Mixture-of-Expert Large Language Models for text-based Personality Assessment from Asynchronous Video Interviews

Tianyi Zhang, *Senior Member, IEEE,* Shan Liang*, Wenming Zheng, *Senior Member, IEEE,*
Antonis Koutsoumpis, Janneke K. Oostrom, and Reinout E. de Vries

Abstract—In selection and assessment, Large Language Models (LLMs) are deemed suitable for personality assessment of Asynchronous Video Interviews (AVI) due to their advanced linguistic understanding and semantic interpretation capacities. However, most of the previous works have focused on personality trait *classification* rather than *regression*. Since current LLMs are trained on massive corpora, they are more attuned to text-based structures than to numeric data. The classification-centered approach fails to take into account the fact that personality traits are continuous, instead of categorical, variables. In addition, LLMs suffer from low rating validity due to their tendency to be over-lenient, assigning higher than average personality scores to a large number of individuals (i.e., issues of *Positivity Biases*). To address these challenges, we designed a text-based, two-stage, Mixture-of-Experts based personality assessment framework (*MoE-Personality*) to provide fine-grained personality ratings and regulate the positivity biases of LLMs. The designed model first rates the coarse-grained personality category of the individual (*classification stage*). After that, the model rates fine-grained personality scores by merging the obtained personality category and empirical score ranges of different personality categories. Inspired by the the way human annotators rate personality traits, each stage comprises multiple annotator LLM experts and one aggregator LLM expert to promote validity. Our experiments in two datasets show that the designed framework outperformed open-sourced, medium-sized LLMs (e.g., Llama 3.1-8B, qwen2-7B) and achieved comparable results with close-sourced, large-sized LLMs (e.g., GPT-3.5 and GPT-4).

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

1 Introduction

The rapid development of Artificial Intelligence (AI) technology has markedly impacted traditional processes of job interviews, a change that was further accelerated by the COVID-19 pandemic [1], [2]. Over previous years, Asynchronous Video Interviews (AVIs), a one-way online job interview, where job applicants provide video-based responses to questions presented via their laptops, tablets, or smartphones - have solidified their status as a recognized tool in personnel selection procedures. As a strong predictor of job-related outcomes, personality traits are evaluated in AVIs by vendors to obtain a comprehensive understanding of a candidate's potential fit within the company culture and alignment with job expectations [3], [4], [5]. Although personality traits are frequently manually evaluated by recruiters based on the AVI responses of candidates, AI systems, especially those that use deep learning techniques, are increasingly being used to automatically assess personality traits from AVIs due to their advantages in time and economic efficiency [6], [7]. For example, Hirevue (one of the largest human resources management company) reported that they have "*hosted more than 26 million video interviews and 5 million AI-based candidate assessments*" already in 2022 [8].

Among many well-developed AI models, Large Language Models (LLMs), such as GPT-4 [9], Llamma-3.1-8B [10] and Gemma-2-9B [11], are particularly capable of assessing personality traits from AVIs. Due to their exhaustive training on large datasets, they have reached remarkable depths of understanding human language [12]. Moreover, their zero-shot capabilities allow them to perform accurate assessments without the need to rely on training data from the vendors. This advantage mitigates potential privacy concerns and makes LLMs especially suitable for tasks involving sensitive and private data such as personality traits [13], [14], [15].

Although previous work has shown high performance for LLM-based zero-shot personality assessment, most of this work [15], [16], [17], [18], [19] performed classification instead of regression tasks for personality traits. Classification has long been a common formulation in personality assessment due to factors such as data availability and accessibility, and LLMs are particularly well-suited to handle linguistic structures [9], [12], making categorical predictions. Thus, they are naturally suitable for classification (e.g., 'low', 'medium', 'high') instead of regression tasks (e.g., a Likert-type 5-point scale). However, classification approaches are limited in that (a) they artificially convert continuous variables into categorical, and (b) they lose personality variance information during this conversion. Thus, even though the use of LLMs in classification tasks might be convenient, it is necessary to employ regression approaches to align computational assessment with rigorous psychological measurement standards (e.g., [20], [21]).

In addition, as LLMs are trained through Reinforcement

1. Tianyi Zhang and Wenming Zheng are with the Key Laboratory of Child Development and Learning Science (Ministry of Education), School of Biological Sciences and Medical Engineering, Southeast University, Nanjing, China. emails: {t.zhang, wenming_zheng}@seu.edu.cn
2. Shan Liang is with the Business School of NanJing XiaoZhuang University, Nanjing, China. email:2025013@njxzc.edu.cn
3. R. E. de Vries and A. Koutsoumpis are with the Department of Experimental and Applied Psychology, School of Business and Economics, respectively, Vrije Universiteit Amsterdam, the Netherlands. e-mail: {re.de.vries, A. Koutsoumpis}@vu.nl
4. J. K. Oostrom is with the Department of Social Psychology, Tilburg University, Tilburg, the Netherlands emails:J.K.Oostrom@tilburguniversity.edu
∗ corresponding author

Learning from Human Feedback (RLFH), they tend to assign socially desirable scores across key personality dimensions [12], [13]. This is known as *positivity bias* [12], that is, when there is not enough information available for LLMs to base their personality recognition on, they are more likely to assign higher (instead of average, or lower) personality scores to participants, artificially inflating the scores of the evaluated personality traits. Therefore, it is necessary for LLM-based personality assessment models to integrate some form of regularization mechanisms to mitigate the effect of positivity bias.

To address the aforementioned research gaps, this paper proposes a novel method to assess personality traits employing a text-based, two-stage, Mixture-of-Expert framework (*MoE-Personality*) in fine level of granularity (regression). The designed model first performs coarse-grained personality classification tasks to leverage the text-centric advantages of LLMs, resulting in personality trait categories (i.e., the classification stage). After that, the model evaluates fine-grained personality scores by integrating the identified personality category with empirical score ranges specific to different personality categories (i.e., the regression stage). In this second stage, the identified personality categories are used as benchmarks to regulate the scores and curtail excessive positivity bias.

When psychologists code personality traits of an individual, a typical approach is that multiple annotators independently assess the traits, and those scores are subsequently averaged [20], [22]. This procedure can improve the validity and reduce the subjectivity of personality trait evaluations. In a similar vain, when a multiple-rater approach is applied to LLMs, previous work [23] shows that integrating the annotations from various LLMs can lead to superior performance, compared to relying on a single LLM. Inspired by the experiences from psychology research and previous works from AI community, we propose a MoE module in each stage to promote the validity of personality measurement. In each stage, multiple annotator LLM experts will first independently assess and generate personality ratings based on the prompt, and then one aggregator LLM expert will review and consolidate these individual assessments to get a final, unified rating.

The present work makes the following contributions in the affective computing community:

- We propose *MoE-Personality*, a two-stage, mixture-of-experts LLM framework, designed to rate the personality traits in fine-level of granularity (regression), providing researchers in affective computing and psychologists with a tool to precisely analyze, predict, and validate personality traits with zero-shot training data. This can mitigate privacy issues as it allows for comprehensive and precise personality trait analysis without the direct involvement of sensitive personal data, thereby reducing the risks associated with data misuse or breaches.
- We tested the proposed framework on two AVI datasets, OPVA (n=685) and AVI-6 (n=646), and compared the results with LLM-based and Non-LLM based methods. The results showed that by only using medium-sized LLMs, the designed framework largely promoted the rating validity of LLM-based methods and achieved comparable results with close-sourced, large-sized LLMs such as GPT-3.5 and GPT-4o.
- We conducted experiments that provided us with a better understanding of self-regularization from two stages and the internal mechanism of MoE when rating personality traits. Our ablation study showed that coarse-grained personality categories can indeed offset the effects of positivity bias and promote the validity of personality trait ratings.

2 Related Work

In this section, we first introduce the HEXACO personality model, which is used to model the personality traits in our work. Secondly, we review the literature of personality assessment from Asynchronous Video Interviews (AVIs) to highlight the technologies used to quantify personality traits from video data. After that, we review the rapid development of Large Language Models (LLMs). In this part, we discuss both their advantages and shortcomings in the context of personality assessment. Finally, we summarize previous works on the MoE LLMs approach and explore the possibility of this strategy to enhance the validity of personality assessments.

2.1 HEXACO Personality Model

The HEXACO model of personality [24] is a widely adopted framework to assess personality traits [25], [26], comprising of six core dimensions: Honesty-Humility (H), Emotionality (E), eXtraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O). Originating from the lexical tradition and developed as an extension of the Big Five model, HEXACO introduces Honesty-Humility as a distinct factor, while redistributing variance from Neuroticism and Agreeableness into Emotionality, Agreeableness, and Honesty-Humility [27]. Each dimension is further subdivided into four facets, enabling a granular analysis of personality traits. This fine-grained structure enhances the model's predictive capacity across domains such as occupational performance [28], interpersonal relationships [29], and mental health outcomes [30]. Thus, the HEXACO model is employed in this study as the personality model to quantify personality traits.

2.2 Audio-visual personality assessment

Audio-visual personality assessment, which focuses on extracting personality-relevant cues from both visual and acoustic modalities, has emerged as the dominant approach in automatic personality recognition [31], [32], [33]. The availability of large-scale annotated datasets, such as those from the ChaLearn First Impressions challenges [34], has played a crucial role in advancing this paradigm by enabling systematic evaluation of models on standardized benchmarks. These datasets, typically labeled with Big Five trait scores, contain short video clips in which subjects display spontaneous or semi-scripted behaviors. Pervious works on audio-visual personality assessment can be divided into two categories. The earliest approaches [35], [36], [37] rely on handcrafted descriptors, where low-level features are computed from each modality separately before being fed into shallow regressors such as support vector regression or random forests. For example, Gurpinar et al. [35] fused LBP-TOP facial texture features, openSMILE prosodic features, and scene descriptors, achieving a mean accuracy of 0.889 on the ChaLearn 2016 dataset.

The second category encompasses deep learning methods that operate end-to-end, learning hierarchical representations directly from raw frames and audio waveforms. For example, Aslan et al. [38] proposed a multimodal fusion algorithm that extracts features from the scene, the subject's face, and voice in videos. Their approach achieved a mean accuracy of 0.917 on the ChaLearn First

Impressions V2 dataset [39], representing a 0.4% improvement over the best single-modality (face-only) model at 0.913. Moreover, Liao et al. [31] conducted a benchmark analysis for 26 non-verbal-based deep learning networks for personality assessment under a standardized evaluation pipeline and found that non-verbal behaviors contributed differently to predicting various personality traits. For example, visual-only networks [40], [41], [42] in their study achieved up to 0.905 mean accuracy for Extraversion, while audio-only networks [43], [44] performed best for Agreeableness at 0.892, highlighting trait-specific modality dependencies.

In parallel, the broader personality computing community has examined related multimodal recognition tasks in interactive settings. For instance, Giorgi et al. [45] reported the results of the WASSA 2024 shared task on empathy and personality detection in interactions. Their results show that multimodal models incorporating both audio and visual streams improved performance over unimodal baselines in conversational contexts.

