, though ChatGPT's evaluations did not perfectly mirror any single rater's scores.

As expected, inter-rater correlations among the four human evaluators were very high (rs = .81 to .93, ps < .001) (see Patel et al., (2025)), reflecting strong agreement in their scoring of interview responses. For instance, the correlation between Rater2 and Rater3 was (r = .93, p < .001), indicating excellent consistency. Overall, ChatGPT's ratings demonstrated lower (but still substantial) convergence with human evaluations when compared to the near-ceiling inter-rater agreement among humans (see **Figure 3** for a visualization of this data).

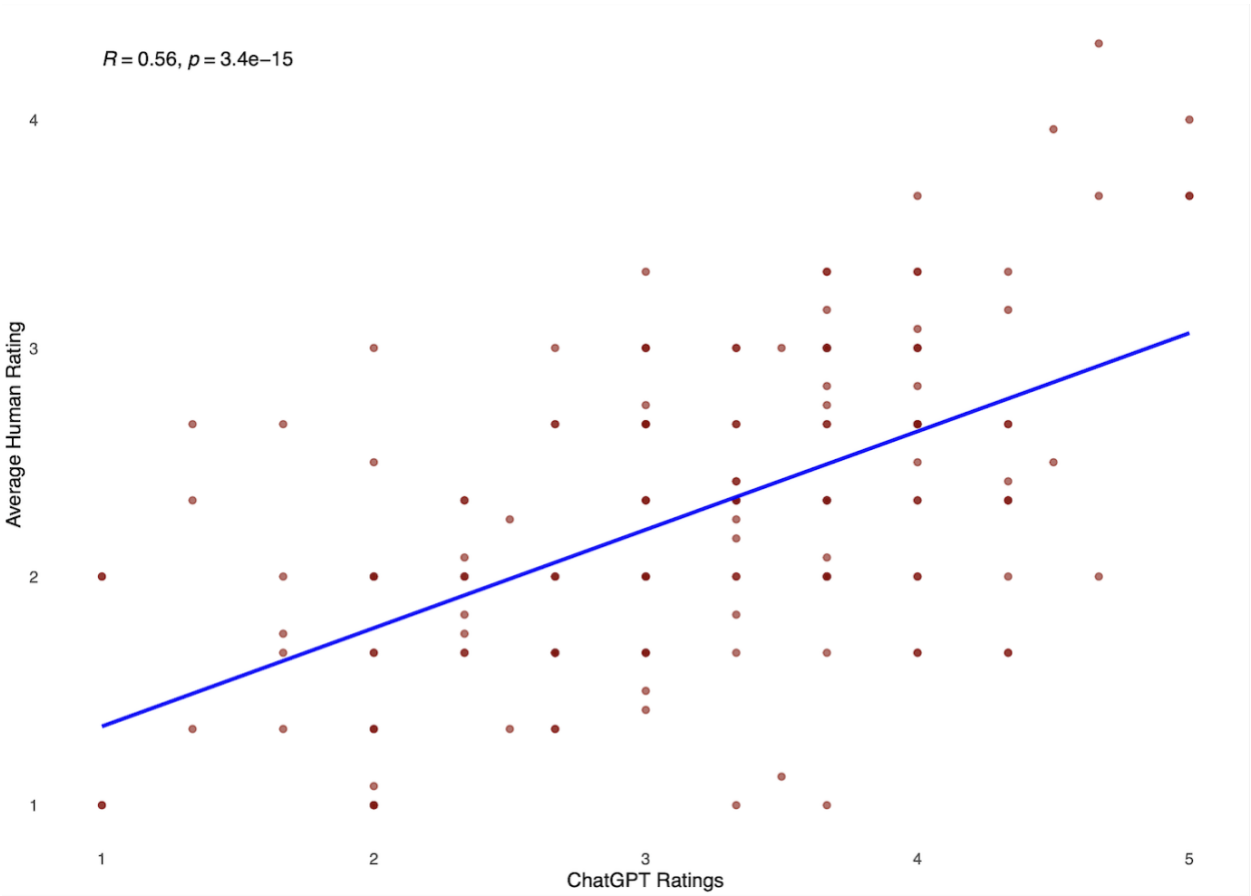

Figure 3: Correlation between ChatGPT & Human Ratings

Scatterplot illustrating the relationship between ChatGPT-generated interview ratings and average human-assigned ratings. Each point represents a single participant, and the blue line depicts the least-squares regression line. The correlation coefficient (r) and corresponding p-value are displayed in the upper left corner. Human ratings reflect the average of four trained raters, who demonstrated high inter-rater reliability and consistent evaluation standards.

