Can Large Language Models Assess Personality from Asynchronous Video Interviews? A Comprehensive Evaluation of Validity, Reliability, Fairness, and Rating Patterns

Tianyi Zhang*, *Member, IEEE,* Antonis Koutsoumpis, Janneke K. Oostrom, Djurre Holtrop, Sina Ghassemi, and Reinout E. de Vries

Abstract—The advent of Artificial Intelligence (AI) technologies has precipitated the rise of asynchronous video interviews (AVIs) as an alternative to conventional job interviews. These one-way video interviews are conducted online and can be analyzed using AI algorithms to automate and speed up the selection procedure. In particular, the swift advancement of Large Language Models (LLMs) has significantly decreased the cost and technical barrier to developing AI systems for automatic personality and interview performance evaluation. However, the generative and task-unspecific nature of LLMs might pose potential risks and biases when evaluating humans based on their AVI responses. In this study, we conducted a comprehensive evaluation of the validity, reliability, fairness, and rating patterns of two widely-used LLMs, GPT-3.5 and GPT-4, in assessing personality and interview performance from an AVI. We compared the personality and interview performance ratings of the LLMs with the ratings from a task-specific AI model and human annotators using simulated AVI responses of 685 participants. The results show that LLMs can achieve similar or even better zero-shot validity compared with the task-specific AI model when predicting personality traits. The verbal explanations for predicting personality traits generated by LLMs are interpretable by the personality items that are designed according to psychological theories. However, LLMs also suffered from uneven performance across different traits, insufficient test-retest reliability, and the emergence of certain biases. Thus, it is necessary to exercise caution when applying LLMs for human-related application scenarios, especially for significant decisions such as employment.

Index Terms—Large Language models, Personality Recognition, Asynchronous Video Interviews, Personnel Selection

1 Introduction

The utilization of Artificial Intelligence (AI) technologies has substantially affected traditional job interview procedures and has been further accelerated by the COVID-19 pandemic [1]. Asynchronous Video Interviews (AVIs) have become an established personnel selection tool in the past years. An AVI is an online interaction where interviewees offer their video responses to questions presented on their computing devices, tablets, or smartphones. Vendors often assess personality in AVIs, as ample theories and research [2] tie it to workplace behaviors, such as job performance or organizational citizenship behavior [3]. Although AVIs are typically evaluated by employers, AI technologies such as deep learning methods have been deployed for personality and interview performance evaluation in AVIs as they can save time and cost for candidate selection [4], [5]. For example, Suen et al. [6] developed an AI-based platform to predict interviewees' communication skills and personality traits in a structured interview context. The use of AI to assess AVIs is becoming increasingly popular and is being embraced by major companies. For example, *Hirevue* reported that they have "*hosted more than 26 million video interviews and 5 million AI-based candidate assessments*" in 2022 [7]. Even though vendors can develop custom-built AI models, one disadvantage of those models is that they need to be trained on a large number of observations, which is costly and time-consuming. Recently, the emergence of Large Language Models (LLMs), such as *ChatGPT* and *Bard*, has set a milestone in the AI community for their powerful capability on universal tasks [8], [9]. The good zero-shot performance [9] of LLMs makes them highly useful tools for personality and interview evaluation [10]; by prompting the interviewees' answers from AVIs, recruiters can easily obtain ratings of personality traits and interview performance without designing and training an AI model by themselves.

However, our understanding remains limited regarding whether the LLM-based evaluation of AVIs adheres to the psychometric standards that typify the assessment methodologies employed by human evaluators. Although LLMs are fine-tuned using Reinforcement Learning from Human Feedback (RLHF) [9], they are still unsupervised generative models, instead of supervised discriminative models, that can suffer from "hallucinations" [9], [11] (i.e., provide unreliable information). In addition, LLMs are designed to comprehend generic language tasks (i.e., task-unspecific) rather than specific tasks. The generative and task-unspecific characteristics of LLMs may introduce potential risks and biases [9]. For example, they can exacerbate bias toward groups with specific gender or educational backgrounds [12]. The understanding of these risks and biases is crucial and indispensable prior to the deployment of any AI algorithm for human-related

- *Tianyi Zhang is with the Key Laboratory of Child Development and Learning Science (Ministry of Education), School of Biological Sciences and Medical Engineering, Southeast University, Nanjing, China.*
- *Tianyi Zhang, S. Ghassemi, and R. E. de Vries are with the Department of Experimental and Applied Psychology, Vrije Universiteit, Amsterdam, Netherlands. E-mails: { t.zhang, s.ghassemi, re.de.vries} @vu.nl*
- *J. K. Oostrom, A. Koutsoumpis and D. Holtrop are with the Department of Social Psychology, Tilburg University, Tilburg, Netherlands Email: J.K.Oostrom,A. Koutsoumpis,D.J.Holtrop@tilburguniversity.edu*

∗ *corresponding author*

applications.

Although some initial efforts [10], [13] have been taken to evaluate the ability of LLMs for personality recognition, they mostly focus on the validity (i.e., accuracy) of LLMs. However, previous works [14], [15]have argued that for human-related AI systems, reliability and fairness are more crucial than accuracy. Furthermore, previous work did not address the question of how the LLMs rate personality traits and interview performance scores (i.e., the question of interpretability). This lack of interpretability can lead to skepticism and mistrust among users even if the prediction accuracy is high [16].

To address the aforementioned research gaps, we conducted a comprehensive assessment [1] of GPT-3.5 [2] and GPT-4 [3], two of the most widely-used LLMs, for Automatic Personality and AVI assessment (AP-AVI). The assessment was conducted on the AVI answers and corresponding HEXACO [17] personality ratings from 685 subjects who participated in a mock management traineeship application. In the interview, we focused on the traits of Extraversion and Conscientiousness, two of the most valid personality traits for workplace behaviors [18]. We not only asked the LLMs to rate personality traits and interview performance scores but also asked them to provide verbal explanations of why they made these assessments. This enabled us to gain insights into the rating patterns of LLMs. We compared the performance between GPT-3.5 and GPT-4 to explore whether the larger model size and training data of GPT-4 can enhance the validity, reliability, and fairness for AP-AVI. Our study is conducted to answer two research questions: 1) *Whether LLMs can provide valid, reliable, and fair predictions for personality traits and interview performance?* (**RQ1**, the question of **Performance**) and 2) *Whether the LLMs follow similar rating patterns as human annotators when rating personality traits and interview performance?* (**RQ2**, the question of **Interpretability**).

To answer the question of *Performance*, we assessed the validity of LLMs by comparing the coefficient of determination (R^2) of GPT-3.5 and GPT-4 with a task-specific AI model. This comparison can help us explore whether LLMs, which do not contain the information from our dataset (i.e., the zero-shot performance), can outperform an AI model which is specifically developed and trained using our dataset. We additionally implemented a simulation of candidate selection to evaluate whether LLMs can function as adequate recruitment agents within a screening process. Reliability was assessed by both the repeated measures and the test-retest correlation for a sub-group of participants who participated in the same study twice. Finally, we explored the effect of four potentially biasing variables (i.e., gender, age, attractiveness, and educational background) to analyze the fairness of LLMs. To answer the question of *Interpretability*, we first compared the performance of LLMs with and without providing meta-information (i.e., psychological guidelines about which personality traits or facets the questions are related to). This can give us insight into whether psychological guidelines can promote the rating performance of LLMs, similar to how they help human annotators for rating personality traits. Additionally, we analyzed the extent to which the linguistic pattern of the explanations from LLMs was in line with the conceptual operationalization of the corresponding personality items in the HEXACO-60 inventory [19]. By analyzing the verbal explanations from LLMs, we can decode their decision-making process and find out whether it follows similar rating patterns as human annotators. The exploration of these two research questions can help us understand the potential response motivation and thinking mode of LLMs, thereby facilitating the development of more trustworthy and human-friendly LLMs for human-related applications.In general, our work contributes to the affective computing community by providing a comprehensive evaluation for LLM-based personality recognition:

- We assessed the validity of two LLMs (GPT-3.5 and GPT-4) for rating personality traits and interview performance by comparing their performance with a classic machine learning baseline method. Our results show that LLMs achieve higher or similar construct validity with zero-shot training data compared with the baseline method which is trained on our dataset. However, their performance is unbalanced in predicting Extraversion and Conscientiousness compared with the baseline method.
- We use repeated measures and test-retest experiments to evaluate the reliability of LLMs for predicting personality traits and interview performance. We found that the average reliability of LLMs for personality recognition is similar to the baseline model while their reliability is still much lower than the recommended values for job selection.
- We explored the effect of four potentially biasing variables (i.e., gender, age, attractiveness, and educational background) to analyze the fairness of LLMs. Our results indicated that AI models (both LLMs and the baseline method) can reproduce and sometimes further increase biases existing in human observers.
- We conducted experiments on the meta-information and the linguistic patterns of LLMs to analyze whether the LLMs follow similar rating patterns as human annotators. We found that LLMs do show some similarity with human annotators. However, the linguistic analysis shows that they may not be equally adept at capturing all facets of human personality.

2 Related Work

In this section, we first introduce the HEXACO personality model, which is used to quantify the ground truth labels of the study. After that, we review the rapid development of LLMs, deliberating both their advantages and shortcomings for personality recognition. Finally, we discuss previous works on evaluating the capability of LLMs for personality recognition.

2.1 HEXACO personality model

The HEXACO personality model [17] is one of the most widely used models to quantify personality traits [20], [21]. The HEXACO model delineates human personality through six major *factors*: Honesty-Humility (H), Emotionality (E), eXtraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O). Being grounded in the same lexical research tradition, the HEXACO model has emerged as a correction and expansion of the Big Five personality model, incorporating an additional factor, Honesty-Humility, and re-partitioning the variance of Neuroticism and Agreeableness from the Big Five model into Emotionality, Agreeableness, and Honesty-Humility in the HEXACO model [22]. Each factor in the HEXACO model is further divided into

[1] The codes and data for this assessment are available at: https://github.com/Tianyi-Zhang-TZ/LLMs_Personality

[2] https://platform.openai.com/docs/models/gpt-3-5

[3] https://platform.openai.com/docs/models/gpt-4

four *facets*, providing a more detailed representation of individual personality differences. This division allows for a rich exploration of personality by dissecting each major trait into more specific tendencies and behaviors. The faceted structure of the HEXACO model offers a comprehensive view of personality, capturing a wide array of individual differences through a detailed breakdown of each major personality factor into more specific and focused facets [23]. The detailed structure of the HEXACO model also aids in predicting a variety of outcomes related to job performance [23], relationship quality [24], and mental health [25], thus holding a significant edge over other personality models in predictive validity. Thus, we use the HEXACO personality model to quantify the ground truth (both self-reported and observer-reported) ratings.

2.2 Large language model-based personality assessment

LLMs are pre-trained generative neural networks that can perform a variety of natural language processing (NLP) tasks [26]. LLMs are normally pre-trained by large amounts of data to understand the implicit patterns in natural languages. Thus, LLMs are capable of generating responses which are coherent and contextually relevant to the responses from human beings [9]. The rapid development of LLMs (e.g., Bard, ChatGPT, Claude [4]) has benefited various applications such as content creation and customer service, and reshaped several industries and aspects of our daily lives [27].