Overall, audio-visual methods offer several advantages. They are relatively language-independent, avoid the variability and cultural bias introduced by textual content, and, unlike verbal information that individuals can deliberately control, capture spontaneous paralinguistic and non-verbal behaviors that are considered difficult to control (e.g., voice timbre is biologically determined based on the length of one's larynx). Fusion of multiple non-verbal channels enables the capture of complementary cues, leading to robust improvements over unimodal baselines by 1–2% in mean accuracy [31]. However, as noted in the recent twenty-year review of the field by Celli et al. [46], these systems remain sensitive to dataset conditions such as lighting, background clutter, and demographic imbalance. The visual channel's dominance can bias predictions toward traits that are more overtly expressed in facial behavior such as Extraversion. Gains from adding more modalities tend to plateau beyond a certain point. Addressing these challenges requires not only advances in adaptive fusion strategies, but also exploration of alternative modalities such as text.

Pervious works show [7], [20], [22] that verbal cues are potentially yield more interpretable and transferable personality predictions across diverse contexts. Thus, they may mitigate the visual-audio-specific biases of audio-visual systems, offer richer access to cognitive and affective cues.

2.3 Text-based Personality assessment from AVIs

Asynchronous Video Interviews (AVIs) are designed to capture a candidate's responses to job-related questions and other desirable outcomes, such as job competencies or individual differences [1]. To maximize the accuracy of measuring personality traits from AVIs, researchers typically design interview questions that trigger the desired personality traits, for instance, by directly asking participants to describe situations in which they displayed behaviors related to specific personality traits. These trait-activating interview questions ensure that the responses to AVIs are more likely to reflect the candidate's true personality traits, as the questions are specifically designed to activate these traits [7], [20]. This approach can provide employers with a deeper understanding of candidates' personality traits and potentially lead to better hiring decisions.

Unlike the audio-visual personality assessment methods reviewed in section 2.2, verbal-based methods analyze the content and linguistic features of a candidate's spoken responses to infer personality traits [47]. Due to reliable and consistent relationships between linguistic performance and personality traits [48], [49], many researchers use linguistic features to train text classifiers to assess personality traits. For example, Amoux et al. [50] combined LIWC and MRC linguistic features with X (i.e., Twitter) profile statistics to predict personality traits. For the deep-learning-based methods, Majumder et al. [51] first proposed using one-dimensional convolutional networks for personality analysis. Additionally, Sun et al. [52] proposed a model named C2W2S4PT, which combines the advantages of CNN and RNN to achieve deep embedding from sentence-to-sentence group.

In the context of AVIs, interview questions are often carefully designed to activate personality traits [7], [20], following the premises of Trait Activation Theory [53]. However, in generic online videos (e.g., YouTube) or daily conversations, personality traits are not deliberately activated, something that has implications for the measurement of those traits. More specifically, two previous studies [20], [54] employing a similar multimodal recognition algorithm demonstrated that when personality traits are activated via trait-activating interview questions (in AVIs), the verbal features (i.e., spoken text) of the AVI were the most significant contributors, compared with other modalities [20]. Conversely, in tasks where traits were not activated (in online YouTube videos [34]), audio and visual features exerted the most substantial influence [54].

As previous research shows that verbal features are the most important contributors in assessing personality traits in AVIs, our present work focuses on exploring the activated personality traits in AVIs through linguistic interactions (i.e., through interview questions that have been developed to activate specific personality traits, and through verbal responses). Thus, we use a verbal- (instead of a non-verbal) based method as the primary approach to assess personality traits in AVIs, by analyzing the spoken text of participants.

2.4 LLMs-based personality assessment

Large Language Models (LLMs) are neural networks pre-trained on extensive datasets to perform various natural language processing tasks [55]. LLMs can generate coherent and contextually relevant responses without requiring task-specific training data, which is known as zero-shot performance [9]. This capability makes LLMs particularly useful for text-based personality assessments, as they bypass the need for extensive data collection and model training, and cut costs related to data annotation. Additionally, LLMs enhance privacy protection by eliminating the need to collect and store large amounts of personal data for model training [13], [14], [15].

Due to the strong capabilities of linguistic comprehension and good zero-shot performance of LLMs, many researchers have developed automatic personality assessment algorithms utilizing these models. For example, the study conducted by Wen et al. [12] represented the first comprehensive assessment of the application of LLMs in personality assessment. This research introduced a framework with questionnaire-based prompting strategies to predict the Big Five personality traits. Ji et al. [16] compared ChatGPT with traditional machine learning models (RNN and RoBERTa). They found that LLMs exhibited enhanced inference capabilities for personality traits when employing varied prompting strategies. Peters et al. [17] validated GPT-4's ability to infer personality traits from short-term interactions and found that the complexity of the prompt design significantly affected

the outcomes. Follow-up research [56] assessed the personality and psychological trait inference capabilities of GPT-3.5 and GPT-4 using social media text. In this study, they emphasized the necessity for further refinement and optimization of dynamic interaction and prompt construction for personality assessment and psychological trait inference.

Recently, some of the researchers have also developed fine-tuned LLMs for personality assessment. For instance, Shen et al. [57] introduced PersLLM, a PEFT-based framework that leverages LLaMA to extract high-dimensional personality representations from user-generated text. Their design improves task flexibility compared with full-model fine-tuning. Similarly, Hu et al. [58] proposed an LLM-enhanced text mapping approach that jointly exploits the embedding and generative capacities of LLMs to improve personality classification from textual data. For AVI-based personality assessment, the collection of annotated AVI data is highly resource-demanding [12], [59]. Personality traits are typically assessed by trained psychologists following standardized behavioral coding protocols [33]. Thus, it requires substantial time investment and domain expertise to interpret open-ended interview responses.

Although previous work has developed several LLM-based methods for personality assessment, most of these methods focus on personality classification rather than regression [15], [19], [58]. Given the linguistic and generative nature of LLMs, they are typically designed to categorize personality traits as "high" or "low" rather than providing numeric ratings [16], [60], [61]. These broad categorical ratings fail to capture the full spectrum of individual personality traits due to their lack of granularity [7], [20]. For instance, two individuals with conscientiousness scores of 3.1 and 2.9 could be classified as high and low in conscientiousness, respectively, if the threshold for categorization is set at 3. This can be a serious misclassification issue, as personality traits are normally distributed, and categorizing personality traits without considering how these values are distributed can significantly underrepresent the interpretation of those personality traits. The fine level of granularity is crucial for AVI-based personality assessment because imprecise assessment can lead, for instance, to inaccurate hiring decisions [62], [63].

In addition, LLM-based personality assessment suffers from low rating validity due to their tendency to align with human preferences. LLMs constantly assign high personality scores to a disproportionately large number of individuals. This issue is referred to as the "*positivity bias*" by Wen et al. [12] or the problem of "*overly lenient recruiters*" by Zhang et al [13]. This bias can lead to the overestimation of candidates' personality traits, resulting in poor hiring decisions that affect organizational performance and success [64], [65].

Because of the aforementioned limitations, current LLM-based personality assessment methods are not yet suitable for AVI analysis and recruitment processes. Thus, it seems necessary to develop LLM-based personality assessment methods that overcome these challenges.

2.5 Mixture-of-Expert LLMs

The Mixture of Experts (MoE) framework is an approach that integrates multiple experts to address complex tasks and enhance decision-making processes [23]. In this framework, distinct expert models are employed to manage various components of a problem, with a gating mechanism determining the appropriate expert for each specific input.

Previous works [23], [66], [67] have demonstrated that leveraging the collective expertise of multiple LLMs can lead to the development of more capable and robust models for a variety of tasks. For instance, Wang et al. [23] proposed an architecture in which multiple LLM experts are organized into layers. In this architecture, each expert uses the outputs from experts in the preceding layer as auxiliary information to generate its response. Their evaluation on AlpacaEval 2.0 [68], MT-Bench [69], and FLASK [70] demonstrated substantial improvements in response quality and achieved state-of-the-art performance. Li et al. [67] further extended the MoE structure by introducing sparsity in expert interactions and incorporating role-playing to foster diverse thinking among experts. Their method consistently improved performance across alignment, reasoning, and fairness benchmarks, achieving results comparable to MoE while significantly reducing toxicity and stereotypes. In psychological research [7], [20], personality traits are typically assessed by multiple annotators who provide independent ratings, and a average score is subsequently calculated. For example, in the work by Koutsoumpis et al. [49], four independent psychologists rated the personality traits in AVIs. This design ensures the objectivity and reliability by incorporating diverse perspectives and minimizing individual biases in the assessment results.

Inspired by the work mentioned above, we designed a framework that leverages multiple experts (i.e., LLMs) to evaluate complex tasks, similar to how multiple annotators assess personality traits. Each expert in our framework acts as an independent evaluator, offering its unique perspective on the task at hand. By integrating the outputs of these experts, we aim to achieve a balanced and comprehensive decision-making process. By drawing parallels between psychological assessment methods and the MoE framework, we seek to harness the strengths of collective intelligence to tackle personality assessment effectively.

3 Methodology

In this section, we introduce a two-stage Mixture-of-Expert framework, *MoE-Personality*, designed for the assessment of personality traits. Unlike autonomous generative experts that can plan, use tools, or interact with environments [71], [72], the '*experts*' in our system function more like task-specific annotators operating within a coordinated pipeline. Each large language model contributes complementary perspectives, and their outputs are aggregated through a structured decision process. In this sense, our approach is better understood as an ensemble strategy guided by the MoE principle and tailored for personality assessment tasks.

To facilitate the generation of personality predictions by LLMs, we employed Google Cloud's speech-to-text transcription service to automatically transcribe the video responses from the AVIs. The proposed model comprises two modules: (1) ***The Classification Module***, wherein the model initially engages in conventional coarse-grained personality classification tasks. This module capitalizes on the text-centric strengths of LLMs to obtain textual personality trait categories. (2) ***The Regression Module***, which subsequently refines the evaluation of the classification module, by calculating fine-grained, numeric personality scores. This is achieved by integrating the identified personality categories with empirical score ranges pertinent to each category. During this phase, the identified categories function as benchmarks to modulate the numeric scores and mitigate excessive positivity bias. We implement a MoE aggregation mechanism within each

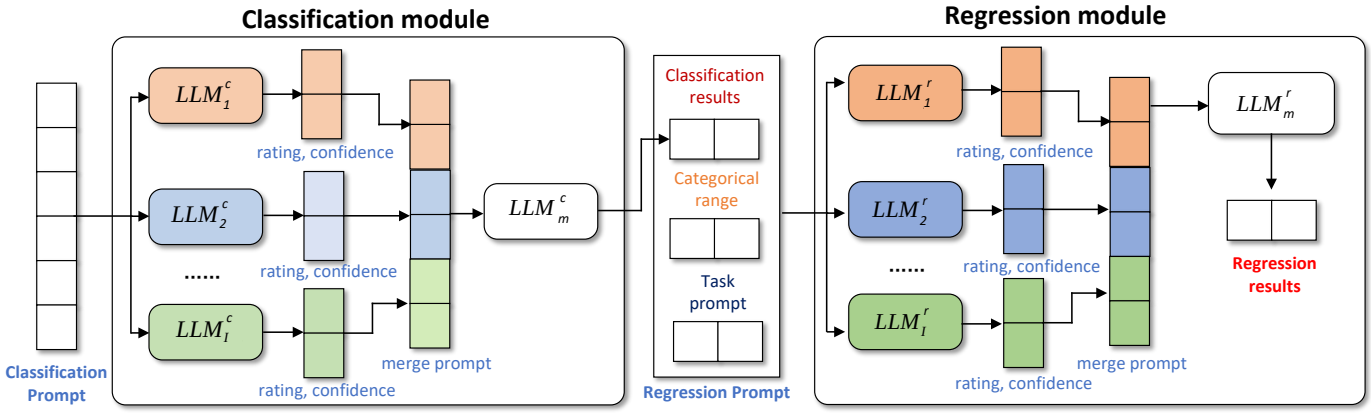


Fig. 1. Overview of the MoE-Personality pipeline. Video responses are first transcribed into text. The classification module assigns coarse trait categories, while the regression module refines them into numeric scores using empirical ranges. A Mixture-of-Expert (MoE) aggregation mechanism combines multiple annotator LLMs' outputs into unified predictions.

module to enhance the validity of the ratings. Within each module, multiple annotator LLM experts independently evaluate and generate personality ratings based on the provided prompts. These individual assessments are then reviewed and aggregated by an aggregator LLM expert to produce unified ratings. The structural design of the framework is shown in Fig. 1.