RQ3: How does AI-human agreement vary across demographic subgroups (gender, race, age)?

To assess differences in the alignment between ChatGPT 4o-generated and human-assigned interview ratings across various demographic subgroups, correlation comparisons were conducted using Fisher's *r*-to-*z* transformations and associated significance tests. Table 1 (Appendix) provides detailed correlation coefficients (*r*), 95% confidence intervals (CI), and significance values (*p*) for all ChatGPT 4o and human rating relationships to demographic subgroups, along with the magnitude and significance of differences (Δr) between them.

Results revealed notable differences in correlation strength between rating sources for *two* demographic subgroups, with accompanying statistical evidence suggesting these differences are unlikely to be due to chance.

Specifically, membership in the Black subgroup exhibited a moderate negative correlation with ChatGPT-generated ratings, (r = -.27, 95% CI [-.40, -.13], $p < .001$), which aligns with Bosco et al.'s (2015) benchmarks for moderate correlation effect sizes. Human-assigned ratings showed a negligible correlation with Black subgroup membership, (r = -.02, 95% CI [-.17, .14], p = .832). This resulted in a significant correlation difference of (Δr = -.29, 95% CI [-.43, -.16], $p < .001$), suggesting that ChatGPT ratings systematically differed from human ratings with respect to Black participants.

Participants identifying as South Asian showed a positive correlation with ChatGPT ratings (r = .19, 95% CI [.04, .33], p = .013), whereas the correlation with human ratings was smaller (r = .07, 95% CI [-.09, .22], p = .388). Although the difference between these correlations (Δr = .13, 95% CI [-.02, .26], p = .083) points to a possible distinction in effect

sizes, the associated p-value suggests that the evidence for such a difference remains inconclusive.

Among age subgroups, participants aged 45-54 exhibited a small positive correlation with ChatGPT ratings (r = .15, 95% CI [.00, .29], p = .057), and a near-zero correlation with human ratings (r = -.02, 95% CI [-.17, .14], p = .847). The difference between these correlations (Δr = .20, 95% CI [.05, .33], p = .007), suggests that ChatGPT ratings for this age group diverged meaningfully from those of human raters, with statistical evidence indicating that this difference is unlikely to be due to random variation.

For all other subgroup comparisons, including gender (female, agender, genderfluid), the remaining ethnic groups (Latin, Southeast Asian, West Asian, Arab, Indigenous), and age categories (under 25, 35-44, 55+), differences in correlation patterns between ChatGPT and human ratings were generally small, with p-values exceeding the conventional alpha threshold and confidence intervals that included zero, suggesting limited evidence of divergence in these cases. The reported statistics above are visualized in **Figure 4.**

To complement the correlational findings, independent samples *t*-tests were then conducted to examine whether average interview ratings differed between the two groups with significant AI-human rating divergence: Black participants and participants aged 45-54.

Black participants received significantly lower ChatGPT interview ratings (M = 2.68, SD = 0.92, n = 44) compared to non-Black participants (M = 3.27, SD = 0.87, n = 138), $t(59.41) = -3.55$, p = .001, 95% CI [-0.92, -0.26]. In contrast, human raters assigned statistically similar ratings to Black (M = 2.25, SD = 0.64) and non-Black participants (M = 2.28, SD = 0.70), $t(65.68) = -0.22$, p = .82, 95% CI [-0.27, 0.21]. The mean difference

between the two raters in regard to participants in this subgroup was M_{diff} = -0.43, 95% CI [-0.64, -0.21].

For the age 45-54 subgroup, ChatGPT gave significantly higher ratings to participants in this age range (M = 3.42, SD = 0.78, n = 32) than to other participants (M = 3.07, SD = 0.94, n = 150), t(50.93) = 2.15, p = .036, 95% CI [0.02, 0.67]. Again, human raters showed no such difference, rating participants aged 45-54 (M = 2.25, SD = 0.73) and all others (M = 2.28, SD = 0.68) similarly, t(40.87) = -0.19, p = .85, 95% CI [-0.32, 0.27]. The mean difference between the two raters regarding participants in this subgroup was M_{diff} = 1.17, 95% CI [0.08, 2.26].