One of the advantages of LLMs is that they can perform a wide variety of NLP tasks without task-specific training data (i.e., good zero-shot performance) [9], [10], which makes them useful tools for text-based personality assessment. The unique challenge of applying deep learning networks to personality recognition is the considerable annotation workload (in terms of time and cost) necessitated to compile a sufficient training dataset for the model. For observer-ratings, it usually requires at least three external annotators [28] to get a meaningful agreement (e.g., high Intraclass Correlation Coefficient (ICC)). The annotators are supposed to be personality experts who went through a special training program designed by the researchers. However, with LLMs, researchers can simply prompt the verbal answers of participants (e.g., interviewees) and obtain personality ratings without collecting and training models by themselves. Thus, many researchers [13], [29], [30] have started to develop tools or platforms to rate personality traits using LLMs. For example, Rao et al. [13] devised unbiased prompts, subject-replaced queries, and correctness-evaluated instructions to enable LLMs for a reliable assessment of personality traits. Their experimental results showed that GPT-4 can independently assess personality traits while their results are sensitive to prompt biases. Ji et al. [31] compared three different type of prompting for LLM-based personality assessment. They found that ChatGPT with chain-of-thought prompting exhibits good personality recognition ability and is capable to provide verbal explanations through text-based logical reasoning.

Although the previous works mentioned above demonstrated the potential for LLM-based personality assessment, the LLMs are still originally designed and trained to understand generic language tasks (i.e., task-unspecific), rather than conduct one specific task. For example, GPT-4 is a Transformer-style model [32] pre-trained to predict the next token in a document. Despite their powerful capability on generative tasks, their performance on other specific tasks, such as predicting personality traits, is somewhat constrained and necessitates further evaluation. In addition, although most LLMs are fine-tuned using RLHF, they are still unsupervised generative models instead of supervised discriminative models. Thus, they can suffer from low reliability (hallucinations) [9], limited knowledge of professional fields [33], and biases in their outputs [31]. The generative and task-unspecific characteristics of LLMs introduce potential risks [26] and biases [31], especially in fields deeply intertwined with human behavior and psychology. Thus, the understanding of these limitations is pivotal for ensuring ethical and safe utilization in any human-related domains.

2.3 Evaluating LLMs for personality assessment

Concerning the necessity and importance of evaluating LLMs for personality assessment, initial studies [10], [13], [31], [34] have been commenced to examine the validity of personality evaluations facilitated by LLMs and to explore potential biases. For example, Ganesan et al. [10] conducted an evaluation of LLMs' performance in estimating the Big five personality traits derived from social media postings. Their results show that the zero-shot GPT-3 performance is close to the performance of a classic personality recognition model [35]. Amin et al. [34] evaluated GPT-3.5 and GPT-4 on 13 affective computing tasks, including personality assessment using the Chalearn First Impressions dataset [36]. Similar to the findings from [29], [30], the recognition accuracy of GPT-3.5 is worse than the baseline method (Bag-of-Words, BoW).

Although there are some preliminary studies to evaluate LLMs' performance for personality recognition, most of these works focus on the validity (i.e., accuracy) of LLMs. However, the reliability and fairness of LLM-based personality recognition may hold great significance as well [14], [15], [37]. For example, Ji et al. [31] found that GPT-3.5 exhibits unfairness to some sensitive demographic attributes on text-based personality recognition tasks, with men and elderly people having a higher likelihood to be rated as low on Openness when compared to women and young people respectively. In addition, the black-box nature of LLMs has led to skepticism and mistrust [16], [38] among users when applying them to human-related applications. Thus, the question of how LLMs rate personality traits (i.e., their rating patterns), while remaining limitedly discussed, is of paramount importance.

To address the aforementioned limitations of previous works, we conducted a comprehensive and psychological evaluation of the validity, reliability, fairness, and rating patterns for LLM-based personality assessment. That is, our work not only conducted a deep analysis of the issues of concern in previous works [10], [31], [34] (i.e., the validity), but also expanded the discussion on issues related to potential risks and biases for LLM-based personality assessment.

3 Methodology

3.1 Experiment data

3.1.1 Participants

We assessed the performance of LLMs for AVI analysis using the data [39], [40] collected from 685 participants. The participants first read the job post of a management traineeship for a fictitious company and provided video responses to eight interview questions using their web camera and microphone. Participants were recruited via the crowdsourcing platform Prolific[5] (n = 889) and

[4] https://bard.google.com, https://chat.openai.com, https://claude.ai/

[5] https://www.prolific.co/

via a snowball sampling procedure (n = 58). Participants were excluded from the dataset if 1) the audio, video, or annotation file was corrupted (n = 100), 2) they failed to pass the attention check (n = 28), 3) their scores on the personality inventory showed too small ($SD < 0.70$) or too large variability ($SD > 1.60, n = 92$) [41], and/or if 4) participants were manually flagged by observers as non-compliant (e.g., not taking the interview seriously; n = 42). Among these participants, 231 self-identified as men, 447 as women, and 7 as non-binary. Participants were on average 31.08 years old (SD = 11.52). For their educational background, 4 had a level lower than high school, 55 were high school graduates, 183 held a lower-than-college degree, 305 held a college degree and 138 held a postgraduate degree. Each participant was paid $7 for participating in the experiment.

3.1.2 Interview questions

The AVI consisted of eight interview questions (in Appendix A) designed by personality experts to collect information about the candidates' personality facets (i.e., a specific and fine-grained aspect of one personality trait) of *Extraversion* (facets: Sociability, Social Self-esteem, Social Boldness, and Liveliness) and *Conscientiousness* (facets: Perfectionism, Prudence, Diligence, and Organization), respectively. All eight questions were past-behavioral [42] interview questions and contextualized in the workplace [43]. For the question development, a set of 26 questions was developed initially, with 14 and 12 questions for Extraversion and Conscientiousness, respectively. These questions were evaluated by three personality experts in two rounds to ensure they effectively elicited 1) the targeted traits (with an Intraclass Correlation Coefficient ($ICC(2, 3)$) of 0.84 in the first round) and (b) facets (with an $ICC(2, 3)$ of 0.99 in the second round). The final AVI consisted of one question per facet (four activated Extraversion, four activated Conscientiousness), with absolute agreement ($ICC(2, 3) = 1$).

3.1.3 Ground truth annotations

In the most thorough research on personality and interview evaluation, Hickman et al. [5] found that AI-based recognition models achieved much higher construct validity on observer-reported personality traits than self-reported traits. However, their task-specific models were trained and tested on these two "*ground truth*" annotations, respectively. In our study, we want to explore whether the LLMs, which do not have information on observer-reported or self-reported annotations, will still reproduce this finding. This approach can help us validate Hickman et al.'s findings by eliminating the potential bias and overfitting from the training data. Thus, in our study, we collected both self-reported personality traits as well as observer-reported traits and facets to compare them with the ratings from LLMs.

The self-reported personality traits were collected using the HEXACO-60 inventory [19], which measures six personality factors. For each factor, participants rated their own personality traits on 10 items, using a 5-point scale (1-5). For observer-reported personality traits, participants were rated by four independent raters using a Behavioral Anchored Rating Scale (HEXACO-BARS, 5-point scale). Trained raters scored the personality facets and factors after watching the answers to the four questions activating each factor, and the answers to all eight questions in the AVI, respectively. The inter-rater agreement for Extraversion and Conscientiousness was $ICC(1, 4) = 0.91$ and 0.77, respectively.

For the annotation of interview performance, six professional recruiters (2 for each participant) were assigned to rate four job-related competences (i.e., Communication Flexibility, Persuasiveness, Quality Orientation, Development Orientation) and overall Interview Performance scores. The four competences were aligned with the traits of Extraversion and Conscientiousness and taken from the manual of a Dutch consultancy company[6] [40]. The averaged inter-rater agreement across the five competences is $ICC(1, 2) = 0.56$, ranging from $ICC(1, 2) = 0.49$ (Persuasiveness) to 0.64 (Communicative Flexibility).

We also collected annotations of first impression attractiveness to analyze the potential bias of different models. Attractiveness was also annotated using a 5-point scale. The annotations were provided after raters had watched thin slices of the interview (three random snippets of 4-5 seconds) to avoid being influenced by the content of the AVI. The average attractiveness score was 3.11 (SD = 0.49), and inter-rater agreement was $ICC(2, 4) = 0.75$.

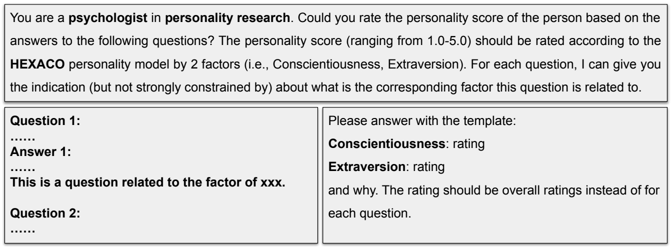
3.1.4 Test-retest reliability

To examine the test-retest reliability, we invited 145 participants to complete the same AVI again (T2). The time difference between the two studies (T1 and T2) ranged from 7 to 24 months (M = 11.80, SD = 6.50, Median = 7 months). We decided to measure test-retest reliability over a longer duration (instead of e.g., two weeks [5]) to avoid artificially attenuating reliability estimates from participants (e.g., participants might remember their previous responses and think that researchers are interested in new information [44]).

3.2 LLMs implementation

To obtain the LLM personality predictions, we automatically transcribed the AVI responses using Google Cloud's speech-to-text transcription service. Then, we provided the transcribed responses (and interview questions) to the two LLMs using the prompt in Fig 1. We created an independent session for every participant to make sure that LLMs do not provide ratings based on their previous response (no context information). We set the temperature of the model to be 1 in our analysis to make the model neither too deterministic nor too random: the model selects words based on its calculated probabilities without significant alteration. Moreover, this level of temperature is suitable as it provides a good mix of predictability and novelty, making it versatile for a wide range of tasks.

The prompts sent to LLMs were designed to mimic the human-based annotation procedure of our study. Here, we use chain-of-thought prompting strategy (i.e., ask the LLMs to think about and output why it rates the personality traits like this) inspired by previous works [45], [46].



You are a **psychologist** in **personality research**. Could you rate the personality score of the person based on the answers to the following questions? The personality score (ranging from 1.0-5.0) should be rated according to the **HEXACO** personality model by 2 factors (i.e., Conscientiousness, Extraversion). For each question, I can give you the indication (but not strongly constrained by) about what is the corresponding factor this question is related to.

Question 1:
......
Answer 1:
......
This is a question related to the factor of xxx.

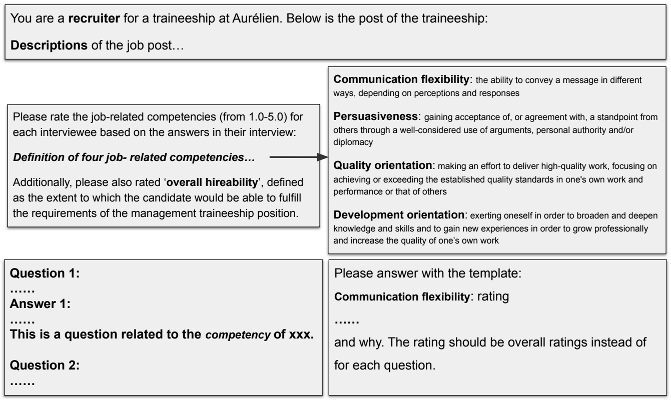
Question 2:
......

Please answer with the template:
Conscientiousness: rating
Extraversion: rating
and why. The rating should be overall ratings instead of for each question.

