

Inferring personality from a live interaction with an LLM: Predictions are internally consistent and show convergent validity

Merlin Monzel

`merlin.monzel@uni-bonn-diff.de`

University of Bonn

Sebastian Jung

University of Bonn

Martin Reuter

University of Bonn

Article

Keywords: large language model, personality assessment, Big Five, prompting, response biases

Posted Date: August 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10495841/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Personality assessment typically relies on psychometric questionnaires, which are efficient but overlook the contextual richness of natural behavior. Here, we investigate whether large language models (LLMs) can infer Big Five personality traits from live interactions. In a preregistered study, 117 participants engaged in a 10-minute conversation with an LLM instructed to provoke information useful for personality assessment. Personality was assessed using non-contextual and contextual prompting, one providing no additional information while the other provided the items of the Big Five Inventory II for additional context. LLM-derived trait scores showed high internal consistency comparable to self-reports and demonstrated convergent validity, particularly for neuroticism and agreeableness. Predictive performance showed weaker results for extraversion and openness, consistent with lower linguistic observability. Non-contextual prompting proved more robust, whereas contextual prompting suffered from higher variability and technical failures. Further analyses revealed that specific subfacets, such as depression and compassion, drove prediction accuracy. While the LLM exhibited a leniency bias towards socially desirable traits, no evidence of gender bias was found, indicating that participants' genders could be successfully concealed from the LLM. Together, these findings suggest that LLMs can extract personality indicators from conversational data, while highlighting constraints related to prompting, trait visibility, and systematic biases.

Introduction

Personality assessment is commonly conducted using psychometric questionnaires (e.g., BFI-2) due to their cost-efficiency, ease of administration, and straightforward conversion of responses into numerical scores through Likert-type scales¹. However, self-report measures are also susceptible to introspection errors² and various response biases (e.g., social desirability³), which may compromise the accuracy of the obtained personality estimates. In addition, such instruments often fail to capture contextual nuances and rich qualitative information that may emerge in textual data and natural language reports⁴.

In contrast to psychometric questionnaires, qualitative assessment formats remain relatively underutilized because they present substantial practical challenges. These include the transcription, coding, and quantification of textual data⁵, as well as concerns regarding reliability stemming from potential disagreements among human raters⁶. Automated natural language processing approaches address some of these limitations but frequently rely on predefined dictionaries, which may lack the flexibility required to interpret language within its specific context⁷.

Recent advances in large language models (LLMs) offer a promising alternative. Wright et al. demonstrate that LLMs can leverage contextual embeddings to capture semantic meaning more effectively, thereby improving predictive validity in personality assessment⁸. At the same time, Monzel et al. found that the predictive performance of LLM-based assessments may depend on the nature of the input data and the extent to which it aligns with the intended diagnostic objective⁹. In addition, Matšenas et al. suggest, based on a small-scale sample, that real-time guided conversations may be more advantageous than passive text analysis¹⁰.

Building on this emerging literature, the present study examines whether predictive validity in the assessment of the Big Five personality traits can be enhanced through a live interaction with a chatbot specifically prompted to elicit information relevant to the Big Five dimensions. By enabling an adaptive and context-sensitive conversational exchange, this approach may provide richer personality-relevant information than the text sources deployed by Wright et al. (i.e., spontaneous streams of thought and daily video diaries)⁸. In addition, we will explicitly address whether the limitations of traditional self-report questionnaires, such as response biases, can be circumvented with LLMs, and which subfacets of the Big Five personality traits can be estimated most accurately from conversational data.

Personality assessment with large language models

The Big Five personality model conceptualizes personality along five broad dimensions: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism¹¹. These traits have been widely adopted in psychological research as a robust framework for describing and comparing personality characteristics across individuals. Openness reflects curiosity and appreciation towards novel experiences and consists of aesthetic sensitivity, intellectual curiosity, and creative imagination. Conscientiousness describes goal-directed behavior, self-discipline, and reliability, including organization, productivity, and responsibility. Extraversion captures social engagement and energetic behavior, seen in sociability, assertiveness, and energy level. Agreeableness represents the degree of interpersonal adaptability, including compassion, respectfulness, and trust. Neuroticism depicts emotional instability and vulnerability, such as anxiety, depression, and emotional volatility¹.

Previous studies suggest that the Big Five personality traits may be reflected in linguistic behavior, giving rise to what is often referred to as linguistic fingerprint¹². Language use can provide indirect indicators of personality through patterns in vocabulary, syntax, emotional expression, and communication style. For example, neuroticism is usually identified through more first-person language and display of negative emotions, while extraversion is typically shown by positively connoted and social words. Conscientiousness is associated with structural and formal language, while agreeableness is connected to prosocial and warm language. Finally, openness is indicated by more complex and abstract wordings¹³.