To contextualize these group-level differences in ratings, standardized mean differences were computed using Cohen's d (Cohen, 1988). For the age 45-54 subgroup, ChatGPT assigned higher ratings to participants in this group compared to others, yielding a small-to-moderate effect size (d = 0.38). Human raters, however, showed minimal difference in their evaluations across the same comparison (d = -0.04).

A similar pattern emerged in the comparison of Black versus non-Black participants. ChatGPT ratings were lower for Black participants, with a moderate-to-large effect size (d = -0.67). In contrast, human ratings for this comparison showed negligible differences (d = -0.04). These standardized effect sizes reinforce the observed divergence between ChatGPT and human raters in how specific demographic groups were evaluated, suggesting that the discrepancies in mean scores are not only statistically detectable but also practically meaningful in the context of algorithmic fairness and bias. The results reported above are visualized in **Figure 5.**

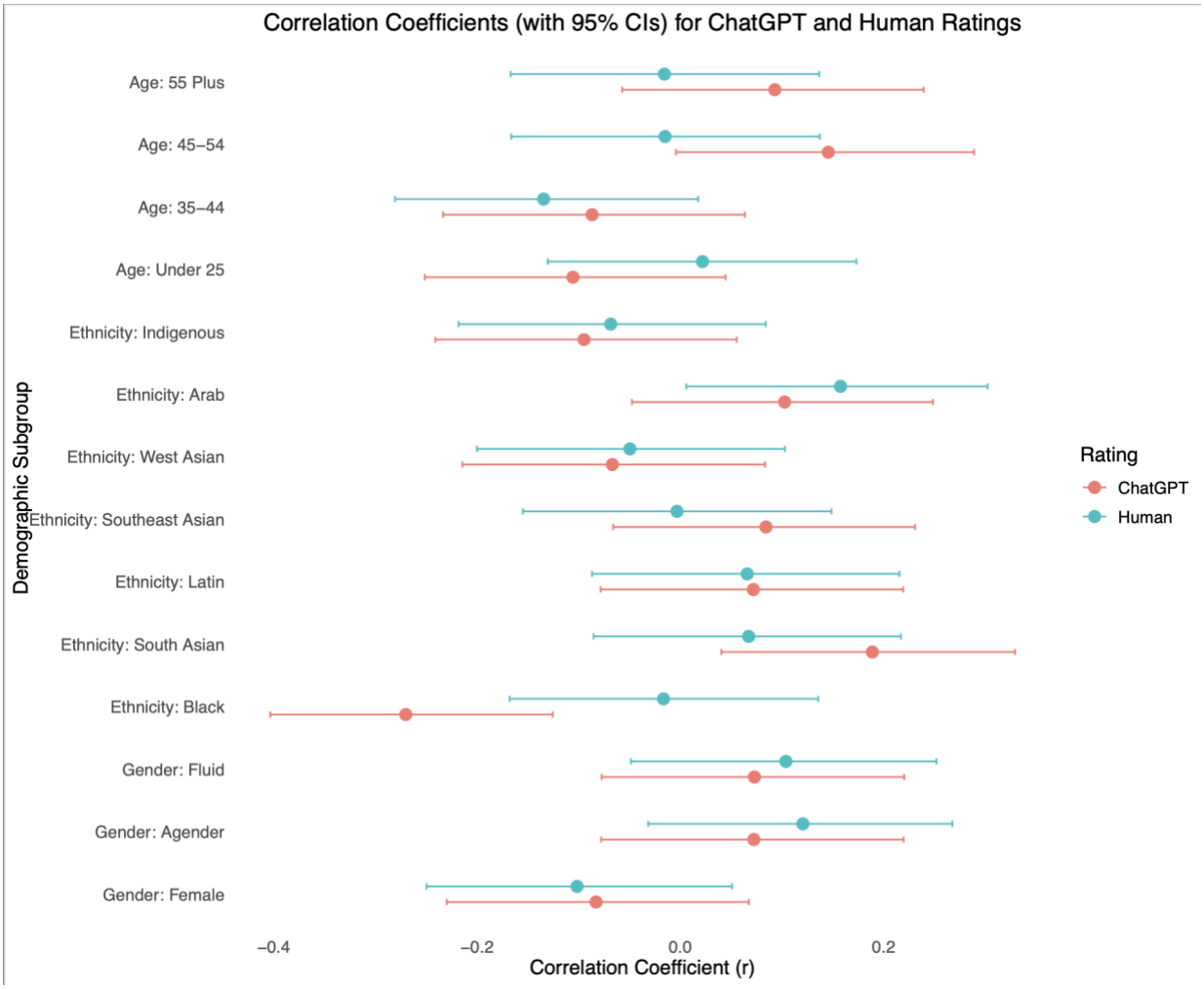

Figure 4: Difference in Correlations (.r) Between ChatGPT and Human Ratings, By Demographic Subgroup

Correlation coefficients between demographic subgroup membership and interview ratings, separately for ChatGPT-generated and human-assigned scores. Each point represents the Pearson correlation (*r*) for a given demographic subgroup, with horizontal bars indicating 95% confidence intervals. Subgroups are ordered vertically by category, and rating sources are color-coded and shown side by side for direct comparison.