Fig. 1. The prompt template for personality rating

[6] https://ltp.nl/

More specifically, regarding personality traits, we asked LLMs to pretend to be a psychologist in personality research and rate the personality factors or facets (ranging from 1.0 - 5.0) of the participants according to the HEXACO personality model. The LLMs were prompted to rate personality across all eight interview questions, instead of giving one rating for each question. Since human annotators are aware of the personality factor that each interview question corresponds to, we also provided this meta-information (i.e., the correspondence between interview questions and personality factors) to the LLMs, to mimic as much as possible the rating procedure between human annotators and LLMs.



You are a **recruiter** for a traineeship at Aurélien. Below is the post of the traineeship:

Descriptions of the job post…

Please rate the job-related competencies (from 1.0-5.0) for each interviewee based on the answers in their interview:

Definition of four job- related competencies…

Additionally, please also rated ‘**overall hireability**’, defined as the extent to which the candidate would be able to fulfill the requirements of the management traineeship position.

Communication flexibility: the ability to convey a message in different ways, depending on perceptions and responses

Persuasiveness: gaining acceptance of, or agreement with, a standpoint from others through a well-considered use of arguments, personal authority and/or diplomacy

Quality orientation: making an effort to deliver high-quality work, focusing on achieving or exceeding the established quality standards in one's own work and performance or that of others

Development orientation: exerting oneself in order to broaden and deepen knowledge and skills and to gain new experiences in order to grow professionally and increase the quality of one's own work

Question 1:
……
Answer 1:
……
This is a question related to the *competency* of xxx.

Question 2:
……

Please answer with the template:

Communication flexibility: rating

……

and why. The rating should be overall ratings instead of for each question.

Fig. 2. The prompt template for interview performance

Regarding interview performance, we followed a similar approach. We asked the LLMs to be a recruiter for a management traineeship position, requiring them to rate four job-related competences. In the prompt, we provided detailed definitions of these four competences. We also asked the LLMs to give an “overall interview performance” score indicating to what extent the candidate would be able to fulfil the requirements of the traineeship position. Fig. 2 shows the prompt template for rating job-related competences and overall interview performance scores.

3.3 Baseline model

To set a baseline model for LLMs, we compared GPT-3.5 and GPT-4 with a supervised regression model (BoW-SBERT) which was specifically trained on our dataset. We selected a deep learning model which is similar to the model (i.e., the transformer) used by GPT-3.5 and GPT-4 to minimize the bias for comparison. The verbal features are extracted in an unsupervised manner and then fed into a supervised network for regression. Thus, the main difference between LLMs and BoW-SBERT is that BoW-SBERT was trained and validated on our dataset, while the LLMs do not have any information from our dataset.

We first processed the transcribed text from AVIs using two groups of features: 1) Bags of Words (BoW) and 2) Sentence Bidirectional Encoder Representations from Transformers (S-BERT) [47]. The first group (BoW) counted the number of occurrences of the 512 most common words (including stop words) in the text. The second group consisted of sentence embedding extracted from a pre-trained S-BERT model. The model was pre-trained on a large corpus of English data in a self-supervised manner. The feature vectors from all sentences in the text were aggregated using mean and max pooling. In general, the BoW focuses on individual words, while S-BERT takes into account the context of the words. This approach allows these feature sets to complement one another, thereby yielding a more comprehensive representation of verbal information. After extracting the features, we trained a multi-layer neural network ($hiddenlayer = (256, 32)$) for the regression using 10-fold cross-validation.

3.4 Analysis

3.4.1 Construct validity

The construct validity is assessed both for observer-reported and self-reported personality factors, facets, job-related competences, and overall interview performance scores. The evaluation metric used to measure the construct validity of LLMs is the coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\widehat{y_i} - y_i)^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2} \tag{1}$$

where y_i is the ground truth for i^{th} observation, $\widehat{y_i}$ is the model prediction, and $\overline{y}$ denotes the mean across all N observations. We use R^2 as 1) it is the most widely used metric for psychological studies and 2) it compares the unexplained variance (i.e., the variance of prediction errors) with the total variance of the data.

3.4.2 Reliability

We assessed the reliability of the LLMs for personality and interview performance assessment in two ways. First, we asked LLMs to predict the ground truth measures (i.e., personality, interview performance) twice based on the same text (i.e., transcribed interview response). Then, we calculated the Pearson r correlation between the two predictions and, in this way, we calculated whether LLM predictions are consistent when analyzing the exact same text. Second, since some participants took the AVI twice (T1 and T2; see section 3.1.4), we asked LLMs to predict the ground truth measures for T1 and T2 and calculated the correlation between the two time points. In this way, we calculated the test-retest reliability of LLMs over time.

3.4.3 Fairness

Four potential biasing variables were selected to analyze the fairness of LLMs. Previous works suggest that these four variables, i.e., gender [12], age [48], attractiveness [49], and educational background [50], might be directly detectable by AI models and introduce bias for the AP-AVI. The model bias is characterized by disparities in ratings across different groups (e.g., people with different genders). While such disparities may reflect existing real-world biases, the LLMs can reproduce (further increasing or decreasing them) the existing, real-life bias or introduce group differences that did not exist in real life [12]. Thus, we compared the differences between the predictions of LLMs, the baseline model, human-based observer-reported ratings, and self-reported ratings to find out whether the LLMs, which do not learn the distribution of our dataset, increase or decrease the personality scores compared to the baseline model, observer-reported and self-reported ratings. In order to examine the variations among different raters (i.e., GPT-3.5, GPT-4, BoW-SBRET, observers, and self) with respect to biasing according to the four variables, we employed a linear mixed-effects model (lmer) that regresses general ratings on the interaction between the raters and the four variables. According to the inherent structure of the data, we added random intercepts of participants as the nesting structure.

3.4.4 Candidate selection

In order to evaluate whether LLMs can function as adequate recruitment agents, we implement a simulation for candidate selection. Suppose we have N interviewees and we seek to hire K individuals for the traineeship position, we first rank their predicted interview performance scores from the highest to the lowest and identify the Kth highest score C. Since interviewees may have identical scores, we select L ($L \geq K$) candidates whose scores are greater or equal to C. After that, we compare the selected candidates by both LLMs and human recruiters. Suppose LLMs and human recruiters select L_l and L_h candidates, respectively, we define the true positive rate (REC_{select}) and precision (PRE_{select}) using human recruiters as ground truth for LLMs:

$$REC_{select} = \frac{L_l \cap L_h}{L_h}, PRE_{select} = \frac{L_l \cap L_h}{L_l} \tag{2}$$

where $L_l \cap L_h$ are the numbers of candidates selected by both LLMs and human recruiters respectively. The true positive rate indicates the proportion of candidates selected by human recruiters that were also selected by the LLMs, providing a measure of the LLMs' alignment with human decision-making on selection. Instead, the precision reveals the fraction of candidates selected by the LLMs that were also chosen by human recruiters, offering insights into the model's ability to avoid redundant candidates in the selection process.

3.4.5 Effect of meta-information

Since LLMs are quite novel, it is still unclear whether certain task design choices might affect their performance. To explore the rating pattern of LLMs compared to human annotators, we evaluated the effect of the meta-information for AVI questions including and excluding this meta-information. During the rating process of human observers, the meta-information of which personality factor and facet this question elicits is provided. This information can help human annotators to understand the context of certain statements in the answers and identify personality traits more accurately. Thus, we want to find out whether this information can also help LLMs in providing better predictions.

Specifically, we defined the predicted personality factors P (i.e., Extraversion and Conscientiousness) as a function of answers from AVIs A and meta-information M, i.e., $P = f(A, M)$. Here, we delineate M into three levels M_0, M_1, and M_2. M_0 (without-meta) signifies a condition in which no meta-information is provided and LLMs solely depend on the answers A from the AVIs for prediction. Conversely, M_1 (with-factor) provides the information on which factor each answer is associated with. M_2 (with-facet) provides a more detailed level of meta-information, indicating the specific facet each answer relates to. For M_2, we instruct the LLMs to initially rate each personality facet, and subsequently, compute the mean of the facet ratings within the same factor to derive the factor ratings. The procedure of M_2 is identical to the approach used by human annotators for rating personality factors.

3.4.6 Linguistic analysis

Except for the ratings for personality traits and interview performance, we also asked the LLMs to provide verbal explanations for their predictions (e.g., "*... The individual's responses indicate a modest level of Extraversion as they tend to be reserved and quiet in social situations, but also comfortable initiating conversations and contributing to group discussions...*"). These explanations allow us to gain insight into how the LLMs rate the personality traits and interview performance. To do this, we implemented a close-vocabulary linguistic analysis to compare the verbal explanations with the items in the HEXACO-60 (e.g., "*The first thing that I always do in a new place is to make friends*" -example Extraversion item; "*I plan ahead and organize things, to avoid scrambling at the last minute*"-example Conscientiousness item). The explanations of LLMs and HEXACO-60 items are first transformed into sentence embeddings using a model pre-trained on a large amount of text, SentenceTranformer. After that, we calculated the l^2 Euclidean distance between each sentence embedding with each item embedding in the HEXACO-60. Suppose we have N explanations for N participants and each explanation contains $M_i, i \in [1, n]$ sentences, the l^2 distance between sentence S_{M_i} and item I_j is:

$$l^2_{(S_{M_i}, I_j)} = \left(E_{S_{M_i}} - E_{I_j}\right)^T \left(E_{S_{M_i}} - E_{I_j}\right) \tag{3}$$

$E_{S_{M_i}}$ and $E_{I_j}, j \in [1, 60]$ are the normalized embedding vectors for S_{M_i} and I_j respectively. Then we normalized (using min-max normalization) the distances for each sentence along 60 items to get the similarity between S_{M_i} and I_j:

$$R_{(S_{M_i}, I_j)} = 1 - \frac{l^2_{(S_{M_i}, I_j)} - \min l^2_{(S_{M_i}, I)}}{\max\left(l^2_{(S_{M_i}, I)}\right) - \min\left(l^2_{(S_{M_i}, I)}\right)} \tag{4}$$

At last, we average $R_{(S_{M_i}, I_j)}$ for all the S_{M_i} to get the R_{I_j} distances for each I_j:

$$R_{I_j} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{M_i} \sum_{j=1}^{M_i} R_{(S_{M_i}, I_j)} \tag{5}$$

The R_{I_j} indicates the averaged similarity between the explanations of LLMs with each HEXACO-60 item I_j. Thus, higher R_{I_j} could suggest that the LLMs rate personalities and interview performance that align closely with specific personality factors or facets measured by the HEXACO-60 item.

4 Results

4.1 Construct Validity

Fig 3 (a-b) shows the R^2 comparison between GPT-3.5, GPT-4, and the baseline method (BoW-SBERT). We found that all three methods explained substantially more variance in observer-reported ratings of Extraversion and Conscientiousness than in self-reported ratings. For the factor of Extraversion, the two LLMs outperformed the BoW-SBERT both on observer-reported and self-reported ratings. However, the LLMs explained less or failed to explain any variance (i.e., $R^2 < 0$) in Conscientiousness on observer-reported and self-reported ratings, respectively. On the facet level, the LLMs also failed to explain variance in observer-reported Prudence and Perfectionism (both facets of Conscientiousness). When comparing the LLMs' ratings for the four job-related competences and overall interview performance scores provided by human recruiters, we found that LLMs are unable to explain variance in job-related competences and overall interview performance (all $R^2 < 0$; not presented in Fig 3).