3.1 Mixture-Of-Expert LLMs

A MoE framework consists of a stack of modules known as MoE layers (i.e., in our case, the classification (c) and regression (r) module). As shown in Fig. 1, each module comprises a set of I annotator LLMs ($LLM_1^j, LLM_2^j \cdots LLM_I^j, j = c, r$) and one aggregator LLM ($LLM_m^j, j = c, r$). Our method does not require fine-tuning and only utilizes the interface of prompting and generation to get ratings of the personality traits. Below, we introduce the design of annotator and aggregator LLMs.

3.1.1 The Annotator LLMs

To construct the prompt for the annotator LLMs, we organize it into three main parts: 1) ***Annotator Instruction***, 2) ***Question Input*** and 3) ***Answer Instruction***. As shown in Fig 2, each part incorporates the different key components we designed to guide the model effectively.

1) **Annotator Instruction:**

- **Role Definition:** The annotator instruction begins by defining the model's role as a psychologist specializing in personality research. This sets the context and expertise required for the task.
- **Task Specification:** It then outlines the primary task: to rate the personality score of an individual based on their responses to a series of questions. The instruction specifies the use of the HEXACO personality model [24], detailing the framework for evaluation.
- **Score Range:** We restrict the score range for the personality rating in the instruction part to make sure the annotator LLMs output the corresponding ratings. The score range can be a categorical label (high/medium/low) for the classification module, or a numeric value for the regression module.
- **Rating Confidence:** The instruction includes the requirement for the model to provide a confidence score for each assessment. The rating confidence component is integral to the assessment process, as it compels the annotator LLMs to critically evaluate and express the certainty of their personality trait ratings. This self-assessment not only enhances the reliability and accuracy of the model's outputs, but also provides a quantifiable measure for aggregators in subsequent layers.

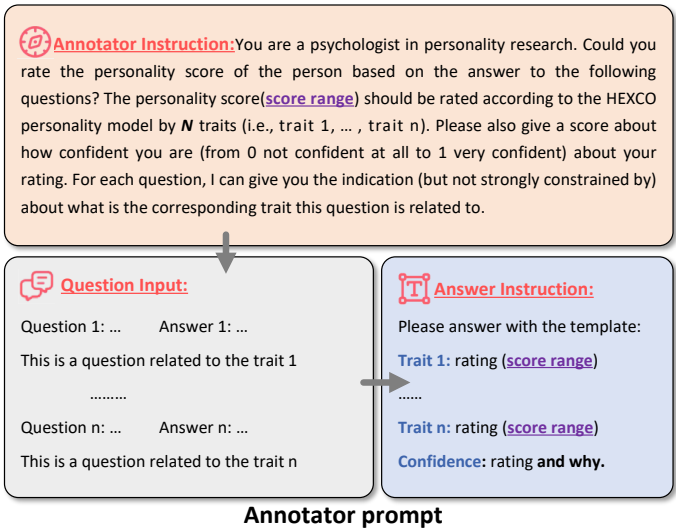


Fig. 2. The construct of annotator prompts

2) **Question Input:**

- **Question-Trait Indication:** In this part, the questions and answers of participants are provided to the annotator. Each question is accompanied by an indication of the corresponding HEXACO trait it relates to, offering guidance to streamline the evaluation process while allowing interpretative flexibility.
- **Contextual Information:** This additional component allows users to provide any relevant background or contextual information about the individual being assessed. This context can help the annotator better understand the linguistic cues of the responses and tailor its evaluation accordingly.

3) **Answer Instruction:**

- **Template:** This part provides a structured format for the

model's output. It includes placeholders for each HEXACO trait rating and the confidence score, ensuring the response is organized and avoid hallucinations [9], [73]. The template facilitates easy interpretation of the results and align with the structured framework provided by the HEXACO personality model.

- **Verbal explanation:** This additional component encourages the model to provide a text summary of the personality assessment. It can enhance the interpretability of the annotator's personality assessment, making it easier for aggregators to evaluate the quality of the annotator's outputs. By providing a text summary, it offers a clear rationale for the assessments, which serves as a basis for aggregators to assess and integrate the results in later stages. This added component of explanation supports the development to get a more coherent and comprehensive synthesis for future aggregation.

The annotator prompts were designed to mimic the human-based annotation procedure of previous psychological studies [7], [49]. Here, we use chain-of-thought prompting strategy (i.e., ask the LLMs to think about and output why they rate the personality traits like this) inspired by previous works [18], [74] from the LLMs community. Each annotator LLM is instantiated from a different family or size of models to introduce intrinsic diversity due to differences in training corpora, parameterization, and pre-training objectives. Such heterogeneity ensures that the annotators do not rely on the same inductive biases when mapping responses to trait scores. To further encourage diversity, we adopt two design choices. First, the prompts are phrased in a flexible manner (with reference to HEXACO traits but without enforcing strict rules or fixed templates). This open-ended formulation allows each model to weigh the linguistic cues in slightly different ways. Second, we inject randomization in the decoding settings to prevent deterministic replication of answers across annotators.

3.1.2 The Aggregator LLM

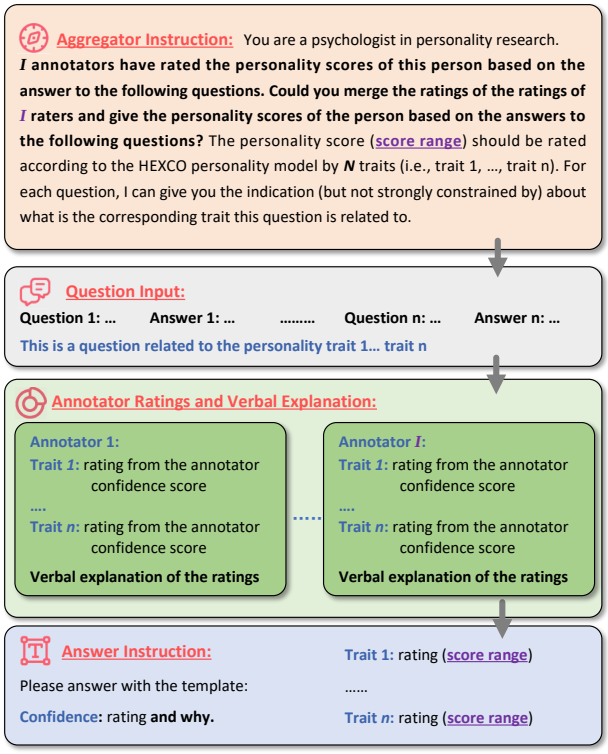
After getting the personality ratings from I annotator LLMs ($LLM_1^j, LLM_2^j \cdots LLM_I^j, j = c, r$), the aggregator LLM ($LLM_m^j, j = c, r$) is used to merge the ratings. Suppose x_i is the input prompt of the annotator LLM_I, the output of the LLM_m, g_m, can be expressed as follow:

$$g_m = \oplus_{i=1}^{I}[A_i(x_i)] + x_0 \tag{1}$$

where x_0 is the shared module (i.e., part of the annotator instruction, user input and answer template module of the annotator prompt) $A_i, i = [1, I]$ shown in Fig 2. $\oplus$ means the construct of the aggregator prompt illustrated in Fig 3.

The aggregator LLM prompt is designed to synthesize ratings from multiple annotator LLMs into a unified personality assessment. It comprises of four main components: 1) ***Aggregation Instruction***, 2) ***Question Input***, 3) ***Annotator Ratings and Verbal Explanation***, and 4) ***Answer Instruction***. Each component plays a crucial role in guiding the aggregator LLM to effectively merge the individual ratings.

1) **Aggregator Instruction:**
   - **Role Definition:** Similar to the annotator prompt, the aggregator prompt begins by defining the model's role for synthesizing individual ratings into a unified personality score. This establishes the contextual framework and objectives for the task.



Aggregator prompt

Fig. 3. The construct of aggregator prompt

   - **Merging Strategy:** After that, the strategy for amalgamating ratings is elucidated by emphasizing the integration of both ratings and confidence scores provided by the annotator LLMs. The application of statistical or heuristic methodologies is specified to ensure a precise and balanced aggregation.
   - **Quality Assurance:** Guidelines are then incorporated for evaluating the consistency and reliability of the aggregated outcomes. That ensures a high degree of confidence and accuracy in the final synthesis.

2) **Question Input:**
   - **Questions and Context:** Similar to the annotator prompt, original questions addressed to the individual are furnished alongside the pertinent contextual information. This aids the aggregator in comprehending the foundation of the ratings and the context within which responses were elicited.
   - **Trait Indication:** Indications of the corresponding HEXACO personality traits (i.e., dimensions that are used to describe and assess an individual's personality) associated with each question are again input to the aggregator LLM. It offers guidelines to streamline the evaluative process while allowing for interpretative flexibility.

3) **Annotator Ratings and Verbal Explanation:**
   - **Ratings Compilation:** The ratings and confidence scores from each annotator LLM for each HEXACO trait are compiled, enabling the aggregator to conduct a comprehensive analysis of each annotator's rating.
   - **Verbal Explanations:** The verbal explanations of the personality ratings generated by the annotator LLMs are combined and input into the aggregator LLM. This provides the aggregator LLMs with insights into the rationale underpin-

ning each rating and assists the aggregator in grasping the context and subtleties of the assessments.

4) **Answer Instruction:**
   - **Output Format:** A structured format is defined in this part for presenting the final aggregated personality scores, including a summary of each HEXACO trait rating to ensure clarity and consistency.
   - **Confidence Assessment:** The aggregator LLM is also required to output the confidence score for aggregation. This ensures that the reliability of the outputs from previous layers is considered across different modules (i.e., the classification and regression module). This can enhance the effectiveness of subsequent modules in utilizing prior information (e.g., the categorical ratings provided by the classification module).
   - **Synthesis Explanation:** The aggregator is also asked to provide a succinct explanation of the derivation of the final scores. This is set to enhance transparency and understanding of the aggregation process.

This structured approach ensures that the aggregator LLM can effectively merge individual ratings into a coherent and reliable personality assessment, leveraging the detailed input from annotators while addressing potential inconsistencies.

3.2 MoE-based personality assessment

The proposed assessment framework combined two layers of MoE (i.e., the classification and regression module) to get fine-grained personality ratings. The details of how we construct the two modules are illustrated below.

3.2.1 Classification Module

The classification module consists of I annotator LLMs and one aggregator LLM, forming a typical MoE framework as described in section 3.1. In this module, all annotators are tasked to classify personality traits as 'high', 'medium', or 'low' based on interviewees' responses to corresponding questions. Subsequently, the aggregator LLM merges the ratings and verbal explanations from the annotator LLMs to produce aggregated personality ratings. The categorical personality ratings derived from this module provide guidelines for fine-grained ratings in the regression module to reduce the positivity bias.