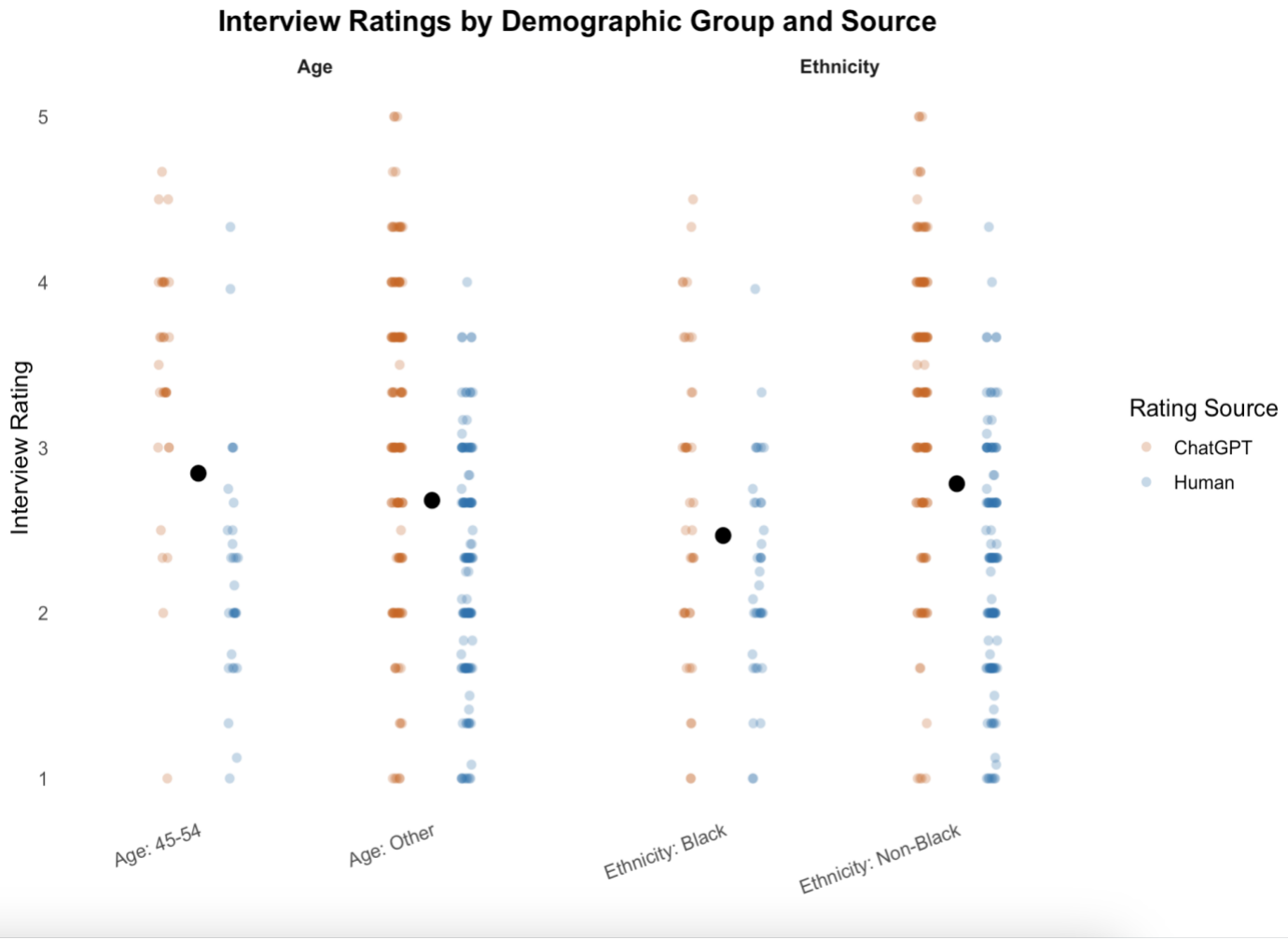

Figure 5: Mean Interview Ratings by Demographic Subgroup Membership & Rating Source

Individual interview ratings assigned by ChatGPT and human raters are shown for participants grouped by demographic category. The plot includes two focal subgroups: ethnicity (Black vs. Non-Black) and age (45-54 vs. Other). Each dot represents an individual participant's rating, jittered for visibility and color-coded by rating source. Black circular markers indicate the group mean for each condition. This visualization highlights both average trends and the distribution of ratings across demographic groups.

Discussion

AI-Human Alignment: Interpreting Agreement and Divergence

The findings of this study indicate that ChatGPT demonstrated a moderate-to-strong degree of alignment with human evaluators when scoring candidate responses in asynchronous video interviews (AVIs), with an intraclass correlation coefficient (ICC3k) of 0.94 and a Pearson correlation (r) of 0.56. These values approach accepted thresholds for inter-rater reliability in human-human evaluations (Koo & Li, 2016; McHugh, 2012). Additionally, ChatGPT's self-disagreement rates (7.47% in the full-context condition and 10.02% in the reduced-context condition) fell within the range typically observed in human rating discrepancies. These results suggest that LLMs can function as effective raters when provided with structured evaluation tools, such as Behaviorally Anchored Rating Scales (BARS), and relevant contextual information. The relative stability of these results across both conditions does indeed suggest that the model's ratings are not excessively sensitive to prompt variability, reinforcing its potential reliability for repeated or large-scale assessment tasks.

However, systematic score inflation, where ChatGPT consistently rated candidates higher than human evaluators (Cohen's d = -1.08), does raise important questions about the model's interpretation of rating criteria. To make ChatGPT-generated ratings truly comparable to those given by human evaluators, this inflation bias should be systematically addressed through calibration or scaling methods. While relative rankings remained stable, the upward bias complicates practical use of LLM evaluators in hiring. In many selection contexts, absolute thresholds, not just relative comparisons, are what determine outcomes.