Among all three AI models, GPT-4 achieved the highest R^2 in most of the cases, except for self-reported Conscientiousness as well as observer-reported Prudence and Perfectionism ($R^2 < 0$). On average, GPT-4 performed 32.5% and 47.2% better than GPT-3.5 and BoW-SBERT on observer-reported Extraversion,

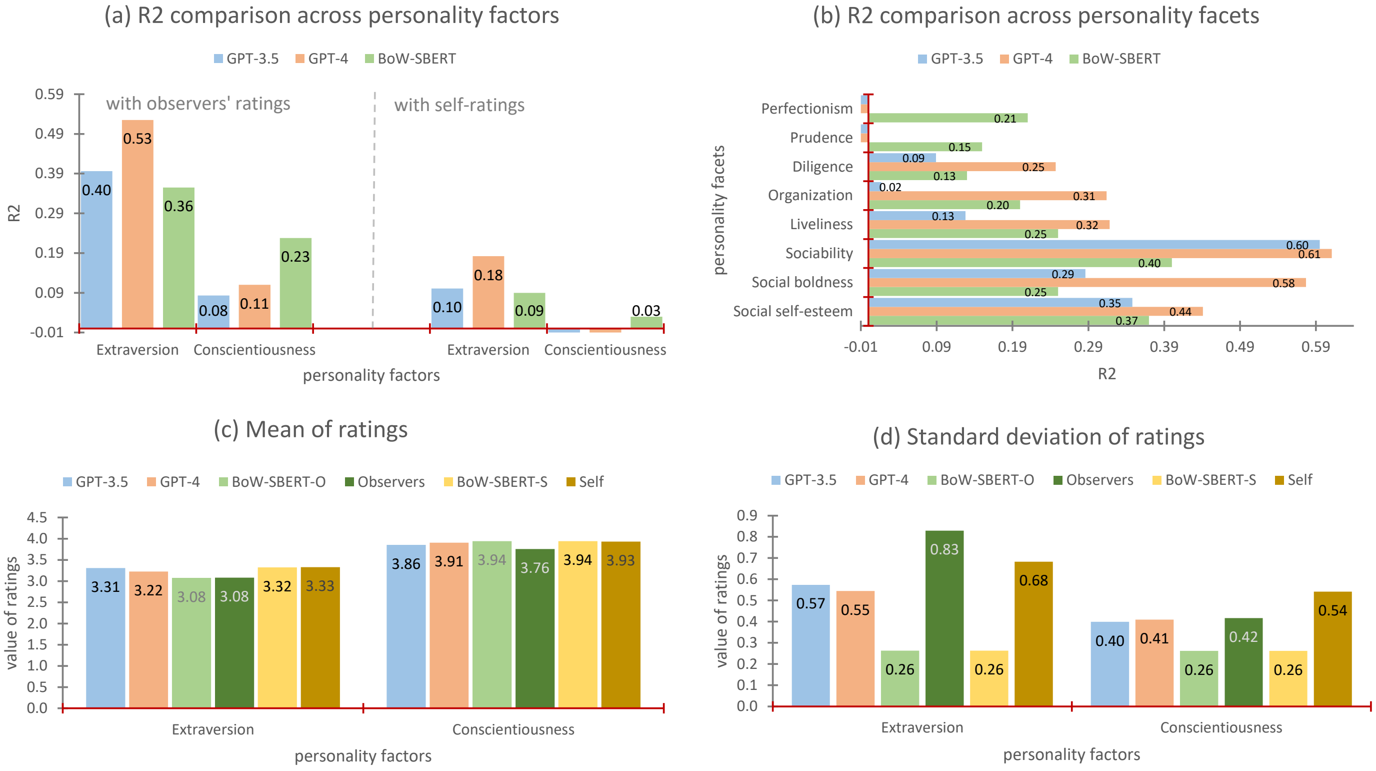


Fig. 3. R^2 comparison between (a) mean observer-reported and self-reported personality factors, (b) observer-reported personality facets as well as the (c) mean, and (d) standard deviation of predicted and annotated personality factors. Negative R^2 signifies the inefficiency of the model to predict the target variable (worse than just a simple average); for $R^2 > 0.01$ values are statistically significant at $p < 0.05$; $R^2 > 0.02, p < 0.01$; and for $R^2 > 0.03, p < 0.001$; "BoW-SBERT-O" and "BoW-SBERT-S" means BoW-SBERT trained using observer-reported and self-reported ratings as ground truth labels respectively. All R^2 for job-related competences and overall interview performance are < 0

respectively. Both GPT-3.5 and GPT-4 had good zero-shot performances for explaining the variance of observer-reported ratings on Extraversion: both of them not only achieved higher R^2 on the personality factors but also on the four personality facets (except for GPT-3.5 on Liveliness) of Extraversion. However, the BoW-SBERT demonstrated a more generalized performance across various personality factors and facets (all $R^2 > 0$) compared with the two LLMs.

For the statistical comparison of different ratings, we first ran the Shapiro-Wilk test to check the assumption of normality. We found that the Extraversion ratings from GPT-3.5, GPT-4, observers, and self were not normally distributed ($p > 0.05$), while the ratings from BoW-SBERT were ($p < 0.05$). For Conscientiousness, all the ratings were not normally distributed ($p > 0.05$) for either model. We first compared the mean values of each model. The results from the Mann-Whitney U-test with Bonferroni correction for pairwise comparisons showed that the mean values of LLMs were significantly different (all $p < 0.05$) from the observer-reported and self-reported ratings. However, both the mean values of observer-reported ratings and BoW-SBERT_O, as well as self-reported ratings and BoW-SBERT_S, did not show significant differences (all $p > 0.05$).

As shown in Fig 3 (d), the standard deviation of the ratings from BoW-SBERT was much lower than those of the scores of the LLMs. Thus, we used pairwise Levene's Test to test whether their variances are significantly different. We used Levene's Test because it is robust to violations of the assumption of normality. We found that the variance of the two LLMs (i.e., GPT-3.5 and GPT-4) does not differ from each other ($F = 0.51$, $p = 0.48$; $F = 0.70$, $p = 0.40$ for Extraversion and Conscientiousness respectively). However, the variances of the BoW-SBERT scores and the LLMs did significantly differ (all $p < 0.05$), with BoW-SBERT variance being significantly narrower. Furthermore, the variances of the ratings from all three AI models significantly differed from those of both the observer-reported and self-reported ratings (all $p < 0.05$).

Discussion: We found that all three AI models explained more variance in observer-reported than self-reported personality traits. This finding is consistent with prior research [5], [28], [51], [52] demonstrating that verbal features have a stronger correlation with observer-reported, as opposed to self-reported personality measures. The observers and AI algorithms relied on a similar pool of information, (i.e., verbal and nonverbal information for observers, and verbal information for AI algorithms). However, self-reported personality relied on previous thoughts, feelings, and behaviors, which are not necessarily displayed during the AVI. Thus, the self-reported personalities have a broad scope and are significantly different (*Kruskal-Wallis H test*, $p < 0.01$) from observer-reported personalities.

We also found that both our baseline method and two LLMs explained more variance in Extraversion than Conscientiousness. In particular, both LLMs cannot explain the variance in self-reported Conscientiousness. The results are in line with LLMs' performance on facets: they also failed to explain variance in Perfectionism and Prudence (both facets of Conscientiousness). This is coherent with the finding that Conscientiousness is more difficult to assess than Extraversion during job selection [53]. The two LLMs also failed to explain the variance of job-related competences and overall interview performance, which makes them unsuitable for selecting candidates for employment purpose (more details on the experiment and discussion in section 5, candidate selection). Thus, explaining variance in Conscientiousness and job-related competences requires at least some annotations from human experts to train the deep learning models.

4.2 Reliability

Table 1 shows the Pearson r for repeated measures and test-retest. For repeated measures, we input the same AVI answers from the participants twice into the LLMs to get their personality and interview performance (i.e., same participants, same content, the input does not change). BoW-SBERT is excluded from the repeated measures because discriminative models output the same values if the input does not change ($corr = 1$). In the repeated measures, both GPT-3.5 and GPT-4 showed significant correlations for Extraversion, Conscientiousness, and overall interview performance. The correlation was stronger in GPT-4 for all three factors compared to GPT-3.5. In summary, GPT-4 appears to be the most reliable model for the two traits and interview performance, showing the highest correlations in the experiments for repeated measures and test-retest. GPT-3.5 is less reliable, especially for Conscientiousness in the test-retest experiment. BoW-SBERT shows strong reliability for Extraversion and Conscientiousness in the test-retest experiment but is less reliable for overall interview performance.

TABLE 1
The correlation (Pearson r) between AI-predicted values in the repeated measures and test-retest experiment (n = 685 for T1 and n = 145 for T2); $*p < 0.05$;$**p < 0$ 0.01; Boldfaced indicates the highest correlation across all three models.

	Repeated measures		
	Extraversion	**Conscientiousness**	**Interview**
GPT-3.5	0.68**	0.48**	0.75**
GPT-4	**0.79****	**0.61****	**0.83****
	Test-retest		
GPT-3.5	0.48**	0.19*	0.78**
GPT-4	**0.59****	0.39**	**0.85****
BoW-SBERT	0.58**	**0.40****	0.59**

Discussion: The average reliability of LLMs for personality recognition is similar (GPT-4) or worse (GPT-3.5) than the baseline model. Both the correlation of repeated measures and test-retest for GPT-3.5 is much lower than the recommended values for job selection (i.e., $r = 0.80$ [54]). The main reason for that is the generative nature of LLMs. The LLMs learn to predict a sequence of words given a preceding sequence of words, effectively modeling the probability distribution of language tasks. Discriminative models, however, learn to model the high-dimensional mappings from input data to output ground truth labels. Thus, when the input data are the same (i.e., repeated measures), the discriminative models will always (except for including random factors in the model, e.g., the dropout layers) output the same results while generative models will generate different outputs every time based on the language distributions they learned.

However, GPT-4 appears to have similar test-retest reliability compared with our baseline model. A reasonable interpretation of this finding is that the Reinforcement Learning from Human Feedback (RLHF) is more effectively used in GPT-4 compared with GPT-3.5. RLHF can help improve the consistency and reliability of the model's outputs by allowing it to learn from human feedback [9]. Thus, we observed improvement both in repeated measures and test-retest reliability.

For the predictions of overall interview performance, we found that, although the construct validity is very low ($R^2 < 0$), the reliability of the predictions is much higher ($r \geq 0.80$) than the predictions for personality traits. In addition, the test-retest correlations for LLMs on overall interview performance are also higher than our baseline method. Since the model structure and RLHF paradigms of LLMs are close-sourced and unknown to the public, we cannot analyze whether LLMs rate interview performance scores and personality traits differently. However, we assume LLMs may know that reliable ratings of job-related competencies are necessary. Therefore, LLMs may implement stronger constraints on predicting interview performance scores using RLHF. These strong constraints may not lead to the accurate representation of distributions of specific dataset. That is why the predicted ratings cannot explain the variance of the job-related competences and overall interview performance rated by the human recruiters. Thus, we believe that additional fine-tuning (with data and annotated labels) should be implemented before using the LLMs for job-related predictions.