3.2.2 Regression Module

Similar to the classification module, the regression module also consists of I annotator LLMs and one aggregator LLM. However, all annotators are tasked to rate the personality traits on a 5-point Likert-type scale, from '1.0' to '5.0'. The categorical ratings provided by the classification module are input to both the annotator and aggregator LLMs in the regression module. Specifically, we modify the annotator/aggregator and answer instruction in the prompt, which is illustrated in Fig 4.

As shown in Fig. 4, $C_1, C_2, \cdots, C_n$ are the categorical ratings (high/medium/low) for personality traits $trait_1, trait_2, \cdots, trait_n$ respectively. We merge the empirical score range of high/medium/low personality ratings in the prompt to regulate the output of the regression module. These score ranges correspond to the 10th, 50th, and 90th percentiles (representing the percentage of respondents whose scores fall below a given value) based on previous psychological studies [24], [75] conducted in a large population. The 10th, 50th, and 90th percentiles and the corresponding score range for each categorical personality rating are shown in Table 1.

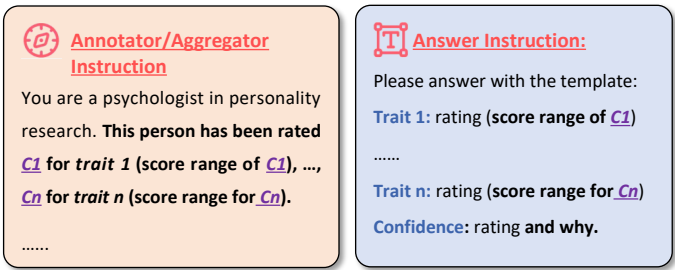


Fig. 4. The modification of annotator/aggregator instruction and answer template for the regression module

TABLE 1
The percentiles of HEXACO personality traits and the score range we set to regulate different categorical labels for the regression module

Personality Traits	percentiles			score range		
	10th	50th	90th	Low	Medium	High
Honesty-Humility(H)	2.4	3.2	4.0	1.9-2.9	2.7-3.7	3.5-4.5
Emotionality(E)	2.6	3.3	4.0	2.1-3.1	2.8-3.8	3.5-4.5
eXtraversion(X)	2.7	3.5	4.2	2.2-3.2	3.0-4.0	3.7-4.7
Agreeableness(A)	2.2	3.0	3.7	1.7-2.7	2.5-3.5	3.2-4.2
Conscientiousness(C)	2.7	3.5	4.2	2.2-3.2	3.0-4.0	3.7-4.7
Openness(O)	2.5	3.3	4.1	2.0-3.0	2.7-3.8	3.6-4.6

For each categorical label, we take these percentiles as the midpoint and expand by 0.5 points on either side to define the score range for each label. This approach results in some overlap between the score ranges of different categorical labels. Given that the classification module may occasionally produce errors, this overlap enhances the robustness of our framework: even if the classification module does not categorize accurately, the regression module's results are less likely to deviate significantly from the true values.

At last, the regression module outputs fine-grained personality ratings based on the participants' responses and the score ranges provided by the previous module. Our framework generates verbal explanation and confidence scores of the ratings. It can help human experts understand and evaluate the rationale behind the ratings, facilitating more effective use of these assessments in real-life applications such as personnel selection, personalized feedback, and psychological research.

4 Datasets

To assess the effectiveness of *MoA-Personality*, we conducted evaluations on two AVI-based personality assessment datasets: *OPVA* [20] and *AVI-6* [76] dataset. The OPVA dataset includes AVI data from 685 individuals who took part in a simulated management traineeship application. In this dataset, eight questions (four per personality trait) were designed by psychologists to evoke the HEXACO personality traits of Extraversion and Conscientiousnesss, two of the most valid personality traits for workplace behaviors [77]. Similarly, the AVI-6 dataset comprises AVI data from 646 participants in mock job application interviews. In this dataset, four questions were designed by psychologists to elicit the HEXACO personality traits of Honesty-Humility, Extraversion, Agreeableness and Conscientiousness. In addition, participants

also answered two generic questions which are frequently asked in selection interviews.

We selected these two datasets for performance evaluation because the interview questions were specifically designed by psychologists to reflect real-world selection contexts and activate personality traits based on *Trait Activation Theory* [53]. In contrast, most other publicly available datasets [34], [39], [78] are sourced from the internet (e.g., YouTube) and do not provide clarity on whether or which personality traits are activated. Therefore, we choose the OPVA and AVI-6 datasets as they provide structured interview data specifically designed to elicit and measure personality traits in an interview setting. This ensures that the performance evaluation is grounded in a context where the activation of personality traits is intentional and aligned with theoretical frameworks.

Since our method is zero-shot and does not involve model training or fine-tuning on human data, we do not perform any additional processing beyond what is allowed by the dataset licenses. We have clarified dataset access restrictions, permissible use conditions, and attribution requirements in the revised paper. Furthermore, we confirm that no personally identifiable information (PII) beyond the video frames included in the datasets was used.

5 Experiments and results

In this section, we first introduce the implementation details of *MoE-Personality* on OPVA and AVI-6 datasets. We then evaluate the regression performance of *MoE-Personality* by comparing the ratings between *MoE-Personality* and human annotators. After that, we compare the performance of *MoE-Personality* with both LLM-based and non-LLM based personality assessment methods. At last, we conducted ablation studies to verify the effectiveness of each step and component of *MoE-Personality*.

5.1 Implementation details

In the experiments conducted in this section, we utilize open-sourced small-sized (model size $\approx$ 10B) LLMs as both annotators and aggregators. Deploying large-sized LLMs (model size $\geq$ 40B) locally demands high hardware specifications. Therefore, we opted for small-sized LLMs to strike a balance between computational efficiency and model performance. We selected open-source models (Llama-3.1, Qwen-2.5, and Gemma-2) primarily to ensure reproducibility and accessibility of our results, allowing the research community to replicate and extend our experiments without reliance on proprietary APIs. While our approach employs zero-shot prompting and could also be applied to closed-source models such as GPT-3.5 or GPT-4, resource constraints prevented such evaluations in this study. We regard this as an important avenue for future work to explore the potential performance ceiling of the proposed framework.

Specifically, we employ the llamma 3.1 (8B) [10], mistral-nemo (12B) [79], and qwen2 (7B) [80] as the annotator LLMs, and use gemma2 (9B) [11] as the aggregator LLM. The impact of varying the number of annotators and the choice of different models for the aggregators is discussed in detail in sections 6.2. We used ollama framework to implement the *MoE-Personality*. All our experiments are performed on a desktop with NVIDIA RTX 4090 GPU (24GB).

5.2 Evaluation metrics

The evaluation metrics for model performance are the Mean Squared Error (MSE) and the Pearson correlation coefficient (*corr*). MSE is a widely adopted metric in regression tasks that calculates the average squared differences between predicted and actual values. This metric is particularly well-suited for evaluating continuous traits and behaviors (e.g., personality traits from AVIs). MSE provides a clear and interpretable measure of how well the model's predictions align with the true values. A lower MSE indicates more accurate predictions of personality traits and interview performance. By squaring the errors, MSE places more weight on larger discrepancies between predicted and actual values. This helps in identifying and minimizing significant prediction errors. Specifically, the MSE is calculated as follows:

$$MSE = \frac{1}{M}\Sigma_i(y_i - x_i)^2 \tag{2}$$

where y_i and x_i are the ground truth and predicted personality traits respectively. M is the number of participants in one dataset.

Furthermore, *corr* is a statistical measure that evaluates the linear relationship between the ground truth and predicted values. *corr* is calculated as the ratio of the covariance of the two variables to the product of their standard deviations

In our experiment, the ground truth personality traits are the *mean observer-reports* conducted by four trained human annotators. Compared with self-reported personality traits, using observer-reported personality traits as ground truth for evaluation can reduce individual bias and increase the validity of the evaluation.

5.3 Results

As shown in Table 2, the proposed model (i.e., *MoE-Personality*) demonstrates strong performance across all HEXACO personality traits, with notable variations depending on which traits were behaviorally activated in each dataset.

In the OPVA dataset, where eXtraversion (X) and Conscientiousness (C) were explicitly activated in AVIs, our model achieved better MSE and *corr* compared with these non-activated traits. For X, it attained the lowest MSE (0.30) and highest correlation (0.71), indicating highly accurate predictions of expressive and social behaviors captured in interview settings. Similarly, for C, it achieved the second-best MSE (0.42) and highest correlation (0.46), reflecting precise assessment of task-oriented behaviors.

In the AVI-6 dataset, where Honesty-Humility (H), X, Agreeableness (A), and C were activated in AVIs, the model excelled across these focal traits. It achieved the lowest MSE for H (0.23) and highest correlation (0.24), demonstrating strong alignment with social perception measures. For X, it maintained robust performance with the second-best MSE (0.36) and highest correlation (0.58). Similarly, for C, it achieved the best MSE (0.40) and highest correlation (0.58). While E and O were not activated in this dataset, they still performed reasonably well (E: 0.45 MSE, 0.32 corr; O: 0.36 MSE, 0.30 corr), though not as strongly as the activated traits.

The results empirically validate trait activation theory [53]. The performance of *MoE-Personality* peaks on traits explicitly triggered by AVI design, while non-activated traits lag despite the model's generalizability.