Moreover, while the inclusion of job context modestly improved consistency (reducing disagreement by 2.5%), both conditions demonstrated relative stability. This supports the reliability of ChatGPT's evaluation process, indicating that the underlying Behavioral Anchored Rating Scale (BARS) framework remains stable even when applied in a more generalized manner. However, the presence of contextual cues does appear to enhance alignment with intended evaluation criteria, reinforcing findings that language models benefit from structured guidance when interpreting complex, domain-specific responses (Ooi et al., 2023). The job description likely serves as a guiding scaffold that reduces ambiguity and helps anchor the model's evaluations in role-relevant criteria.

Without this anchor, the model may rely more heavily on generalized patterns from its training data, leading to a broader range of interpretations and, consequently, a higher disagreement rate. This finding underscores the importance of contextual information in guiding language models' assessments, particularly in scenarios where a nuanced understanding of role-specific expectations is critical.

Philosophically, this highlights the tension between generalizability and specificity in algorithmic assessment. As models rely more on statistical regularities learned during training, the absence of domain-specific framing increases reliance on heuristics rather than principled evaluation. This reinforces the role of prompt design not only as a technical tool but as an epistemic bridge for aligning AI outputs with human evaluative frameworks.

Bias and Fairness: Divergence Across Demographic Subgroups

While aggregate agreement between AI and human raters was generally strong, subgroup analyses revealed troubling disparities that accompany serious ethical concerns. ChatGPT's ratings for Black participants were notably lower than those for non-Black participants (d = -0.67), a difference which not apparent in human ratings (d = -0.04). A similar pattern emerged for participants aged 45-54, who received higher BARS ratings from ChatGPT (d = 0.38) than from human evaluators (d = -0.04). These discrepancies reflect meaningful differences in rating patterns, suggesting that AI-human agreement *does* vary across demographic subgroups in ways that may have practical, and serious, ethical consequences.