4.3 Fairness

For the comparison between observer-reported and self-reported ratings, the external human observers significantly increased gender and attractiveness biases when rating Extraversion but not Conscientiousness. All AI models and human raters (observers and self) favored women, whose scores were higher on both traits, as indicated by the negative Cohen's d-values. GPT-3.5 increased this gender bias on Extraversion compared to self-reported ratings. However, GPT-4 did not significantly increase or decrease the gender differences compared with both observer-reported and self-reported ratings. The correlations with age and educational background were generally weak and not significant across all AI models and human raters. For attractiveness, all three AI models significantly decreased the biases in observer-reported ratings, except for the Conscientiousness ratings for BoW-SBERT.

Discussion: Our results indicated that AI models can reproduce and sometimes further increase biases existing in human observers. For example, BoW-SBERT increased the gender bias in Extraversion compared to self-reported ratings. In comparison, the LLMs attenuated those differences to some extent (compared to BoW-SBERT). These results indicate that discriminative AI models trained on a comparatively small dataset can suffer from existing group differences in the dataset. BoW-SBERT was trained and validated using our dataset. Hence, it learned the probability distributions on the dataset to minimize the loss (i.e., R^2 in our case). If the training data contain bias, the probability distributions the model learned will reflect this same bias. The LLMs (e.g., GPT-4), however, are pre-trained on publicly available data, which are much larger than the data we collected. Thus, the predictions they make contain less bias than our baseline method.

However, it is not correct to say that LLMs are inherently fair for personality assessment in AVIs. For example, Lee & Ashton [55] showed that women typically score significantly higher on

TABLE 2
Bias analysis for gender, age, attractiveness, and educational background. For gender, Cohen's d is listed. Positive values refer to higher scores for men; negative values refer to higher scores for women. For the other three variables, Pearson r is listed. Boldfaced values indicate that the bias of AI rated values is statistically significant ($*p < 0.05, **\,p < 0.01$) compared to human observer-reported (+ -, increase or decrease) or self-reported (↑↓, increase or decrease) ratings using the linear mixed-effects model. For example, for Extraversion rated by GPT-3.5 ($-0.374\uparrow*$), the bias significantly (*) increased compared with self-reports (↑), but did not differ from the bias in observer-reports (no + or -).

AI ratings					
Models	Factor	Gender	Age	Attractiveness	Education
GPT-3.5	Extraversion	**-0.374 ↑***	-0.035	**0.101 -****	0.065
	Conscientiousness	**-0.308 -***	0.043	**0.017 -****	0.143
GPT-4	Extraversion	-0.312	0.008	**0.118 -****	0.074
	Conscientiousness	-0.395	0.035	**0.068 -***	0.113
BoW-SBERT	Extraversion	**-0.443 ↑***	-0.070	**0.131 -****	0.137
	Conscientiousness	-0.684	-0.041	0.162	0.207
Human ratings					
Raters	Factor	Gender	Age	Attractiveness	Education
Observers	Extraversion	**-0.311 ↑****	-0.060	**0.246 ↑****	0.067
	Conscientiousness	-0.479	-0.011	0.154	0.160
Self	Extraversion	-0.101	-0.004	0.115	0.097
	Conscientiousness	-0.423	0.073	0.063	0.106

Honesty-Humility ($d = -0.37$) and Emotionality ($d = -0.84$), while there are no significant gender differences in Extraversion ($d = -0.08$). Yet, we found that GPT-3.5 significantly increased gender differences in Extraversion compared with self-reported ratings. Furthermore, based on the ratings of GPT-4, older candidates received higher Extraversion ratings, while Ashton & Lee [56] found that three out of four Extraversion facets show an upward age-related trend. In general, while LLMs have the potential to decrease some bias by virtue of their extensive pre-training on diverse datasets, they can still reflect and even increase certain biases present in their training data. Therefore, the utilization of LLMs on specific applications (e.g., personality assessment) should be accompanied by appropriate bias-mitigation strategies and fairness checks.

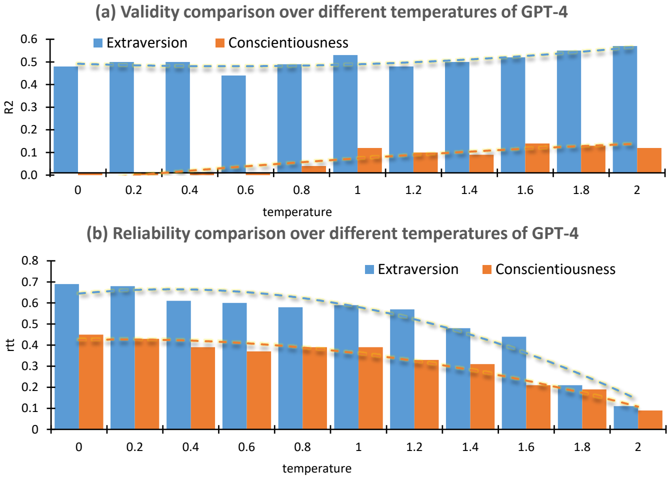


Fig. 4. The validity (indicated by R^2 compared with observer reports) and reliability (indicated by r_{tt} of test-retest experiments) comparison between different values of temperature of GPT-4

4.4 The effect of temperature

The temperature parameter of LLMs, ranging from 0 to 2, controls the degree of randomness in the model's predictions. When generating text, the model calculates probabilities for the next word or token based on the input it receives. Temperature affects how these probabilities are used to make the final choice. A lower temperature biases the model towards more probable responses, making its outputs more predictable and less varied. A higher temperature increases the weighting of less probable choices, introducing more variability and creativity [9]. Thus, it is important to examine how the temperature parameter affects the construct validity and reliability for LLM-based personality recognition.

As shown in Fig 4, the R^2 for Extraversion generally increases when the temperature increases. For Conscientiousness, the model cannot explain the variance when the temperature is smaller than 1, indicating by a negative R^2 compared with observer reports. When the temperature is greater than 1, the construct validity appears to fluctuate with temperature changes.

The correlation for the test-retest experiments for Extraversion shows a decreasing trend as the temperature increases. This suggests that while the model outputs become more valid in terms of capturing the essence of Extraversion at higher temperatures, the reliability of these predictions decreases, indicating less consistency over time. Similarly, the r_{tt} for Conscientiousness also decreases with higher temperatures, which suggests a decline in the reliability of predictions for this trait. As the temperature rises, the outputs become less consistent in rating both Extraversion and Conscientiousness.

Discussion: Most of the previous works [31], [45], [46] have set the temperature to 0 to produce more deterministic and focused responses. However, previous works [57], [58] have also found that setting the temperature to 0 can still lead to variation in LLMs responses, which means it cannot guarantee the unique mapping between inputs and outputs. In addition, lower temperature settings can also limit the capacity of the model by introducing an undesirable clustering of ratings around specific values (the phenomenon of" *streaking*" defined by Han et al. [59]) Thus, the ideal temperature value is dependent on the specific use case [58] and has not yet been defined for the personality rating.

Our findings suggest that there is a trade-off between the diversity of the model's outputs and the stability of personality trait predictions. Increasing the "creativity" (temperature) of the model may lead to more accurate but not reliable predictions. Researchers using LLMs for personality assessments must consider this trade-off when choosing temperature settings.

4.5 Candidate Selection

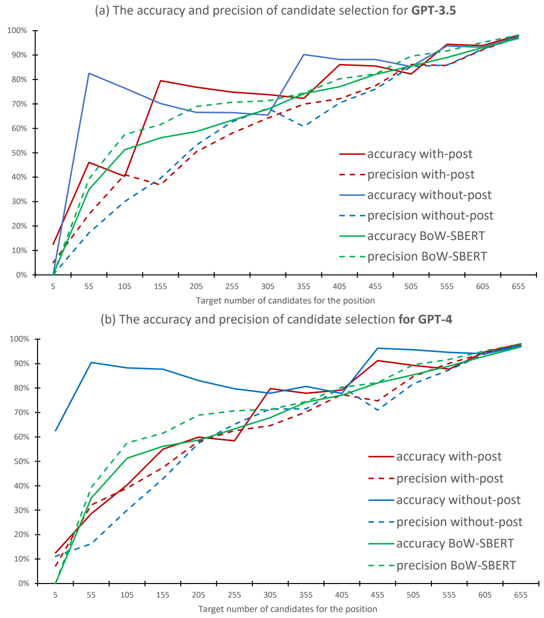


Fig. 5. The accuracy and precision of GPT-3.5 (a) and GPT-4 (b) for candidate selection. "with-post" indicates that we input the job post (description about traineeship position) into the prompt. "without-post" means that we did not input the job post into the prompt. The accuracy and precision of our baseline method (BoW-SBERT) are also plotted for comparison.

Since both GPT-3.5 and GPT-4 cannot explain the variance in job-related competences and overall interview performance, we further explored their ratings on the interview scores by a simulated candidate selection experiment. Fig 5 shows the accuracy and precision for candidate selection by GPT-3.5 and GPT-4 (using the interview performance scores from human recruiters as ground truth). Furthermore, for exploratory reasons, we also ran an experiment to compare the selection results with and without the job descriptions and requirements (i.e., the job post, in Appendix C) as a part of the prompt. As shown in Fig 5, the accuracy and precision of LLMs for candidate selection are low ($< 10\%$) when the position requires only a few (< 10) candidates. Both the accuracy and precision increased when the number of required candidates increased. However, the increase in precision is lower than that in accuracy. Furthermore, the precision of BoW-SBERT is higher than that of the two LLMs. Thus, compared with our task-specific baseline model, both GPT-3.5 and GPT-4 include a large number of unqualified candidates (low precision) in the selection when the number of required candidates increases.

Regarding the comparison between GPT-3.5 and GPT-4, GPT-4 showed a better performance by including fewer unqualified candidates in the selection (the precision and accuracy curves are more similar). GPT-3.5 achieved better accuracy by predicting more candidates with similar and high scores. Thus, the precision of GPT-3.5 is lower. However, when excluding the job post from the prompt, the precision of both LLMs increased more slowly, indicating that the LLMs predict more similar scores and include more unqualified candidates in the selection.

Discussion: In our study, it was observed that the LLMs could not effectively identify the top-performing candidates in a given pool. This deficiency was particularly pronounced when the candidate pool was relatively small. Specifically, it was only when the number of candidates exceeded 200 that the model's accuracy and precision metrics rose above 50%. This can be attributed to the tendency of LLMs to assign similar and high scores to a large number of candidates.

This tendency became even more pronounced under conditions of information scarcity. When the job post (the description of the required skills and qualifications) is not included in the prompt, LLMs tended to assign high scores to an even greater number of candidates (as shown in Fig 6 (a-e)). Our findings suggest that LLMs have the tendency to function as "*overly lenient recruiters*" when they lack sufficient information for their ratings (see the comparison between with and without post in Fig 6 (f)). This approach leads to the inclusion of a surplus of candidates deemed suitable for the position, which can result in an inefficient recruitment process due to the high number of false positive selection decisions. Therefore, while LLMs hold promise for automating aspects of the recruitment process, further refinements are necessary to increase their precision and ability to differentiate among candidates, particularly in situations of limited information.

4.6 Effect of meta-information

Fig 7 demonstrates the comparison of R^2 between different levels of meta-information. For the comparison between mean observer-reported ratings, we found that GPT-3.5 and GPT-4 have similar trends. For Extraversion, the R^2 is highest when no meta-information is provided, followed by the mean of facets and with meta-information of factors. For Conscientiousness, the higher the level of meta-information that is provided to the model, the higher the R^2 the LLMs can achieve. When no meta-information is provided, the R^2 for Conscientiousness is about zero for both GPT-3.5 and GPT-4, indicating their ability to explain variance in Conscientiousness is low.