TABLE 2
The results of MoE-Personality and baseline methods

Mean Square Error (MSE)															
Models		OPVA							AVI-6						
		H	E	X	A	C	O	**Avg**	H	E	X	A	C	O	**Avg**
Open source LLMs	llama3.1 [10]	0.59	0.64	1.04	1.17	0.61	0.42	0.75	1.26	0.46	0.73	1.13	0.82	0.50	0.82
	mistral-nemo [79]	1.38	0.77	0.89	0.98	1.09	0.43	0.92	1.63	0.79	0.54	0.60	0.78	0.45	0.80
	qwen2 [80]	0.97	0.62	1.21	1.17	0.98	0.57	0.92	1.22	0.56	0.51	1.08	1.05	0.78	0.87
	gemma2 [11]	0.49	0.62	0.69	0.68	0.48	0.41	0.56	0.31	0.50	0.48	0.49	0.59	0.43	0.47
	deepseek-r1 [81]	0.71	0.56	0.78	0.75	0.80	0.64	0.71	1.00	**0.36**	0.60	0.69	0.76	0.42	0.64
	vanilla MoA [23]	0.49	0.48	0.33	0.62	0.52	0.44	0.48	0.31	0.43	0.38	0.45	0.51	0.40	0.41
Close source LLMs	gpt-4o-mini [82]	0.49	0.51	0.44	**0.57**	0.50	**0.32**	0.47	0.48	0.89	0.61	0.95	0.76	0.60	0.72
	gpt-3.5 [83]	0.41	0.40	0.76	0.72	0.80	0.47	0.60	0.48	0.92	0.80	0.83	0.66	0.47	0.69
	gpt-4 [9]	0.40	0.38	0.37	0.75	0.49	0.36	0.46	0.44	0.90	0.73	0.55	0.80	0.58	0.67
LLM-based personality assessment	Derner et al. 2024 [15]	0.42	0.51	0.68	0.71	0.66	0.57	0.59	0.51	0.81	0.72	0.81	0.67	0.43	0.66
	Hu et al. 2024 [58]	0.49	0.43	0.53	0.78	0.61	0.51	0.56	0.39	0.72	0.78	0.88	0.52	0.41	0.62
	Cao et al. 2024 [19]	0.44	0.41	0.49	0.69	0.71	0.47	0.54	0.33	0.79	0.71	0.78	0.88	0.52	0.67
	Yan et al. 2024 [14]	0.41	0.44	0.61	0.72	0.59	0.48	0.54	0.41	0.77	0.66	0.53	0.51	0.57	0.58
Non-LLM based personality assessment	Affective-NLI [60]	0.42	0.29	0.33	0.78	0.69	0.52	0.51	0.34	0.58	0.44	0.48	0.57	0.40	0.56
	Personality BERT [84]	0.54	0.33	0.57	0.70	0.50	0.52	0.53	0.24	0.48	0.57	0.49	0.56	0.38	0.45
	PANDA [85]	0.52	0.29	0.59	0.69	0.53	0.71	0.56	0.27	0.52	0.50	0.47	0.41	0.39	0.43
	Hussain et.al 2025 [86]	0.42	0.31	1.59	0.72	2.32	0.63	1.00	0.26	0.51	0.42	0.45	0.42	0.44	0.42
Our method	simple average	0.67	0.52	0.54	0.88	0.67	0.42	0.62	0.34	0.56	0.67	0.72	0.48	0.50	0.55
	MoE-Personality	**0.30**	**0.28**	**0.30**	0.67	**0.42**	0.40	**0.39**	**0.23**	0.45	**0.36**	**0.44**	**0.40**	**0.36**	**0.37**
Pearson correlation coefficient (corr)															
Open source LLMs	llama3.1 [10]	0.15	0.25	0.54	0.20	0.35	0.22	0.29	0.17	0.21	0.48	0.40	0.52	0.28	0.34
	mistral-nemo [79]	0.12	0.11	0.57	0.14	-0.09	0.23	0.18	0.14	0.03	0.47	0.23	0.16	0.24	0.21
	qwen2 [80]	0.18	0.01	0.45	0.26	0.06	0.15	0.19	0.12	0.11	0.41	0.32	0.34	0.24	0.26
	gemma2 [11]	0.16	0.19	0.58	0.23	0.28	0.22	0.28	0.12	0.25	0.51	0.50	0.51	0.19	0.35
	deepseek-r1 [81]	0.14	0.23	0.56	0.16	0.13	0.12	0.22	0.16	0.26	0.43	0.42	0.36	0.30	0.32
	vanilla MoA [23]	0.12	0.17	0.53	0.21	0.42	0.41	0.31	0.21	0.24	0.53	0.52	0.51	0.27	0.38
Close source LLMs	gpt-4o-mini [82]	0.26	0.35	0.67	**0.33**	0.38	**0.48**	0.41	0.21	0.34	0.57	0.41	0.16	**0.38**	0.34
	gpt-3.5 [83]	0.19	0.14	0.66	0.30	0.40	0.38	0.35	0.18	0.25	0.32	0.18	0.11	0.22	0.21
	gpt-4 [9]	0.27	0.14	0.68	0.22	0.40	0.42	0.35	0.19	**0.36**	0.38	0.12	-0.01	0.16	0.20
LLM based Personality assessment	Derner et al. 2024 [15]	0.16	0.15	0.61	0.22	0.40	0.36	0.32	0.11	0.22	0.31	0.19	0.12	0.21	0.19
	Hu et al. 2024 [58]	0.18	0.21	0.58	0.27	0.41	0.31	0.33	0.14	0.24	0.33	0.26	0.33	0.24	0.26
	Cao et al. 2024 [19]	0.22	0.17	0.57	0.23	0.39	0.32	0.32	0.12	0.26	0.26	0.33	0.21	0.22	0.23
	Yan et al. 2024 [14]	0.24	0.15	0.69	0.31	0.40	0.43	0.37	0.21	0.28	0.38	0.14	0.11	0.18	0.22
Non-LLM based personality assessment	Affective-NLI [60]	0.21	0.31	0.52	0.21	0.22	0.43	0.32	0.20	0.19	0.19	0.51	0.51	0.29	0.32
	Personality BERT [84]	0.19	0.33	0.54	0.22	0.31	0.41	0.33	0.22	0.29	0.41	0.42	0.49	0.29	0.35
	Lim et.al 2025 [85]	0.17	0.28	0.62	0.11	0.29	0.21	0.28	0.19	0.20	0.33	0.52	0.56	0.22	0.34
	Hussain et.al 2025 [86]	0.08	0.20	0.13	0.14	0.09	0.24	0.15	0.18	0.19	0.41	0.51	0.51	0.27	0.35
Our method	simple average	0.09	0.11	0.10	0.10	0.22	0.25	0.15	0.14	0.18	0.30	0.33	0.25	0.12	0.22
	MoE-Personality	**0.29**	**0.39**	**0.71**	0.30	**0.46**	0.45	**0.44**	**0.24**	0.32	**0.58**	**0.57**	**0.58**	0.30	**0.43**

5.4 Comparison with baselines

To evaluate the performance of *MoE-Personality*, we compare it against a diverse set of baseline models, including both open-sourced small-sized LLMs (i.e., llama3.1(8B) [10], mistral-nemo(12B) [79], qwen2(7B) [80], gemma2(9B) [11], deepseek-r1(8B) [81]), vanilla MoA [23] and closed-sourced large-sized LLMs from the gpt series (i.e., gpt-3.5 [83], gpt-4 [9], gpt-4o-mini [82]). The comparison between medium-sized LLMs help us to determine whether *MoE-Personality* can outperform similarly sized models while maintaining efficiency and accessibility. These models were selected due to their competitive performance in general language tasks and their suitability for deployment in personality assessment. Comparing *MoE-personality* against close-source large-sized LLMs helps us to figure out whether it can match or exceed the performance of much larger, commercially optimized models due to its mixture-of-expert structure. Both the baseline LLMs and the annotator LLMs were used under comparable conditions, with prompts explicitly designed to produce continuous trait scores. Specifically, the prompt instructed the models to act as psychologists in personality research, to assign HEXACO-based trait ratings on a continuous scale from 1.0 to 5.0, and to use the corresponding interview question as a cue to guide the trait evaluation.

In addition, we also compare *MoE-Personality* with both

LLM-based personality assessment methods [14], [15], [19], [58] and non-LLM-based deep learning methods [60], [84], [85], [86] that were specifically optimized for personality prediction. The LLM-based personality assessment methods go beyond naive prompting by leveraging fine-tuning, augmentation, role-play prompting and embedding-based modeling tailored to personality assessment tasks. The baselines of LLM-based personality assessment methods are obtained by our own re-implementation and re-running of each method on our two evaluation datasets. We adapted each baseline in a minimally invasive way to ensure a fair comparison. We preserved the core inference or feature pipeline of the original method while replaced (a) the personality model with HEXACO, and (b) the output format with six continuous trait scores. The non-LLM-based deep learning methods capture traditional text-based and hybrid architectures widely applied to personality prediction. As most of these methods were also designed as classification models, we adapted their network structures by replacing the final softmax classification layer with a regression head, thereby enabling them to output continuous personality trait values. This ensures that all baselines can be evaluated under the same MSE and correlation metrics as our proposed model.

Compared with open-source and close source LLMs, *MoE-Personality* achieves the lowest average error on both datasets (0.39 on OPVA and 0.37 on AVI-6), outperforming all baselines. Among open-sourced models, gemma2 performs best with an MSE of 0.56 on OPVA and 0.47 on AVI-6. The gpt series, particularly gpt-4o-mini, demonstrates strong performance with an MSE of 0.47 on OPVA, yet *MoE-Personality* still surpasses it by a notable margin. The correlation analysis further reinforces the superiority of *MoE-Personality*, which achieves the highest average correlation (0.44 on OPVA and 0.43 on AVI-6), indicating better alignment with human personality assessments. Among open-sourced models, gemma2 (0.28 on OPVA, 0.35 on AVI-6) and deepseek-r1 (0.22 on OPVA, 0.32 on AVI-6) exhibit moderate correlation strengths. The gpt models show varying performance, with gpt-4o-mini (0.41 on OPVA, 0.34 on AVI-6) being the strongest, though still falling short of *MoE-Personality*.

For the comparison between LLM-based and the non-LLM-based deep learning methods, LLM-based personality assessment methods yield MSE in the range of 0.54 to 0.66 across the two datasets. Although these methods improve upon naive prompting strategies, their performance still lags behind *MoE-Personality*. This demonstrates that the mixture-of-experts design provides complementary benefits by reducing prediction variance and enhancing robustness across datasets. For non-LLM-based methods, although these models are effective in conventional text-based personality assessment, they exhibit higher error rates on AVI-style datasets, particularly on OPVA where MSE typically exceeds 0.50. In contrast, *MoE-Personality* achieves substantially lower error on both datasets, outperforming the strongest non-LLM baseline on AVI-6 by a margin exceeding 15%.

The vanilla MoA method [23], while sometimes outperforming conventional LLM approaches, still performs worse than our *MoE-Personality* framework. This performance gap arises because the original MoA design treats all annotator agents as parallel and independent judges without explicit task specialization or staged reasoning. In contrast, our two-step MoE-Personality approach first decomposes the task into coarse-grained trait-specific reasoning and fine-grained numerical rating. This decomposition allows the model to perform judgments in a staged manner, which effectively constrains the influence of global impression formation and prevents the positivity bias.

These results demonstrate that *MoE-Personality* consistently outperforms all baselines in both MSE and correlation metrics. While the closed-sourced gpt models serve as strong competitors, our method achieves better or similar results. The results show that the MoE approach can enhance performance even when compared to much larger general-purpose models. Among open-sourced alternatives, gemma2 emerges as the closest competitor. However, *MoE-Personality* maintains a clear advantage, particularly in correlation strength. In addition, *MoE-Personality* consistently outperforms state-of-the-art specialized baselines. On OPVA, it improves upon strong baselines Hu et al. [58] (0.56) by roughly 15–30%. On AVI-6, it surpasses Personality-BERT [84] (0.45) and Yan et al. [14] (0.57) by 18–35%. These results illustrate its robustness across different datasets. These findings also underline the promise of mixture-of-experts architectures for achieving accurate, stable, and scalable text-based personality assessment.

5.5 Ablation study

In this section, we first compare the results of simple average of annotators and aggregated results from annotator. Without the comparison between simple average of annotators, it is not possible to determine the contribution of the aggregator when compared to a simple baseline.

As shown in Table 2, the simple average achieves an MSE of 0.62 on OPVA and 0.55 on AVI-6, both notably worse than our proposed MoE-Personality framework (0.39 and 0.37, respectively). A follow-up analysis reveals why simple averaging is insufficient. In practice, annotators LLMs vary in confidence and reliability across traits (i.e., the Confidence Assessment factor in aggregator prompt). Some annotators occasionally output extreme or biased predictions due to low confidence. Simple averaging does not differentiate such outliers from high-quality predictions, which means the final score may be distorted rather than corrected. This often prevents averaging from converging to the true trait value. In contrast, our mixture-of-experts aggregator adaptively weighs annotators based on the text-based explanation provided by annotators about why they rate the traits like this. This can help the aggregator to effectively down-weigh low-quality or outlier predictions. This mechanism explains the substantial improvement of MoE-Personality over simple averaging. To further validate this point, we conducted statistical significance testing (paired t-test) across both datasets. The results confirm that the differences between MoE and simple average are statistically significant ($p<0.01$), reinforcing that our improvements are not due to random variance but to the design of the aggregator itself.