These results echo prior concerns regarding algorithmic output and fairness outcomes (Barocas & Selbst, 2016; Landers & Behrend, 2023), especially when models trained on large, real-world corpora reflect latent societal biases. While some AI proponents argue that machines merely replicate existing human judgment, the results of this study demonstrate that AI can *diverge* from human evaluation patterns in ways that disproportionately disadvantage marginalized groups, even in the absence of explicit intent. These findings reinforce what Matthias (2004) described as the "*responsibility gap*": the growing disconnect and chasm between decision-making authority, and the capacity to be held accountable for those decisions.

The epistemic opacity of LLMs compounds this problem. When disparities arise, such as the ones found in this study, it is often difficult to identify whether they stem from biased training data, flawed prompt engineering, or emergent behavior within the model itself (Burrell, 2016). From a philosophical standpoint, this challenges the legitimacy of AI-based evaluative authority as a concept. As Burrell (2016) argues, genuine evaluative legitimacy requires the ability to justify decisions through transparent, reason-giving procedures. AI assessments, by

contrast, offer probabilistic outputs without epistemic justification (Landers & Behrend, 2023), undermining the moral authority typically vested in human judgment (Zagzebski, 2012; Fricker, 2007).

Practical Implications for AI-Driven Hiring

The strong correlation and ICC values between ChatGPT and human raters do suggest potential utility for LLMs as assistive tools in personnel selection. If properly calibrated and contextualized, these systems could reduce administrative burdens, improve standardization, and offer preliminary screening support. However, these findings also call for extreme caution against overreliance on automated decision-making in such sensitive contexts. The presence of systematic rating inflation, and more importantly, demographic disparities in rating patterns, underscores the need for human oversight and continuous bias auditing.

HR professionals and organizations considering AI (and specifically LLM) integration into their personnel selection procedures must approach these tools with critical scrutiny. AI-generated scores should never replace human evaluation but rather augment it. They should function as one input in a *broader*, *transparent*, and *accountable* decision-making system. In practice, this may include the implementation of hybrid scoring frameworks, AI calibration protocols across demographic subgroups, or the use of algorithmic "red flag" indicators that trigger human review when discrepancies arise.

Without these safeguards, there is a risk of delegitimizing the evaluative authority within hiring processes, by placing too much weight on systems that lack moral reasoning, context sensitivity, or the capacity to engage in evaluative dialogue. These capacities are essential to personnel selection's role as a trust-building, socially significant activity (Conway et. al., 1995).

Evaluative Authority and Moral Accountability

Delegating evaluative authority to LLMs without moral reasoning capabilities intensifies ethical and epistemic concerns. The subgroup disparities observed (where ChatGPT disproportionately underrated Black candidates and overrated middle-aged candidates) likely reflect biases embedded in the model's training data (Barocas & Selbst, 2016) and its reliance on statistical heuristics, rather than principled judgment (Bommasani et al., 2021). In human evaluators, such misalignments can be debated, contextualized, and rectified through deliberation; AI systems, by contrast, lack the capacity for justificatory dialogue or moral reflection (Burrell, 2016).

This absence of "reason-giving" exacerbates the responsibility gap: when an AI's decision produces unfair outcomes, there is no clear agent to hold accountable (Matthias, 2004; Floridi, 2016). Candidates subjected to opaque, automated ratings cannot challenge or even comprehend the basis of determinations that may severely affect their livelihoods, risking testimonial injustice when their credibility is dismissed by inscrutable algorithmic output (Fricker, 2007). Moreover, these models do not possess the virtues (i.e., empathy, humility, prudence) that are integral to and vested in legitimate evaluative authority (Vallor, 2016).

From an epistemic standpoint, trusting AI requires that systems earn epistemic authority by demonstrating transparent reasoning processes and by aligning with diverse stakeholder values (Zagzebski, 2012). Yet, the probabilistic nature of LLM outputs resists straightforward explanation (Landers & Behrend, 2023), undermining procedural legitimacy (Dennis & Aizenberg, 2022). Until AI assessments can incorporate explicit fairness constraints and support

meaningful recourse mechanisms, handing over primary evaluative power to these systems risks eroding the moral foundations of personnel selection itself.