For the comparison between self-reported ratings, we found that GPT-3.5 failed to explain variance in self-reported Conscientiousness in all three conditions (with-facet, with-factor, without-meta). It also failed to explain the variance of self-reported Extraversion when we obtain the factors by the mean of facets. For GPT-4, the R^2 is positively correlated with the level of meta-information provided to the model. However, it also failed to explain the variance in self-reported Conscientiousness when the information of facets is not provided to the model ($R^2 < 0$ for

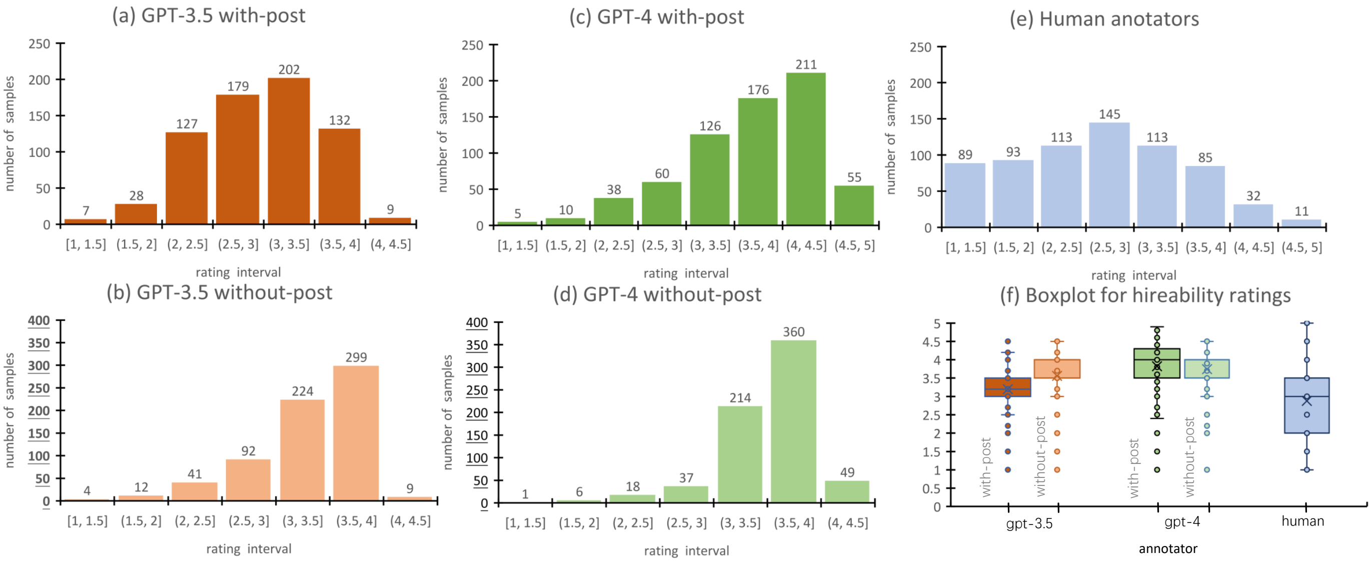


Fig. 6. Histogram (a-e) and boxplot (f) of overall interview performance for GPT-3.5 (a,b), GPT-4 (c,d), human annotators(e).

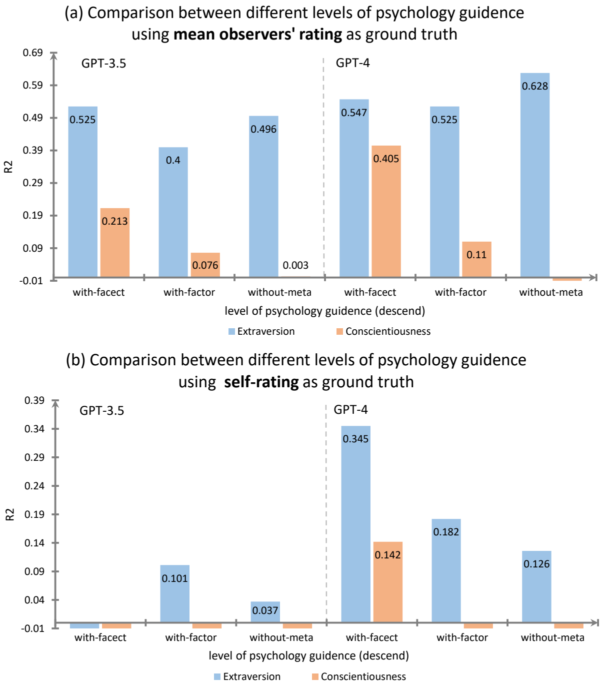


Fig. 7. R^2 comparison between different levels of psychological guidance: "with-facet" means that LLMs know the corresponding personality facets of AVI questions and personality factors are obtained by mean of the facets. "with-factor" means that LLMs know the corresponding personality factors of AVI questions. "without-meta" means no meta information is provided.

with-factor, without-meta).

Discussion: Our results indicate that meta-information can indeed help LLMs perform more accurate personality assessments. This is especially true for Conscientiousness, which was also the most difficult trait to predict in general. Compared to broader personality factors, facets are more specific and fine-grained. These facets might not be overtly observable in an individual's responses during AVIs, but they could be subtly hinted at by LLMs. When the facets are available, they guide the LLMs to pick up on these subtleties, yielding a more thorough analysis of the responses. This conclusion aligns with what we have found in terms of rating validity of other non-activated traits: LLMs may not know how to identify the subtleties from relevant questions to rate them in an accurate manner. Thus, their ratings have negative R^2 values, indicating a poor fit with human annotators. However, when comparing LLMs' performance without meta-information to observer-reported Extraversion, LLMs show a higher R^2 than with meta-information. The verbal cues associated with Extraversion may be prevalent in all questions. Conversely, the verbal cues of Conscientiousness are more subtly embedded, present only in questions that activate this aspect of personality. Thus, LLMs might "over-focus" on specific personality factors, especially in the absence of psychological guidances (the meta-information), leading to an overfitting issue. This means that it may fit the specific traits too closely, capturing the random noise along with the actual personality trait.

In conclusion, the meta-information significantly aids LLMs in achieving better, more balanced performance across different personality factors. Like human annotators, LLMs show similar patterns when annotating personality traits from interviewees' answers. However, unlike humans, LLMs do not genuinely understand the content and context of the responses they analyze. For example, when the information about facets is available for GPT-3.5, it failed to explain variance in self-reported personality traits. Thus, the meta-information or psychological guidance acts as a more substantial constraint on them.

4.7 Linguistic analysis

4.7.1 Personality traits

Fig 8 (a) shows the averaged similarity scores (range [0,1]) for each personality factor. We found that when rating Extraversion and Conscientiousness, the similarity scores for these two factors substantially surpass those of others. Thus, the verbal explanation of LLMs for rating Extraversion and Conscientiousness is coherent with the theoretical description of these two personality factors, as captured by the individual items of the HEXACO-60.

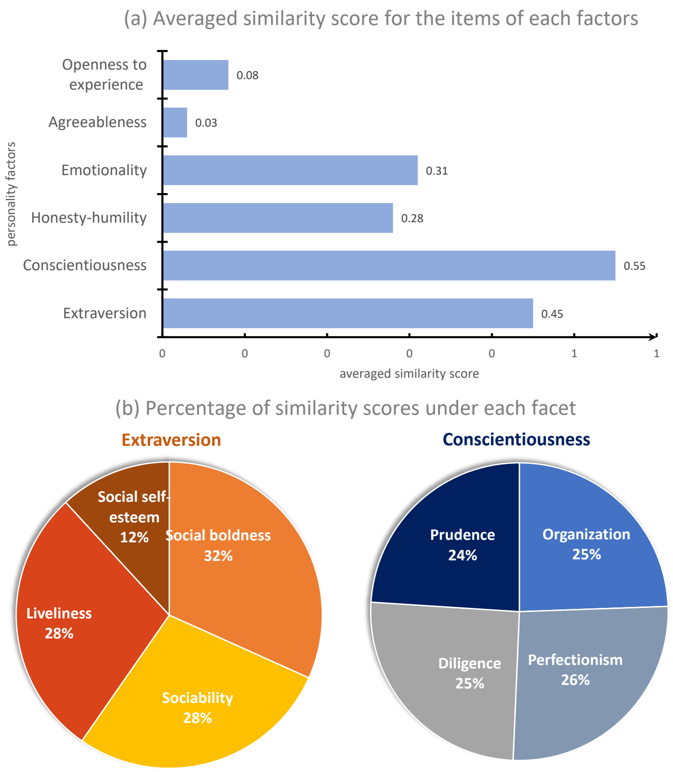


Fig. 8. Averaged similarity scores for the items of personality factors (a) and the percentage of similarity scores for each facet under one factor (b). Both scores are obtained through the task of rating Extraversion and Conscientiousness using GPT-4. The percentage is calculated by dividing the average of similarity scores under each facet by the sum of the average scores under the factor the facet belongs to.

Fig 8 (b) demonstrates the percentage of similarity scores for each facet under one factor. The higher percentage the facet has, the more similar is the verbal explanation to the items of that facet. Here we found that Social Boldness (32%), Sociability (28%), and Liveliness (28%) are considered to have more weight when GPT-4 rates Extraversion. For Conscientiousness, all four facets have similar weights (around 25%). In general, the LLMs assign similar weights for each facet when they predict the corresponding personality factors. The results also indicate that it is easier for LLMs to assess external behavior-based facets (e.g., Sociability) than internal cognitive-based facets (e.g., Social self-esteem). Thus, they seem to rely less on the internal cognitive-based facets when they rate the personality factors.

Discussion: Our results suggest that LLMs are capable of associating specific linguistic patterns with these personality traits in a way that aligns with the HEXACO-60 items related to Extraversion and Conscientiousness. Therefore, when LLMs provide verbal explanations for these personality traits, they are not merely creating plausible-sounding responses; they are producing explanations that are substantively consistent with established psychological theories and research. For example, an explanation for high Extraversion (4.0) and Conscientiousness (4.5) is "*The individual consistently exhibited traits related to both Extraversion and Conscientiousness throughout their responses. They appear to be outgoing, cheerful, and inclusive in social situations (Question 1) and are responsive to others in group discussions (Question 3), indicating a high level of Extraversion. They also consistently demonstrated organization, attention to detail, and a willingness to put in extra effort to achieve success (Questions 2, 4, and 6), indicating a high level of Conscientiousness.*" This explanation highlights the traits detected for social situations ("outgoing, cheerful, and inclusive") and the meticulousness and commitment in task-oriented scenarios ("organization, attention to detail, and willingness to put in extra effort"). Thus, it rates the two personality traits based on the observed linguistic patterns and their alignment with the HEXACO-60 framework.

However, the disparity between externally observable behaviors (such as Sociability) and internal cognitive processes (such as Social Self-Esteem) hints at a crucial limitation of LLMs. They can more readily identify and analyze behaviors that manifest externally and are therefore more visible in linguistic data. However, they may struggle with traits that represent internal thought processes or feelings, as these are not always as readily observable or describable in language. The model's training data, largely consisting of text from the internet, may not adequately represent these more nuanced, subjective aspects of human psychology. This insight underscores the need for caution and a nuanced understanding of human personality when using LLMs to assess personality factors. While LLMs show promise in associating language patterns with personality traits, they may not be equally adept at capturing all facets of human personality.