To further validate the efficiency of each component in *MoE-Personality*, we examine four methodological variants through an ablation study. *REG* (regression) represents our baseline approach that directly employs regression for fine-grained personality assessment without any additional architectural enhancements. This simplest form of our model serves as the foundation for comparison. *CLA+REG* introduces a modification by implementing the two-step assessment process. This approach first performs coarse-grained personality classification (CLA) before proceeding to fine-grained regression (REG) without implementing the mix-of-expert structure. *MoE+REG* explores the impact of the mixture-of-experts (MoE) architecture while maintaining direct regression without classification. The complete *MoE-Personality* framework,

represented by *MoE+CLA+REG*, integrates both the mixture-of-experts architecture and the two-step assessment process.

TABLE 3
Ablation study on different methodological variants of *MoE-Personality*. *REG* = direct regression baseline (*O(n)*); *CLA+REG* = two-step classification + regression pipeline (*O(2n)*); *MoE+REG* = mixture-of-experts with regression only (*O((k+1)n)*); *MoE+CLA+REG* = full framework with mixture-of-experts and two-step process(*O(2n(k+1))*). Here, n denotes the input size of a single forward pass, and k=3 denotes the number of annotators LLMs.

Components	Averaged MSE		Averaged *corr*	
	OPVA	AVI-6	OPVA	AVI-6
REG	0.56	0.47	0.35	0.28
CLA+REG	0.53	0.43	0.39	0.32
MoE+REG	0.48	0.41	0.41	0.38
MoE+CLA+REG	**0.39**	**0.37**	**0.43**	**0.44**

As shown in Table 3, the *REG* variant achieves an averaged MSE of 0.56 (OPVA) and 0.47 (AVI-6), with correlation scores of 0.35 and 0.28, respectively. Introducing coarse-grained classification in CLA+REG reduces MSE by 5.4% on OPVA (0.53) and 8.5% on AVI-6 (0.43), and improving correlation to 0.39 and 0.32. This demonstrates that hierarchical assessment, which categorizes personality traits before refining predictions, helps align the regression process with categorical regulation.

The *MoE+REG* variant, which replaces the single model with a mixture-of-experts architecture, also shows performance improvements: MSE decreases by 14.3% (0.48) on OPVA and 12.8% (0.41) on AVI-6 compared to *REG*. These gains highlight the importance of multi-expert collaboration in modeling diverse personality aspects, as specialized experts collectively refine predictions through complementary perspectives.

However, the full *MoE+CLA+REG* model achieves the most significant improvements, reducing MSE to 0.39 (30.4% lower than *REG*) on OPVA and 0.37 (21.3% lower) on AVI-6, with correlation reaching 0.43 and 0.44. Notably, its performance surpasses the additive effects of *CLA+REG* and *MoE+REG*, indicating synergistic interaction between the mixture-of-experts architecture and the two-step pipeline. The classification step provides initialization for the regression tasks, while the experts' diverse expertise enhances the quality of categorical guidance. This mutual reinforcement is evident in the correlation metric on AVI-6, where the full model outperforms *MoE+REG* by 15.8% (0.44 vs. 0.38).

These results confirm that both the mixture-of-experts structure and the hierarchical classification-regression pipeline are indispensable for optimal performance. The integrated framework effectively balances coarse-grained categorical alignment with fine-grained assessment.

6 Discussion

6.1 Positive Bias: does MoE make a difference?

One of the major issues for LLM-based personality assessment is that they suffer from low rating validity due to their tendency to align with human preference and assign high personality scores to an even greater number of individuals (i.e., issues of *Positivity Biases* [12], [13]). Although our experimental results show that *MoE-personality* outperformed the baseline methods, it remains important to examine whether the Mixture-of-Expert structure can reduce the positivity bias in personality assessment.

To answer this question, we compare the Lower quartile (Q1), average (Avg), median (Med), and Upper quartile (Q3) of the ratings from individual annotator LLMs (Annotators 1–3), the aggregated LLM model (Mixture of Annotator 1–3), and human annotators. Annotators 1-3 are individual LLM-based annotators (i.e., llama 3.1, mistral-nemo, qwen2), each providing their own personality ratings. The "Mixture of Annotator 1 - 3" represents an aggregated result from these three individual LLM annotators using gemma2. The "Human Annotator" serves as a benchmark from human judgment.

TABLE 4
The statistical comparison between individual LLM annotators, mixture of annotator and human annotators

	OPVA				AVI-6			
	Q1	Avg	Med	Q3	Q1	Avg	Med	Q3
Annotator 1	3.50	3.75	3.90	4.50	3.50	3.85	4.00	4.50
Annotator 2	3.00	3.78	4.00	4.50	3.00	3.57	3.50	4.50
Annotator 3	3.50	3.87	4.00	4.50	3.50	3.82	3.80	4.50
Mixure of Annotator 1-3	2.80	3.38	3.45	4.00	2.80	3.35	3.00	4.00
Human Annotator	2.75	3.29	3.25	3.75	2.91	3.18	3.16	3.48

As shown in Table 4, aggregating individual LLM annotators (via the MoE structure) reduces positivity bias compared to standalone models. For OPVA dataset, individual LLM annotators showed consistently high ratings with 75% of ratings clustering at the scale's upper end. A similar pattern emerged in the AVI-6 dataset, where individual LLMs such as Annotator 1 reported high scores (Q1=3.50, Avg=3.85, Med=4.00, Q3=4.50). In contrast, the MoE-aggregated outputs display markedly more conservative scoring behavior. Specifically, the average scores of aggregated LLMs drop to 3.38 (OPVA) and 3.35 (AVI-6). The median also lowers to 3.45 and 3.00. In general, the rating distribution of MoE was broader and more balanced. Thus, it closely resembled the variability and central tendency observed in human annotator ratings. Moreover, MoE reduces the clustering of scores near the upper bound. It recovers finer-grained distinctions among individuals—particularly those with lower trait levels.

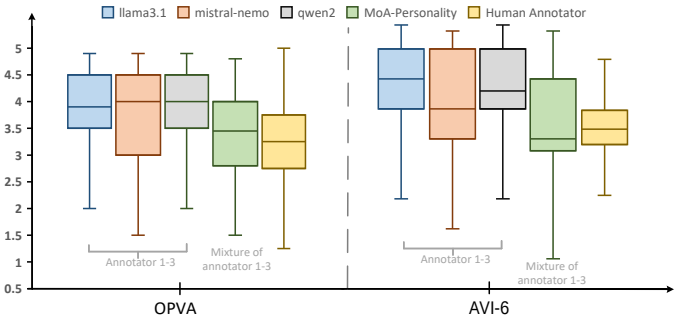


Fig. 5. The boxplot of personality ratings of individual LLM annotators, mixture of annotator and human annotators

These distributional shifts are also clearly visible (Fig 5). The individual annotators show highly compressed boxplots at the upper end of the scale (around 4.2 to 4.5). These patterns illustrate skewed and inflated distributions with limited variance and discriminative capacity across subjects. In contrast, *MoE-Personality* produces a visibly broader distribution. The MoE

boxplot aligns closely with the human annotator distribution, which presents the lowest median and the widest spread among all methods. This observation confirms that MoE does not simply average LLM predictions. Instead, it approximates the human-like sensitivity to inter-individual differences in personality traits.

The reason MoE reduces positivity bias lies in its ability to integrate diverse outputs from multiple LLM annotators while preserving inter-rater variability. Individual LLMs tend to generate inflated scores due to their exposure to human-preference-aligned training data and their tendency to favor socially desirable descriptions. This leads to compressed, skewed distributions that lack discriminative value, especially for individuals with mid-to-low trait levels. MoE addresses this limitation by explicitly modeling the variation across annotators rather than relying on a simple average. It weighs assessment ratings according to shared patterns while suppressing outlier tendencies. This selective aggregation reduces score inflation and reveals a broader and more balanced distribution of personality ratings. Psychologically, this process reflects the principle of consensual validation [87], where averaging across multiple human raters improves accuracy and reduces individual bias. In conclusion, the structure of MoE not only improves predictive performance but also enhances the psychological validity of LLM-based personality assessments by mitigating systematic rating inflation and recovering trait variability that aligns more closely with human judgment.

6.2 Specialization models and number of annotators

Understanding which components contribute most to the effectiveness of MoE-personality is crucial for optimizing both accuracy and efficiency in LLM-based personality assessment. While MoE improves over individual LLMs by aggregating their outputs, its performance depends on two design choices: 1) the selection of the aggregator model and 2) the number of annotators involved. Exploring these factors not only helps explain the internal dynamics of the MoE framework but also provides practical guidance for building more scalable, interpretable, and computationally efficient systems. In this section, we investigate how aggregator specialization and annotator quantity influence overall model performance. In the comparison, we set the aggregator LLM to be gemma2 (the same with the *MoE-Personality*) to make a fair comparison. For the annotator LLMs, we test the combination of all 5 open-source models (i.e., llama 3.1, mistral-nemo, qwen2, gemma2 and deepseek-r1). We report the mean MSE and corr of different combination of different number of annotators. This setup allows us to systematically analyze how the number of annotators influences aggregation performance.

TABLE 5
The performance of MoE-personality using different LLMs as the aggregator. To get a fair comparison, we use all four LLMs in the table as the annotators.

Aggregator	OPVA		AVI-6	
	MSE	*corr*	MSE	*corr*
llama3.1	0.55	0.31	0.44	0.31
mistral-nemo	0.48	0.27	0.41	0.33
qwen2	0.56	0.40	0.48	0.35
gemma2	**0.40**	**0.42**	**0.38**	**0.45**

Table 5 presents the performance of MoE-personality when different LLMs serve as the aggregator, with all four models used as annotators to ensure a fair comparison. The results indicate performance differences across aggregators (recall that we only used open-source medium-size LLMs as aggregators). Gemma2 achieves the lowest mean squared error (MSE) and the highest correlation with human ratings on both datasets. These results suggest that some models are better suited to function as aggregators due to differences in calibration, generalization, or internal representation alignment. It seems that, within the MoE framework, the aggregator acts as a meta-learner that integrates multiple annotator predictions. A stronger aggregator can better detect agreement, filter out inconsistent or noisy predictions, and weight the input more effectively. Thus, selecting a specialized LLM with robust internal reasoning capabilities is critical to maximizing the benefits of MoE.