Limitations and Future Directions

The limitations of this study warrant careful consideration. Firstly, while ChatGPT's textual assessments closely approximated human ratings, the model was limited to processing only transcribed text. In contrast, the human raters from Patel et al.,'s 2025 study evaluated full audiovisual recordings, allowing them to incorporate paraverbal features such as tone, pacing, and hesitations, as well as nonverbal cues like facial expressions and body language. These elements are often essential to forming a holistic impression of a candidate's performance (Zhang et al., 2024). The absence of these modalities in the AI assessment pipeline likely constrained the depth and nuance of its evaluations. Moreover, because paraverbal and nonverbal behaviors can influence perceptions of warmth, confidence, or professionalism (Hickman et al., 2022, Zhang et al., 2024), excluding them may also restrict the fairness and realism of AI-generated assessments, especially in scenarios where such cues carry important interpretive weight.

Secondly, human raters in the dataset did not participate in a formalized adjudication process, or secondary review, meaning that the standard against which AI was compared lacked its own "verification loop" (akin to what the model was instructed to do). This asymmetry could have exaggerated or diminished any observed discrepancies.

Future research should explore the integration of multimodal AI systems capable of analyzing speech tone, visual behavior, and other paralinguistic cues alongside textual content, in comparison to human ratings. Incorporating these dimensions could bring AI evaluations closer to the holistic assessments human raters perform in real-world hiring contexts. Additionally,

expanding the dataset to include more demographically and professionally diverse candidate pools would improve the generalizability of findings across employment sectors and varying cultural contexts.

More longitudinal validation studies are also needed. Tracking whether higher ChatGPT-generated interview ratings correspond to meaningful outcomes (such as job performance, retention, or advancement) could clarify the predictive value of LLM-based assessments and help distinguish between surface-level agreement, versus deeper functional alignment, with human judgment.

Finally, ongoing philosophical inquiry must continue to interrogate the boundaries and nature of evaluative authority regarding AI systems and their associated roles in personnel selection. As automated algorithmic tools become more entrenched in hiring processes, it is crucial to assess not only their technical accuracy, but also the ethical legitimacy of their role as an evaluator. Understanding when, how, and whether evaluative authority can be justifiably delegated to non-human agents is key to preserving accountability & fairness within institutional decision-making frameworks.

Conclusion

This study examined the extent to which ChatGPT-generated interview ratings aligned with human evaluations in the context of automated video interviews (AVIs), focusing on overall agreement, and demographic subgroup variation. The findings indicated that large language models (LLMs) like ChatGPT *can* demonstrate moderate-to-strong alignment with human evaluators, when supported by structured prompts and behavioral rating scales. The model's self-consistency rates also suggest it can operate as a reliable scoring tool across varying contextual conditions, provided appropriate quality controls are in place.

However, the study also identified key limitations that temper enthusiasm about full-scale implementation or transfer of evaluative authority in personnel selection. ChatGPT systematically produced inflated scores relative to human raters, and more notably, demonstrated divergent patterns in its ratings across specific demographic groups. These were most prominently found among Black participants and individuals aged 45-54. These disparities raise urgent concerns about algorithmic bias, fairness, and the ethical risks of delegating evaluative authority in personnel selection to non-human systems (without adequate safeguards in place).

These results carry meaningful implications for both research and industry. For practitioners in HR and personnel selection, the findings suggest that AI tools like ChatGPT may potentially offer value as supplemental evaluators, but only when embedded within hybrid frameworks that incorporate human oversight, ongoing audits, and fairness calibration across subgroups (Landers & Behrennd, 2023). For researchers and ethicists, this study underscores the need for further interdisciplinary inquiry into how epistemic and moral authority is constructed, transferred, and/or undermined in algorithmic decision-making contexts.

Ultimately, while LLMs do offer compelling possibilities for streamlining assessment in high-volume hiring environments, their integration must be approached with the utmost caution and transparency, fostering a commitment to procedural justice. Future work should focus not only on improving technical validity of these models, but also on interrogating the normative frameworks that govern the use of AI in domains where human dignity, fairness, and accountability remain paramount.