4.7.2 Interview performance

Table 3 shows the similarity scores (also compared with HEXACO-60 items) of each personality factor and facet when GPT-4 rates job-related competences and overall interview performance. The verbal explanation is more similar to the items of Extraversion and Conscientiousness but not to the items of Emotionality and Agreeableness. For the facets, Social Boldness, Sociability, and Liveliness (Extraversion) as well as all four facets of Conscientiousness have similarity scores above 0.5, indicating that they have a higher weight when GPT-4 rates job-related competences and overall interview performance.

Discussion: Our results are in line with earlier research on human annotators showing that Extraversion and Conscientiousness are valid personality traits for analyzing workplace behaviors [39]. Honesty-Humility and Openness to Experience are also strong predictors of workplace deviance, although weaker than Extraversion and Conscientiousness [60]. Thus, LLMs seem to demonstrate a similar emphasis on personality traits when rating job-related competencies and overall interview performance. The alignment with human annotators' focus on these traits implies that LLMs may have learned the associations between these traits and job performance from the training data.

In terms of specific facets, the high similarity scores for Social Boldness, Sociability, Liveliness, Prudence, Diligence, Perfectionism, and Organization suggest that these aspects of personality are particularly influential in GPT-4's ratings. These facets align with important qualities in the workplace: Social Boldness, Liveliness, and Sociability (necessary for teamwork, leadership, and social interaction), Diligence and Perfectionism (aspects of conscientious work ethic), as well as Organization and Prudence (ability for efficient planning and decision making).

Notably, GPT-4 does not appear to place much emphasis on Emotionality and Agreeableness. This could be due to the nature of these two factors. Emotionality includes facets like fearfulness,

TABLE 3
The similarity scores under each personality factor and facet when GPT-4 rates job-related competences and overall interview performance scores. The boldfaced portion signifies that the value is greater than the third quartile of all the similarity scores.

factors	score	facets	score
Honesty humility	0.28	Sincerity	0.37
		Fairness	0.07
		Greed Avoidance	0.37
		Modesty	0.38
Emotionality	0.31	Fearfulness	0.13
		Anxiety	0.43
		Dependence	0.37
		Sentimentality	0.37
Extraversion	**0.45**	Social Self-Esteem	0.22
		Social Boldness	**0.59**
		Sociability	**0.52**
		Liveliness	**0.53**
Agreeableness	0.03	Forgivingness	0.00
		Gentleness	0.09
		Flexibility	0.00
		Patience	0.00
Conscientiousness	**0.55**	**Organization**	**0.54**
		Diligence	**0.56**
		Perfectionism	**0.58**
		Prudence	**0.53**
Openness to experience	0.08	Aesthetic Appreciation	0.00
		Inquisitiveness	0.00
		Creativity	0.12
		Unconventionality	0.14

anxiety, and dependence. Agreeableness, on the other hand, encompasses traits such as empathy, cooperation, and a willingness to maintain positive relations with others. These characteristics are valuable and usually activated by interpersonal interactions. However, the AVIs are one-way video interviews that do not contain this interaction. Thus, GPT-4 may not have enough information to rate job-related competence and overall interview performance scores based on the answers of AVIs.

5 Discussion

Our experiments suggest that LLMs can provide valid, reliable, and fair predictions for AP-AVI to some extent, particularly for GPT-4 (**RQ1**). LLMs achieve higher or similar construct validity with zero-shot training data compared with the baseline method (BoW-SBERT) which is trained on our dataset. The test-retest reliability of GPT-4 is similar to BoW-SBERT for rating personality traits. The LLMs also reproduce or attenuated most of the existing biases when predicting personality traits. However, the LLMs have uneven performance across different personality traits. While GPT-4 achieved a higher R^2 in most of the cases, it explained less or failed to explain variance in observer-reported and self-reported Conscientiousness, respectively. In addition, LLMs cannot explain variance in job-related competences and overall interview performance rated by professional recruiters. They have the tendency to make skewed predictions and assign high scores to an even greater number of candidates. The repeated-measure and test-retest correlation of LLMs is also lower than the test-retest recommended values for job selection (i.e., $corr = 0.8$ [54]), which means they cannot be directly used for job-related applications. Finally, LLMs can still reflect and even increase certain biases (i.e., gender for GPT-3.5). The above limitations underscore the need for careful validation and potential fine-tuning when using LLMs for human-centric applications to ensure that they can accurately measure and predict the constructs of interest.

We also find that LLMs do show some similarity with human annotators (**RQ2**). The psychological guidance of interview questions has proven beneficial for LLMs in facilitating improved and more uniformly distributed performance across different personality traits, similar to how it benefits human annotators. The linguistic comparison between the LLMs' explanation and HEXACO-60 items shows that LLMs produce explanations that are in line with established psychological theory. However, LLMs are only good at identifying traits by externally observable behaviors. Thus, their capacity to predict traits associated with internal cognitive processes, such as Social Self-Esteem, is comparatively limited.

In conclusion, our study validated the advantages of using LLMs over developing a classical deep learning system for personality recognition. One of the unique challenges with applying deep learning to personality recognition, compared to other applications, is the extensive burden of annotation required to amass an adequate dataset for training the model. In our case, the process of annotation demanded that 31 annotators spent more than 1,400 hours on this task, and it cost more than 10,000 euros to annotate 685 participants. Typically, researchers are compelled to gather such voluminous annotations to train credible deep-learning models. Such models may not be generalizable to different datasets which could be collected via different platforms, experimental paradigms, and scales (i.e., the problem of generalizability in personality recognition [61]). The LLMs, however, necessitated less than 4 hours and 50 euros (GPT-4) for the entire annotation process, demonstrating an impressive zero-shot performance. Thus, the vast data utilized for the pre-training of LLMs bolsters their generalizability, rendering them more resilient across diverse application scenarios. However, the problems of uneven performance across different traits, insufficient test-retest reliability, and the presence and increase of certain biases compromise their reliability and fairness. Therefore, while LLMs have demonstrated significant potential for AP-AVIs, further research is needed to address these challenges, particularly in scenarios where high-stakes (e.g., job-related) decisions are being made.

6 Limitation and future Work

While our study provided deep insights into the capability of LLMs in terms of AP-AVI, our research has certain limitations. Firstly, our empirical assessment was confined to the evaluation of GPT-3.5 and GPT-4, and not other LLMs. While the evaluation of GPT-3.5 and GPT-4 provided valuable insights, it is important to recognize that other LLMs might offer different perspectives. Future research should therefore aim to encompass a broader range of LLMs to ensure a more comprehensive understanding of their capabilities. In the future, we also plan to expand our research by incorporating audio-visual data and advanced temporal analysis techniques into our models. In addition, we only employed one baseline method (BoW-SBERT) for comparison.

Subsequent studies would benefit from integrating diverse state-of-the-art methods to replicate our conclusions. In the future, we also plan to expand our research by incorporating audio-visual data and advanced temporal analysis techniques into our models.

For the prompting optimization, we utilized one specific psychological guidance for comparison. Given the swift advancements in prompt engineering, future exploration should consider diversifying guidance methods for the most optimized performance for AP-AVI tasks. For example, in-context learning has potential advantages and effectiveness in enhancing the validity of LLM-based personality assessment. In the future, we will explore the use of adaptive prompting techniques, where the context is dynamically adjusted based on the ongoing interaction. We also plan to customize LLMs through fine-tuning of datasets annotated by psychological experts and professional recruiters for better AP-AVI performance. The human guidance and fine-tuning can enable more reliable and trustworthy LLMs for human-centric applications.

For the results of repeated measurement, we only input the same content twice to the model and calculated the correlation coefficients between the two outputs. The purpose of this is to find out how reliable the LLMs are with the same output. Thus, we did not define what is the final output of the model. In the future, we will extend this methodology by conducting repeated measurements with a larger number of iterations, not just twice. This would allow for a more robust statistical analysis of the model's output consistency and reliability. Additionally, exploring different types of input content and varying the length and complexity of these inputs could also provide deeper insights into how these factors influence the consistency of LLM outputs.

Finally, we will develop bias-mitigation strategies as LLMs can reproduce or increase biases existing in human observers. The key strategy to mitigate biases is to incorporate both human experts and LLMs in the loop. This involves establishing a system where psychologists can review and provide feedback on the personality assessments made by LLMs. Additionally, conducting comprehensive user testing with a broad spectrum of users will also help to identify biases and fairness issues that might not be immediately evident from the personality scores rated by LLMs.

7 Conclusion

This study presents a comprehensive evaluation of two widely used Large Language Models (LLMs), GPT-3.5 and GPT-4 for personality and interview performance ratings from interviewees' answers in Asynchronous Video Interviews (AVIs). We evaluated the validity, reliability, and fairness of LLMs toward the task and compared their rating pattern with human annotators. Our results show that LLMs can provide relatively valid, reliable, and fair predictions for personality traits and their rating patterns show some similarity with human annotators. However, LLMs also show uneven performance levels, insufficient reliability, and the presence and increase of certain biases. When rating interview performance from AVIs, LLMs also have the tendency to assign high scores to an even greater number of candidates. Consequently, researchers should exercise caution and consider incorporating guidance from human experts, particularly in situations where critical decisions, such as those related to employment, are being made.

References

[1] E.-R. Lukacik, J. S. Bourdage, and N. Roulin, "Into the void: A conceptual model and research agenda for the design and use of asynchronous video interviews," *Human Resource Management Review*, vol. 32, no. 1, p. 100789, 2022.

[2] A. I. Huffcutt, C. H. Van Iddekinge, and P. L. Roth, "Understanding applicant behavior in employment interviews: A theoretical model of interviewee performance," *Human Resource Management Review*, vol. 21, no. 4, pp. 353–367, 2011.

[3] J. L. Pletzer, J. K. Oostrom, and R. E. de Vries, "HEXACO personality and organizational citizenship behavior: A domain-and facet-level meta-analysis," *Human Performance*, vol. 34, no. 2, pp. 126–147, 2021.

[4] Y. Sun, F. Zhuang, H. Zhu, Q. Zhang, Q. He, and H. Xiong, "Market-oriented job skill valuation with cooperative composition neural network," *Nature communications*, vol. 12, no. 1, p. 1992, 2021.

[5] L. Hickman, N. Bosch, V. Ng, R. Saef, L. Tay, and S. E. Woo, "Automated video interview personality assessments: Reliability, validity, and generalizability investigations." *Journal of Applied Psychology*, vol. 107, no. 8, p. 1323, 2022.

[6] H.-Y. Suen, K.-E. Hung, and C.-L. Lin, "Intelligent video interview agent used to predict communication skill and perceived personality traits," *Human-centric Computing and Information Sciences*, vol. 10, pp. 1–12, 2020.

[7] HireVue. (2022) Explainability statement (white paper). [Online]. Available: https://webapi.hirevue.com/wp-content/uploads/2022/04/HV_AI_Short-Form_Explainability_1pager.pdf

[8] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[9] OpenAI, "Gpt-4 technical report," *ArXiv*, vol. abs/2303.08774, 2023.

[10] A. V. Ganesan, Y. K. Lal, A. H. Nilsson, and H. A. Schwartz, "Systematic evaluation of gpt-3 for zero-shot personality estimation," *arXiv preprint arXiv:2306.01183*, 2023.

[11] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[12] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Bias and fairness in multimodal machine learning: A case study of automated video interviews," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 268–277.