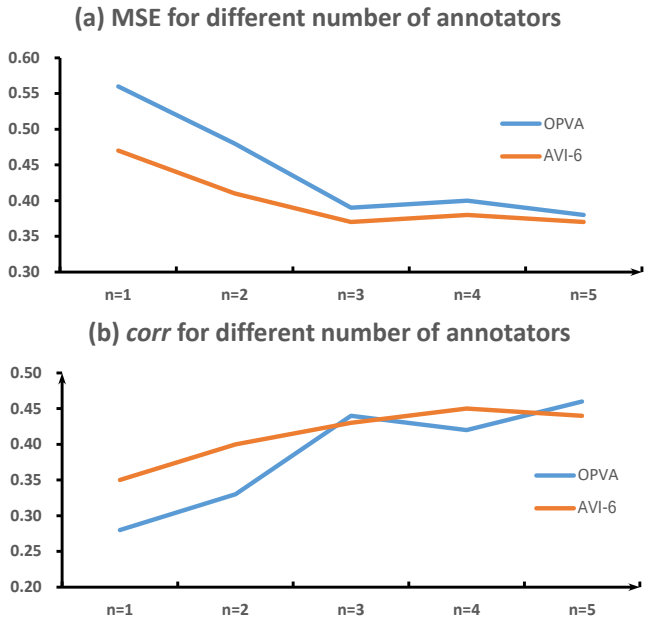


Fig. 6. The results of different numbers of annotators

In addition to the aggregator choice, the number of annotators (i.e., LLMs) in the MoE ensemble significantly impacts performance. The analysis shown in the Fig 6 shows that the performance improves clearly when increasing from 1 to 3 annotators. However, after three annotators the performance seems to reach an asymptote, after which adding more annotators does not contribute to increased performance (or it might even slightly decrease). Thus, it highlights the trade-off between diversity and noise in the MoE-based personality assessment. While increasing annotator count generally adds useful variance, it can also dilute signal quality if weaker or biased annotators are included. The careful selection or weighting is necessary to get validated assessment results when MoE is implemented for personality assessment. This finding aligns with ensemble theory [88], which warns that adding low-quality predictors can degrade overall accuracy if their errors are not sufficiently uncorrelated.

In summary, the discussion emphasizes that the effectiveness of the *MoE-personality* depends not only on the presence of multiple annotators but also on the strategic selection of both the aggregator and the annotator ensemble. While stronger aggregator models such as gemma2 enhance prediction reliability through better calibration and integration of diverse inputs, the number of annotators contributes to performance gains only up to a point.

Beyond three annotators, the benefit plateaus or even fluctuates due to potential noise and redundancy, possibly introduced by less informative models. The key takeaway is that MoE benefits most from thoughtful design: incorporating well-calibrated aggregators and selectively curated annotators yields more valid, robust, and interpretable personality assessments. Future applications should prioritize quality over quantity when assembling annotator ensembles and leverage model specialization to ensure consistent and meaningful trait inference.

6.3 Comparison between fine-tuning LLMs

Fine-tuning has become an increasingly important paradigm in recent LLM applications. It often enables stronger task alignment and improved performance. However, in the context of AVI-based personality assessment, fine-tuning requires substantial computational resources, which limits accessibility and scalability for real-world interview screening systems. Moreover, supervised fine-tuning depends on large volumes of high-quality personality annotations, which are costly and time-consuming to obtain due to the need for trained expert raters.

Despite these constraints, the comparison between zero-shot approaches and fine-tuned LLMs is still essential for understanding the trade-off between annotation effort and model performance. To this end, we conducted experiments using LoRA [89] based parameter-efficient fine-tuning on three representative LLM backbones (i.e., llama3.1, qwen2, and gemma2) and compared them with our MoE-Personality framework on both OPVA and AVI-6 datasets.

For the fine-tuned baselines, we followed the data split protocol proposed in prior work [90], which is specifically designed for AVI-based personality assessment and prevents subject-level leakage between training and evaluation sets. All models were fine-tuned under an identical supervised regression setting to predict continuous personality trait scores, using the same input formatting and evaluation protocol to ensure a fair comparison.

TABLE 6
The comparison between fine-tuned LLMs and MoE-Personality. The reported results are averaged MSE across all personality traits

Model		Dataset	
		OPVA	AVI-6
fine-tuned LLMs	llama 3.1	0.40	0.31
	qwen 2	0.39	0.39
	gemma2	0.50	0.42
MoE-personality		0.39	0.37

As shown in Table 6, fine-tuning can improve performance for certain LLM backbones. However, these gains are not consistent across models or datasets. While llama3.1 benefits from LoRA fine-tuning on both OPVA and AVI-6, gemma2 exhibits limited improvement on AVI-6 and a noticeable performance degradation on OPVA. Qwen2 shows only marginal or dataset-dependent gains. This variability suggests that fine-tuning does not guarantee robust performance improvements for AVI-based personality assessment, particularly under conditions of limited data availability and annotation noise.

In contrast, *MoE-Personality* achieves competitive performance without relying on any task-specific training data, remaining comparable to or outperforming several fine-tuned baselines, especially on AVI-6. These findings highlight the trade-off between annotation effort and performance. Although fine-tuning can yield higher accuracy for selected models, it incurs substantial costs in annotation and computation. *MoE-Personality* therefore offers a more stable, scalable, and deployable alternative for real-world AVI personality assessment scenarios where labeled data are scarce or difficult to obtain.

7 Limitation and future Work

Despite its promising results, *MoE-Personality* has several limitations that point to future improvement. First of all, our approach only uses carefully crafted prompts to guide the LLMs without any task-specific fine-tuning of model weights. This design choice simplifies deployment and allows strong zero-shot capabilities and flexibility. However, the lack of fine-tuning means the LLMs are not explicitly optimized for personality assessment. Future work can explore incorporating fine-tuning to retain flexibility while boosting task-specific accuracy. In addition, our method is not tested using some of the latest large-scale open-source LLMs (e.g., llama 3.1-405B, deepseek-r1-671B) due to their computational cost. Incorporating such powerful models could further improve accuracy and robustness. However, we made a deliberate trade-off for efficiency and resource feasibility: using multiple massive LLMs in parallel (as annotators or aggregators) would greatly increase computational cost and latency, making the system less practical. This resource-conscious design likely sacrifices some raw performance for scalability. Our method is also unimodal and rely solely on linguistic input. Future implementations should explore feeding audio-visual features into the framework alongside text. Such a multimodal MoE-Personality could more holistically evaluate personality, mirroring how human assessors consider both what is said and how it is said. At last, the distributional properties of the HEXACO variables derived from MoE-Personality are not fully aligned with those of the original scale. Addressing this issue constitutes an important direction for future work.

Building on the above limitations, we envision several research directions to enhance *MoE-Personality* both technically and theoretically. On the technical side, integrating fine-tuning could improve task-specific accuracy beyond what prompt engineering alone provides. Expanding the framework to handle multimodal input, such as combining text with audio and visual signals, could offer more complete and human-like assessments. Additionally, improving the robustness and generalizability of prompts, or automating prompt optimization, would make the system more adaptable. On the theoretical front, future work should explore strategies for reducing LLM bias and enhancing model calibration to ensure consistent and fair scoring. Research into optimal annotator selection, inter-model disagreement modeling, and dynamic trait scaling could further strengthen the reliability and interpretability of MoE-based personality assessments. Future work should also integrate audio-visual emotional cues and explore systematic mappings between expressive behaviors and personality traits.

8 Conclusion

In this research, we introduced *MoE-Personality*, a framework for personality assessment using Mixture-of-Experts architecture with multiple annotator LLMs and an aggregator LLM working in concert. Our experimental results showed that *MoE-Personality*

achieved better performance than single-LLM baselines. The proposed system showed good alignment with human ratings and achieved performance competitive with larger closed-source models. In general, our framework outperformed the baseline LLMs by leveraging model diversity without requiring task-specific fine-tuning, making it suitable for scenarios with limited labeled data. Furthermore, *MoE-Personality* provided a privacy-preserving and scalable solution for asynchronous video interviews, offering consistent, fair, and interpretable trait inferences. We also highlight the importance of evaluating positivity bias in LLMs and propose MoE as a structure that mitigates such biases through variance-aware aggregation. Finally, this work lays a foundation for generalizable, multimodal trait assessment, and encourages future development of scalable psychological computing systems grounded in collaborative, human-like evaluation processes.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 62506070), the Natural Science Foundation of Jiangsu Province (Grants No BK20251348), the Research Project of Humanities and Social Sciences of the Ministry of Education (No. 25YJCZH372), the Fundamental Research Funds for the Central Universities (2242025S30059) and Joint Open Project of Southeast University-Jiangsu Province Hospital.

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. C. Veríssimo, P. Oliveira, P. M. Matos, and L. Ribeiro, "Identifying hexaco personality types: what do type characteristics tell us about student misconduct?" *BMC Medical Education*, vol. 25, no. 1, p. 1083, 2025.

[3] 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.

[4] 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.

[5] J. Jin, G. F. Yuan, and Y. An, "Latent profiles and transition of job burnout and their relationship with dispositional mindfulness and reappraisal among firefighters in china," *Personality and Individual Differences*, vol. 221, p. 112544, 2024.

[6] 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.

[7] 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.

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

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

[10] A. D. et.al, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[11] G. T. et al., "Gemma 2: Improving open language models at a practical size," 2024. [Online]. Available: https://arxiv.org/abs/2408.00118

[12] Z. Wen, Y. Yang, J. Cao, H. Sun, R. Yang, and S. Liu, "Self-assessment, exhibition, and recognition: a review of personality in large language models," *arXiv preprint arXiv:2406.17624*, 2024.

[13] T. Zhang, A. Koutsoumpis, J. K. Oostrom, D. Holtrop, S. Ghassemi, and R. E. de Vries, "Can large language models assess personality from asynchronous video interviews? a comprehensive evaluation of validity, reliability, fairness, and rating patterns," *IEEE Transactions on Affective Computing*, 2024.

[14] Y. Yan, L. Ma, A. Li, J. Ma, and Z. Lan, "Predicting the big five personality traits in chinese counselling dialogues using large language models," *arXiv preprint arXiv:2406.17287*, 2024.

[15] E. Derner, D. Kučera, N. Oliver, and J. Zahálka, "Can chatgpt read who you are?" *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 2, p. 100088, 2024.

[16] 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.

[17] H. Peters, M. Cerf, and S. C. Matz, "Large language models can infer personality from free-form user interactions," *arXiv preprint arXiv:2405.13052*, 2024.

[18] 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.

[19] X. Cao and M. Kosinski, "Large language models know how the personality of public figures is perceived by the general public," *Scientific Reports*, vol. 14, no. 1, p. 6735, 2024.

[20] 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.

[21] J. Allik, A. Realo, and R. E. de Vries, "Elusive specific variance: A marginal effect on the accuracy of personality judgment," *Journal of Research in Personality*, vol. 109, p. 104464, 2024.

[22] J. K. Oostrom, D. Holtrop, A. Koutsoumpis, W. van Breda, S. Ghassemi, and R. E. 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*, vol. 97, no. 1, pp. 160–189, 2024.

[23] J. Wang, J. Wang, B. Athiwaratkun, C. Zhang, and J. Zou, "Mixture-of-agents enhances large language model capabilities," *arXiv preprint arXiv:2406.04692*, 2024.

[24] 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.

[25] R. M. Brown, S. G. Roberts, and T. V. Pollet, "Hexaco personality factors and their associations with facebook use and facebook network characteristics," *Psychological reports*, vol. 128, no. 3, pp. 1942–1966, 2025.

[26] M. H. She, R. Ronay, and D. N. Den Hartog, "The sociable and the deviant: a latent profile analysis of hexaco and the dark triad," *Journal of Business Ethics*, vol. 199, no. 3, pp. 529–547, 2025.

[27] 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.

[28] 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.

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

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

[31] R. Liao, S. Song, and H. Gunes, "An open-source benchmark of deep learning models for audio-visual apparent and self-reported personality recognition," *IEEE Transactions on Affective Computing*, 2024.

[32] N. M. Aljuhani, A. A.-M. Al-Ghamdi, H. S. Alghamdi, and F. Saleem, "Convolutional bi-lstm for automatic personality recognition from social media texts," *IEEE Access*, 2025.