Recently, Wright et al. showed that LLMs can detect these linguistic fingerprints, demonstrating significant correlations between self-reported Big Five personality traits and LLM-assessed Big Five personality traits based on 10-minute thought streams and daily diaries⁸. Across all Big Five personality traits, LLMs had average convergent correlations with self-reports of $r = .48$ and $r = .44$, respectively. Results were specifically strong for extraversion, reaching correlations of $r = .57$. Internal consistency was high across all dimensions, with Cronbach's α ranging from .76 to .88, comparable to self-reports.

Possible limitations of LLMs for personality assessment

Despite the potentially promising applications of LLMs in personality assessment, several limitations must be considered. First, LLMs are known to reproduce stereotypes embedded in their training data¹⁴, which may lead to systematic response biases. For example, if an LLM reproduces gender stereotypes, it may be more

likely to assign higher neuroticism scores to female participants. Although women tend to score higher on neuroticism than men at the group level¹⁵, an LLM may overgeneralize this population-level pattern to individual cases, thereby amplifying gender-related biases in personality assessments. Similar concerns apply to other Big Five personality traits as large cross-cultural studies have reported higher levels of agreeableness and extraversion among women¹⁵. An LLM may disproportionately rely on such group-level associations when generating personality judgments, potentially exaggerating observed gender differences and reducing the accuracy of individual-level assessments. However, concealing the participants' gender from the LLM may prevent these effects.

Second, LLMs are known to generate responses that align with user expectations and preferences¹⁶, which may introduce a tendency toward overly favorable or socially desirable evaluations. This leniency bias may inflate scores on socially desirable traits such as agreeableness and conscientiousness while yielding comparatively lower scores on neuroticism, which is often perceived as unfavorable (see ¹⁷). For extraversion and openness, on the other hand, no clearly socially desirable directions exist. While a certain level of these traits is typically perceived as socially desirable, they become maladaptive when expressed excessively^{18,19}.

Hypotheses

Given the above-mentioned observations, we propose the following hypotheses, from which the second was preregistered:

H1: The LLM understands the internal structure of the Big Five personality traits as indicated by internal consistencies in the response patterns of the LLM being similar to those in self-report data.

H2: The self-reported Big Five personality traits can be predicted by the LLM based on a 10-minute live interaction.

H3: The LLM shows a leniency bias, assigning higher scores in agreeableness and conscientiousness and lower scores in neuroticism.

With regard to gender biases, we expect that concealing our participants' gender from the LLM prevented it from reproducing gender stereotypes. These null effects will be examined using Bayesian statistics providing evidence for the null hypothesis:

H4: Gender differences in the LLM-assessed Big Five personality traits are not bigger than in the self-reported Big Five personality traits.

Besides these hypotheses, we will explore the influence of different prompting strategies, that is, a non-contextual prompt asking for a simple evaluation of the Big Five personality traits without further explanation, and a contextual prompt, providing the items of the Big Five Inventory 2, on which the participants are then evaluated by the LLM. Last, we will investigate which Big Five subfacets show the strongest correlations with the LLM's assessments, in order to identify which subfacets are most discernible in linguistic fingerprints.

Results

Internal consistency

Before examining whether our model can predict our participants’ personality traits, we checked whether it can understand the internal structure of our questionnaire. For this, Cronbach’s Alpha for each dimension of the BFI-2 was calculated. As depicted in Table 1, internal consistencies ranged from acceptable to good and were comparable to those observed in self-report data. Extraversion was the only trait showing a notable deviation, with only limited overlap in the 95% confidence intervals. Nevertheless, all BFI-2 subscales demonstrated adequate internal consistency.

Table 1
Cronbach’s Alpha for each BFI-2 subscale for self-report and LLM data as well as correlations between the traits assessed via contextual and non-contextual prompting strategies.

Trait	95% CI			95% CI			$r_{context, no context}$	p
	$\alpha_{Self-Report}$	LL	UL	α_{LLM}	LL	UL		
Neuroticism	.87	.84	.90	.88	.81	.93	.53	< .001
Extraversion	.87	.84	.90	.77	.64	.87	.46	.004
Openness	.85	.81	.88	.83	.73	.91	.20	.237
Agreeableness	.83	.79	.86	.85	.76	.91	.56	< .001
Conscientiousness	.84	.81	.88	.83	.73	.90	.33	.045

Robustness of prompting strategies

From our initial sample of 126 participants, non-contextual prompting led to 117 usable records (= 92.9%), whereas our model failed to provide an answer or encountered technical issues in nine cases (= 7.1%). In contrast, contextual prompting led to 37 usable records (= 29.4%), whereas our model failed to provide an answer or encountered technical issues in 89 cases (= 70.6%). This difference between the prompting strategies was significant, $t(125) = 14.74, p < .001, g = 1.31$.

Next, we checked whether our two prompting strategies led to similar results. The correlations between the traits assessed via a single item and the BFI-2 dimensions are depicted in Table 1. With the exception of openness, each correlation was significant despite the small sample size of the contextual assessments ($n = 37$). Thus, at least some level of convergent validity can be assumed.