References

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104(3), 671–732. http://www.jstor.org/stable/24758720

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., et al. (2021). On the Opportunities and Risks of Foundation Models. *ArXiv.* Retrieved from https://crfm.stanford.edu/assets/report.pdf

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. Journal of Applied Psychology, 100(2), 431–449. https://doi.org/10.1037/a0038047

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12. https://doi.org/10.1177/2053951715622512

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. Journal of Applied Psychology, 80(5), 565–579. https://doi.org/10.1037//0021-9010.80.5.565

Dennis, M. J., & Aizenberg, E. (2022). The Ethics of AI in Human Resources. *Ethics and Information Technology*, 24(3). https://doi.org/10.1007/s10676-022-09653-y

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS One*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Harwood, H. (2024). *Lights, camera, persuasion: Examining the impacts of impression management tactics on predictive validity* (Master's thesis). Saint Mary's University, Halifax, Nova Scotia.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. J*ournal of Applied Psychology*, 107(8), 1323–1351. https://doi.org/10.1037/apl0000695

Hickman, L., Langer, M., Saef, R. M., & Tay, L. (2024). Automated speech recognition bias in personnel selection: The case of automatically scored job interviews. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0001247

Hickman, L., Tay, L., & Woo, S. E. (2024). Are automated video interviews smart enough? behavioral modes, reliability, validity, and bias of machine learning cognitive ability assessments. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0001236

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for Reliability Research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Koutsoumpis, A., Ghassemi, S., Oostrom, J. K., Holtrop, D., van Breda, W., Zhang, T., & de Vries, R. E. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for Personality and Interview Performance Evaluation Using Machine Learning. *Computers in Human Behavior*, *154*, 108128. https://doi.org/10.1016/j.chb.2023.108128

Landers, R. N., & Behrend, T. S. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*, *78*(1), 36–49. https://doi.org/10.1037/amp0000972

Li, L., Lassiter, T., Oh, J., & Lee, M. K. (2021). Algorithmic hiring in practice. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* https://doi.org/10.1145/3461702.3462531

Liff, J., Mondragon, N., Gardner, C., Hartwell, C. J., & Bradshaw, A. (2024). Psychometric Properties of Automated Video interview competency assessments. *Journal of Applied Psychology*, *109*(6), 921–948. https://doi.org/10.1037/apl0001173

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019a). Intraclass correlation – a discussion and demonstration of basic features. PLOS ONE, 14(7). https://doi.org/10.1371/journal.pone.0219854

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McHugh, M. L. (2012). Interrater Reliability: The kappa statistic. *Biochemia Medica*, 276–282. https://doi.org/10.11613/bm.2012.031

Patel, R. D., Powell, D. M., Roulin, N., & Spence, J. R. (2025). Tell me more! examining the benefits of adding structured probing in asynchronous video interviews. *International Journal of Selection and Assessment*, *33*(1). https://doi.org/10.1111/ijsa.12514

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review*, 87(3), 1085–1139.

Selbst, A. D., & Barocas, S. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732. https://doi.org/10.15779/Z38BG31

Spiller, N., Lison, P., & Woldemariam, M. F. (2023). Automatic transcription and annotation of interviews: Accuracy and user perceptions of Whisper. *Behavior Research Methods*, *55*, 1045–1061. https://doi.org/10.3758/s13428-022-02072-4

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial Intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4), 15–42. https://doi.org/10.1177/0008125619867910

Tippins, N. T., Oswald, F. L., & McPhail, S. M. (2021). *Scientific, Legal, and Ethical Concerns about AI-Based Personnel Selection Tools: A Call to Action*. https://doi.org/10.31234/osf.io/6gczw

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190498511.001.0001

Yam, J., & Skorburg, J. A. (2021). From Human Resources to Human Rights: Impact Assessments for hiring algorithms. *Ethics and Information Technology*, 23(4), 611–623. https://doi.org/10.1007/s10676-021-09599-7

Zagzebski, L. (2012). *Epistemic authority: A theory of trust, authority, and autonomy in belief*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199936472.001.0001

Zhang, T., Koutsoumpis, A., Oostrom, J. K., Holtrop, D., Ghassemi, S., & de Vries, R. E. (2024). Can large language models assess personality from asynchronous video interviews? A comprehensive evaluation of validity, reliability, fairness, and rating patterns. *IEEE Transactions on Affective Computing*, *15*(3), 1769–1785. https://doi.org/10.1109/taffc.2024.3374875

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. https://doi.org/10.1037/1082-989X.12.4.399