[33] A. Naz, H. U. Khan, A. Bukhari, B. Alshemaimri, A. Daud, and M. Ramzan, "Machine and deep learning for personality traits detection: a comprehensive survey and open research challenges," *Artificial Intelligence Review*, vol. 58, no. 8, p. 239, 2025.

[34] 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.

[35] F. Gürpinar, H. Kaya, and A. A. Salah, "Multimodal fusion of audio, scene, and face features for first impression estimation," in *2016 23rd*

International conference on pattern recognition (ICPR). IEEE, 2016, pp. 43–48.

[36] H. Kaya, F. Gurpinar, and A. Ali Salah, "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 1–9.

[37] H. Kaya and A. A. Salah, "Multimodal personality trait analysis for explainable modeling of job interview decisions," in *Explainable and interpretable models in computer vision and machine learning*. Springer, 2018, pp. 255–275.

[38] S. Aslan, U. Güdükbay, and H. Dibeklioğlu, "Multimodal assessment of apparent personality using feature attention and error consistency constraint," *Image and Vision Computing*, vol. 110, p. 104163, 2021.

[39] 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.

[40] C. Ventura, D. Masip, and A. Lapedriza, "Interpreting cnn models for apparent personality trait regression," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 55–63.

[41] X.-S. Wei, C.-L. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 303–315, 2017.

[42] C. Suman, S. Saha, A. Gupta, S. K. Pandey, and P. Bhattacharyya, "A multi-modal personality prediction system," *Knowledge-Based Systems*, vol. 236, p. 107715, 2022.

[43] H. Hayat, C. Ventura, and A. Lapedriza, "On the use of interpretable cnn for personality trait recognition from audio," in *Artificial Intelligence Research and Development*. IOS Press, 2019, pp. 135–144.

[44] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," in *European conference on computer vision*. Springer, 2016, pp. 337–348.

[45] S. Giorgi, J. Sedoc, V. Barriere, and S. Tafreshi, "Findings of wassa 2024 shared task on empathy and personality detection in interactions," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 2024, pp. 369–379.

[46] F. Celli, A. Kartelj, M. Đorđević, D. Suhartono, V. Filipović, V. Milutinović, G. Spathoulas, A. Vinciarelli, M. Kosinski, and B. Lepri, "Twenty years of personality computing: Threats, challenges and future directions," *arXiv preprint arXiv:2503.02082*, 2025.

[47] S. Mushtaq and N. Kumar, "Text-based automatic personality recognition: Recent developments," in *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021*. Springer, 2022, pp. 537–549.

[48] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[49] 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.

[50] P.-H. Arnoux, A. Xu, N. Boyette, J. Mahmud, R. Akkiraju, and V. Sinha, "25 tweets to know you: A new model to predict personality with social media," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 472–475.

[51] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE intelligent systems*, vol. 32, no. 2, pp. 74–79, 2017.

[52] X. Sun, B. Liu, J. Cao, J. Luo, and X. Shen, "Who am i? personality detection based on deep learning for texts," in *2018 IEEE international conference on communications (ICC)*. IEEE, 2018, pp. 1–6.

[53] R. P. Tett, M. J. Toich, and S. B. Ozkum, "Trait activation theory: A review of the literature and applications to five lines of personality dynamics research," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 8, pp. 199–233, 2021.

[54] S. Ghassemi, T. Zhang, W. Van Breda, A. Koutsoumpis, J. K. Oostrom, D. Holtrop, and R. E. de Vries, "Unsupervised multimodal learning for dependency-free personality recognition," *IEEE transactions on affective computing*, 2023.

[55] 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.

[56] H. Peters and S. C. Matz, "Large language models can infer psychological dispositions of social media users," *PNAS Nexus*, vol. 3, no. 6, p. pgae231, 2024.

[57] L. Shen, Y. Long, X. Cai, G. Chen, I. Razzak, and S. Jameel, "Less but better: Parameter-efficient fine-tuning of large language models for personality detection," *arXiv preprint arXiv:2504.05411*, 2025.

[58] L. Hu *et al.*, "Llm vs small model? large language model based text augmentation enhanced personality detection model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024.

[59] S. M. H. Motlagh, M. H. Rezvani, and M. Khounsiavash, "Ai methods for personality traits recognition: A systematic review," *Neurocomputing*, p. 130301, 2025.

[60] Z. Wen *et al.*, "Affective-nli: Towards accurate and interpretable personality recognition in conversation," in *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2024.

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

[62] J. G. Rosse, M. D. Stecher, J. L. Miller, and R. A. Levin, "The impact of response distortion on preemployment personality testing and hiring decisions," *Journal of Applied Psychology*, vol. 83, no. 4, p. 634, 1998.

[63] S. Huber, N. B. E. Papenfuß, L. Weppert, V. Wohlfart, J. Basch, O. Happel, and T. Grundgeiger, "Hiring an ai: Incorporating personnel selection methods in user-centered design to design ai agents for safety-critical domains," in *Adjunct Proceedings of the 2024 Nordic Conference on Human-Computer Interaction*, October 2024, pp. 1–9.

[64] A. Fabris, N. Baranowska, M. J. Dennis, D. Graus, P. Hacker, J. Saldivar, and A. J. Biega, "Fairness and bias in algorithmic hiring: A multidisciplinary survey," *ACM Transactions on Intelligent Systems and Technology*, 2024.

[65] A. J. Barends and R. E. de Vries, "Developing and improving personality inventories using generative artificial intelligence: The psychometric properties of a short hexaco scale developed using chatgpt 4.0," *Journal of Personality Assessment*, pp. 1–7, 2024.

[66] W. Li, Y. Lin, M. Xia, and C. Jin, "Rethinking mixture-of-agents: Is mixing different large language models beneficial?" in *Language Gamification-NeurIPS 2024 Workshop*.

[67] D. Li, Z. Tan, P. Qian, Y. Li, K. S. Chaudhary, L. Hu, and J. Shen, "Smoa: Improving multi-agent large language models with sparse mixture-of-agents," *arXiv preprint arXiv:2411.03284*, 2024.

[68] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto, "Length-controlled alpacaeval: A simple way to debias automatic evaluators," *arXiv preprint arXiv:2404.04475*, 2024.

[69] C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, and D. Jiang, "Wizardlm: Empowering large language models to follow complex instructions," *arXiv preprint arXiv:2304.12244*, 2023.

[70] S. Ye, D. Kim, S. Kim, H. Hwang, S. Kim, Y. Jo, J. Thorne, J. Kim, and M. Seo, "Flask: Fine-grained language model evaluation based on alignment skill sets," in *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

[71] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th annual acm symposium on user interface software and technology*, 2023, pp. 1–22.

[72] J. Li, Y. Lai, W. Li, J. Ren, M. Zhang, X. Kang, S. Wang, P. Li, Y.-Q. Zhang, W. Ma *et al.*, "Agent hospital: A simulacrum of hospital with evolvable medical agents," *arXiv preprint arXiv:2405.02957*, 2024.

[73] 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.

[74] 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.

[75] K. Breevaart and R. E. de Vries, "Followers' hexaco personality traits and preference for charismatic, relationship-oriented, and task-oriented leadership," *Journal of Business and Psychology*, vol. 36, no. 2, pp. 253–265, 2021.

[76] A. Koutsoumpis, T. Zhang, J. K. Oostrom, D. Holtrop, and R. E. De Vries, "Open dataset AVI-6: Annotated asynchronous video interviews," https://doi.org/10.17605/OSF.IO/XTUYQ, 2025, oSF.

[77] 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.

[78] J.-I. Biel, O. Aran, and D. Gatica-Perez, "You are known by how you vlog: Personality impressions and nonverbal behavior in youtube," in

Proceedings of the International AAAI Conference on Web and Social Media, vol. 5, no. 1, 2011, pp. 446–449.

[79] S. T. Sreenivas, S. Muralidharan, R. Joshi, M. Chochowski, M. Patwary, M. Shoeybi, P. Molchanov *et al.*, "Llm pruning and distillation in practice: The minitron approach," *arXiv preprint arXiv:2408.11796*, 2024.

[80] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, J. Lin *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[81] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[82] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[83] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[84] D. Jain, A. Kumar, and R. Beniwal, "Personality bert: a transformer-based model for personality detection from textual data," in *Proceedings of international conference on computing and communication networks: ICCCN 2021*. Springer, 2022, pp. 515–522.

[85] S. Lim, S. Lee, D. Min, and Y. Yu, "Persona dynamics: Unveiling the impact of persona traits on agents in text-based games," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 31 360–31 394.

[86] R. Hussain, J. Ma, R. Khandelwal, J. Oltmanns, and M. Gupta, "Personality prediction from life stories using language models," *arXiv preprint arXiv:2506.19258*, 2025.

[87] R. R. McCrae, J. E. Kurtz, S. Yamagata, and A. Terracciano, "Internal consistency, retest reliability, and their implications for personality scale validity," *Personality and social psychology review*, vol. 15, no. 1, pp. 28–50, 2011.

[88] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[89] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[90] T. Zhang, T. Qi, A. Koutsoumpis, Y. Zong, W. Zheng, J. K. Oostrom, D. Holtrop, Z. Luo, and R. E. de Vries, "Assessing personality traits and interview performance from asynchronous video interviews," in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 13895–13900. [Online]. Available: https://doi.org/10.1145/3746027.3762016

Tianyi Zhang (Senior Member, IEEE) 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 affective computing and personality recognition. He is an Associate Editor of *IEEE Transactions on Affective Computing*.

Shan Liang receive her PhD degree at the Management School of Nanjing University of Information Science and Technology. She is now an assistant professor at the Business School of Nanjing Xiaozhuang University. Her research interests include human resource management, Leadership, and AI-based personality assessment.

Wenming Zheng (Senior Member, IEEE) received the B.S. degree in computer science from Fuzhou University, the M.S. degree in computer science from Huaqiao University, and the Ph.D. degree in signal processing from Southeast University. Since 2004, he has been with the Research Center for Learning Science, Southeast University. He is currently a Full Professor with the Key Laboratory of Child Development and Learning Science, Ministry of Education, Southeast University. His research interests include affective computing, pattern recognition, machine learning, and computer vision. n. He has been elected as an IET Fellow since 2022. He is an Associate Editor of *IEEE Transactions on Affective Computing* and the Editorial Board Member of the *Visual Computer*.

Antonis Koutsoumpis works as a postdoc researcher at the School of Business and Economics, at the Vrije Universiteit Amsterdam, the Netherlands. He received his PhD at the Organizational Psychology Section of the Experimental and Applied Psychology Department, Vrije Universiteit Amsterdam. His research interests include automatic personality assessment from asynchronous video interviews as well as the verbal and nonverbal 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 (e.g., asynchronous video interviews, situational judgment tests), with the goal to make them more predictive of future work behaviors, while reducing discrimination against marginalized groups. She received her Master's (2005) and Ph.D. (2010) degree in Work and Organizational Psychology from the Erasmus University Rotterdam. She is the Co-Editor-in-Chief of the International Journal of Selection and Assessment.

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. His current work focuses on the Situation-Trait-Outcome Activation (STOA) model, the Three Nightmare Traits (TNT) of personality, and the automatic assessment of personality using algorithms. Reinout is on the editorial (review) boards of The Leadership Quarterly, the European Journal of Personality, the International Journal of Selection and Assessment, and various other journals, and he has published more than 100 articles on personality and organizational topics in a wide range of scientific journals.