[13] H. Rao, C. Leung, and C. Miao, "Can chatgpt assess human personalities? a general evaluation framework," *arXiv preprint arXiv:2303.01248*, 2023.

[14] M. A. Ricci Lara, R. Echeveste, and E. Ferrante, "Addressing fairness in artificial intelligence for medical imaging," *nature communications*, vol. 13, no. 1, p. 4581, 2022.

[15] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: a review," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–38, 2022.

[16] D. A. Broniatowski *et al.*, "Psychological foundations of explainability and interpretability in artificial intelligence," *NIST, Tech. Rep*, 2021.

[17] M. C. Ashton and K. Lee, "Empirical, theoretical, and practical advantages of the HEXACO model of personality structure," *Personality and social psychology review*, vol. 11, no. 2, pp. 150–166, 2007.

[18] J. M. Beus, L. Y. Dhanani, and M. A. McCord, "A meta-analysis of personality and workplace safety: addressing unanswered questions." *Journal of applied psychology*, vol. 100, no. 2, p. 481, 2015.

[19] M. C. Ashton and K. Lee, "The HEXACO–60: A short measure of the major dimensions of personality," *Journal of personality assessment*, vol. 91, no. 4, pp. 340–345, 2009.

[20] L. Cabrera-Quiros, E. Gedik, and H. Hung, "Multimodal self-assessed personality estimation during crowded mingle scenarios using wearables devices and cameras," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 46–59, 2019.

[21] J. C. J. Junior, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. Van Gerven, R. Van Lier *et al.*, "First impressions: A survey on vision-based apparent personality trait analysis," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 75–95, 2019.

[22] M. C. Ashton, K. Lee, M. Perugini, P. Szarota, R. E. De Vries, L. Di Blas, K. Boies, and B. De Raad, "A six-factor structure of personality-descriptive adjectives: solutions from psycholexical studies in seven languages." *Journal of personality and social psychology*, vol. 86, no. 2, p. 356, 2004.

[23] J. L. Pletzer, J. K. Oostrom, M. Bentvelzen, and R. E. de Vries, "Comparing domain-and facet-level relations of the HEXACO personality model

with workplace deviance: A meta-analysis," *Personality and Individual Differences*, vol. 152, p. 109539, 2020.

[24] J. H. Manson, "Life history strategy and the HEXACO personality dimensions," *Evolutionary Psychology*, vol. 13, no. 1, p. 147470491501300104, 2015.

[25] N. Aghababaei and A. Arji, "Well-being and the HEXACO model of personality," *Personality and Individual Differences*, vol. 56, pp. 139–142, 2014.

[26] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Transactions on Affective Computing*, 2022.

[27] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *arXiv preprint arXiv:2307.03109*, 2023.

[28] A. Koutsoumpis, J. K. Oostrom, D. Holtrop, W. Van Breda, S. Ghassemi, and R. E. de Vries, "The kernel of truth in text-based personality assessment: A meta-analysis of the relations between the big five and the linguistic inquiry and word count (liwc)." *Psychological Bulletin*, vol. 148, no. 11-12, p. 843, 2022.

[29] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt," *IEEE Intelligent Systems*, vol. 38, no. 2, pp. 15–23, 2023.

[30] M. Amin, E. Cambria, and B. Schuller, "Can chatgpt responses boost traditional natural language processing?" *arXiv preprint arXiv:2307.04648*, 2023.

[31] Y. Ji, W. Wu, H. Zheng, Y. Hu, X. Chen, and L. He, "Is chatgpt a good personality recognizer? a preliminary study," *arXiv preprint arXiv:2307.03952*, 2023.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[33] P. Ritala, M. Ruokonen, and L. Ramaul, "Transforming boundaries: how does chatgpt change knowledge work?" *Journal of Business Strategy*, 2023.

[34] M. M. Amin, R. Mao, E. Cambria, and B. W. Schuller, "A wide evaluation of chatgpt on affective computing tasks," *arXiv preprint arXiv:2308.13911*, 2023.

[35] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman, "Automatic personality assessment through social media language." *Journal of personality and social psychology*, vol. 108, no. 6, p. 934, 2015.

[36] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 400–418.

[37] N. Tomašev, J. Cornebise, F. Hutter, S. Mohamed, A. Picciariello, B. Connelly, D. C. Belgrave, D. Ezer, F. C. v. d. Haert, F. Mugisha *et al.*, "AI for social good: unlocking the opportunity for positive impact," *Nature Communications*, vol. 11, no. 1, p. 2468, 2020.

[38] S. Han, H. Huang, and Y. Tang, "Knowledge of words: An interpretable approach for personality recognition from social media," *Knowledge-Based Systems*, vol. 194, p. 105550, 2020.

[39] J. Oostrom, D. Holtrop, A. Koutsoumpis, W. Van Breda, S. Ghassemi, and R. De Vries, "Applicant reactions to algorithm-versus recruiter-based evaluations of an asynchronous video interview and a personality inventory," *Journal of Occupational and Organizational Psychology*, 2023.

[40] A. Koutsoumpis, S. Ghassemi, J. K. Oostrom, D. Holtrop, W. van Breda, T. Zhang, and R. E. de Vries, "Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning," *Computers in Human Behavior*, p. 108128, 2024.

[41] A. J. Barends and R. E. De Vries, "Noncompliant responding: Comparing exclusion criteria in mturk personality research to improve data quality," *Personality and individual differences*, vol. 143, pp. 84–89, 2019.

[42] J. Levashina, C. J. Hartwell, F. P. Morgeson, and M. A. Campion, "The structured employment interview: Narrative and quantitative review of the research literature," *Personnel Psychology*, vol. 67, no. 1, pp. 241–293, 2014.

[43] P. R. Sackett, C. Zhang, C. M. Berry, and F. Lievens, "Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range." *Journal of Applied Psychology*, 2021.

[44] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[45] T. Yang, T. Shi, F. Wan, X. Quan, Q. Wang, B. Wu, and J. Wu, "Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection," *arXiv preprint arXiv:2310.20256*, 2023.

[46] T. Hagendorff, S. Fabi, and M. Kosinski, "Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt," *Nature Computational Science*, vol. 3, no. 10, pp. 833–838, 2023.

[47] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[48] F. P. Morgeson, M. H. Reider, M. A. Campion, and R. A. Bull, "Review of research on age discrimination in the employment interview," *Journal of Business and Psychology*, vol. 22, pp. 223–232, 2008.

[49] M. R. Barrick, J. A. Shaffer, and S. W. DeGrassi, "What you see may not be what you get: relationships among self-presentation tactics and ratings of interview and job performance." *Journal of applied psychology*, vol. 94, no. 6, p. 1394, 2009.

[50] L. Yarger, F. Cobb Payton, and B. Neupane, "Algorithmic equity in the hiring of underrepresented it job candidates," *Online information review*, vol. 44, no. 2, pp. 383–395, 2020.

[51] P. Borkenau, N. Mauer, R. Riemann, F. M. Spinath, and A. Angleitner, "Thin slices of behavior as cues of personality and intelligence." *Journal of personality and social psychology*, vol. 86, no. 4, p. 599, 2004.

[52] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[53] U. J. Wiersma and R. Kappe, "Selecting for extroversion but rewarding for conscientiousness," *European Journal of Work and Organizational Psychology*, vol. 26, no. 2, pp. 314–323, 2017.

[54] A. Evers, W. Lucassen, R. Meijer, and K. Sijtsma, "Cotan review system for evaluating test quality," *Retrieved January*, vol. 25, p. 2023, 2015.

[55] K. Lee and M. C. Ashton, "Sex differences in HEXACO personality characteristics across countries and ethnicities," *Journal of personality*, vol. 88, no. 6, pp. 1075–1090, 2020.

[56] M. B. Donnellan and R. E. Lucas, "Age differences in the big five across the life span: evidence from two national samples." *Psychology and aging*, vol. 23, no. 3, p. 558, 2008.

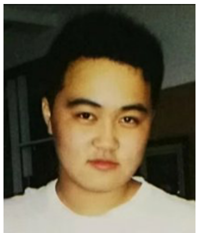
[57] M. Zheng, X. Su, S. You, F. Wang, C. Qian, C. Xu, and S. Albanie, "Can gpt-4 perform neural architecture search?" *arXiv preprint arXiv:2304.10970*, 2023.

[58] F. Antaki, D. Milad, M. A. Chia, C.-É. Giguère, S. Touma, J. El-Khoury, P. A. Keane, and R. Duval, "Capabilities of gpt-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering," *British Journal of Ophthalmology*, 2023.

[59] C. Han, D. W. Kim, S. Kim, S. C. You, J. Y. Park, S. Bae, and D. Yoon, "Evaluation of gpt-4 for 10-year cardiovascular risk prediction: Insights from the uk biobank and koges data," *Iscience*, vol. 27, no. 2, 2024.

[60] J. L. Pletzer, M. Bentvelzen, J. K. Oostrom, and R. E. De Vries, "A meta-analysis of the relations between personality and workplace deviance: Big five versus HEXACO," *Journal of Vocational Behavior*, vol. 112, pp. 369–383, 2019.

[61] L. Tay, S. E. Woo, L. Hickman, and R. M. Saef, "Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining," *European Journal of Personality*, vol. 34, no. 5, pp. 826–844, 2020.



Tianyi Zhang is an associate professor at the School of Biological Sciences and Medical Engineering, Southeast University. Before that, he worked as a postdoc researcher at Vrije Universiteit Amsterdam. He got his PhD degree in Delft University of Technology. He was also associated with the Distributed & Interactive Systems (DIS) group at the national research institute for mathematics and computer science in the Netherlands (CWI). His research interests lie in human-computer interaction, affective computing, and personality recognition.

Antonis Koutsoumpis is a PhD candidate at the Organizational Section of the Experimental and Applied Psychology Department, at the Vrije Universiteit Amsterdam, the Netherlands. He has a background in psychology and his research interests include automatic personality assessment from asynchronous video interviews as well as the verbal and non-verbal behaviors that individuals exhibit depending on their personality traits.

Janneke Oostrom is a Professor of Work & Organizational Psychology at Tilburg University's department of Social Psychology. Her research focuses on understanding and improving psychological assessments, with the goal to make them more predictive of future work behaviors, while reducing discrimination against marginalized groups. She received her Master (2005) and Ph.D. (2010) in Work and Organizational Psychology from the Erasmus University Rotterdam.

Djurre Holtrop is an assistant professor at Tilburg University's department of Social Psychology. He has worked in consulting for psychometric assessments, leading large scale online assessment projects. He completed his PhD at the VU Amsterdam studying the refinement of personality and vocational interest questionnaires. Subsequently, he worked for the University of Western Australia and Curtin University to study the recruitment, motivation, and retention of volunteers. Currently, his research focusses on personnel recruitment and selection and volunteer attraction and engagement.

Sina Ghassemi is a postdoctoral researcher at Vrije Universiteit Amsterdam. He received his PhD in Telecommunication Engineering with specialization in Signal Processing and AI from Politecnico di Torino, Turin, Italy, in 2018. His research interests are in the field of deep learning, signal processing, and affective computing.

Reinout E. de Vries is Full Professor in Organizational Psychology with a chair in 'Personality at Work' at the Vrije Universiteit Amsterdam, the Netherlands. His main areas of interest are the theoretical background, structure, measurement, and effects of personality, leadership, communication styles, and situations.