Predicting self-reports from our model’s assessments

Following the promising results regarding internal consistency, we calculated the preregistered correlations between our participants’ self-reports and our model’s assessments. Since the intercorrelations between the contextual assessments and the non-contextual assessment were limited, we decided to calculate the

correlations for both, the contextual assessments and the non-contextual assessments to examine which prompting strategy is better suited for predicting our participants' personality traits. A comparison between the predictions of the individual traits is depicted in Fig. 1. For the non-contextual assessment ($n = 117$), all correlations were significant (one-tailed p -values ranging from $< .001$ to $< .048$), whereas for the contextual assessment ($n = 37$), openness and extraversion could not be significantly predicted by our model (all other one-tailed p -values ranging from $< .001$ to $< .015$). Correlation comparisons between the prompting strategies did not reach significance for any trait (p -values $> .188$), most likely due to the small sample size of the contextual assessment (see Table S1, Supplemental Material).

Within the non-contextual assessment, most constructs could be assessed with similar precision (see Table S2, Supplemental Material). The only differences were between a more precise assessment of neuroticism compared to extraversion ($p = .017$), a more precise assessment of agreeableness compared to extraversion ($p = .005$) and a more precise assessment of agreeableness compared to conscientiousness ($p = .011$), indicating that neuroticism and agreeableness could be predicted most accurately, whereas extraversion and conscientiousness could be predicted least accurately.

Within the contextual assessment, agreeableness and conscientiousness could be predicted more precisely than the remaining traits (p s ranging from $< .001$ to $.030$). In addition, neuroticism could be predicted more precisely than extraversion ($p = .034$) and openness ($p = .003$). An overview of all correlation comparisons is depicted in Table S3 (see Supplemental Material).

Big Five subfacets driving the prediction performance

To examine which subfacets drove the predictive performance of our LLM, we calculated correlations between the LLM's Big Five assessments (non-contextual and contextual) and the subfacet scores reported by our participants. The results of these calculations are depicted in Table 2.

Table 2
Correlations between the LLM's Big Five assessments (non-contextual and contextual) and the subfacet scores reported by our participants.

Trait	$r_{\text{non-contextual}}$	p	$r_{\text{contextual}}$	p
Neuroticism	.35	< .001	.36	.003
Anxiety	.30	< .001	.22	.199
Depression	.30	.001	.43	.008
Emotional volatility	.28	.002	.31	.065
Extraversion	.15	.049	.20	.119
Sociability	.09	.336	.27	.107
Assertiveness	.14	.128	.08	.645
Energy level	.17	.072	.19	.251
Openness	.25	.003	.09	.302
Aesthetic sensitivity	.29	.002	.19	.265
Intellectual curiosity	.27	.003	.11	.519
Creative imagination	.00	.997	– .02	.889
Agreeableness	.40	< .001	.52	< .001
Compassion	.40	< .001	.47	.004
Respectfulness	.25	.006	.27	.106
Trust	.29	.001	.49	.002
Conscientiousness	.20	.063	.50	< .001
Organization	.17	.071	.33	.048
Productiveness	.14	.131	.50	.002
Responsibility	.11	.220	.47	.004

For the non-contextual assessment, neuroticism, conscientiousness, and extraversion were represented equally across all subfacets (correlation comparison p -values > .354). For agreeableness, the compassion subfacet stood out, at least numerically (p -values > .079). Regarding openness, the subfacet creative imagination received significantly less consideration by the LLM than the subfacets aesthetic sensitivity ($p = .006$) and intellectual curiosity ($p = .010$). All correlation comparisons between the subfacets and their corresponding higher-order Big Five personality traits are displayed in Table S4 (see Supplemental Material).

For the contextual assessment, the subfacet depression stood out in neuroticism compared to anxiety ($p = .002$). For extraversion, the subfacet assertiveness received relatively little consideration by the LLM compared to sociability ($p = .012$). For agreeableness and conscientiousness, the subfacets respectfulness

lagged behind compassion and trust ($ps < .020$), and organization lagged behind responsibility ($p = .037$), although all correlations were still moderate in size.

LLM’s response biases

To assess leniency and gender biases in our model’s assessment, ANOVAs were calculated for each personality trait with gender (female vs. male) and data source (self-report vs. non-contextual assessment) as independent variables. For these analyses, contextual assessments were not considered due to their small sample sizes. Means, standard deviations, and the results of each ANOVA are depicted in Table 3.

Table 3
Two-Way ANOVA Statistics for the BFI-2 personality traits.

Variable	Self-report		LLM		ANOVA				
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	Effect	<i>F</i> (1, 115)	<i>p</i>	η_p^2	<i>BF</i> ₀₁
Neuroticism					Data source	17.79	< .001	.13	< .01
Male	2.56	0.72	2.16	0.65	Gender	7.61	.007	.06	0.20
Female	2.79	0.66	2.53	0.78	Data source × Gender	.981	.324	.01	3.13
Extraversion					Data source	13.90	.001	.11	< .01
Male	3.50	0.72	3.34	0.65	Gender	1.42	.236	.01	< .01
Female	3.53	0.63	3.10	0.65	Data source × Gender	2.87	.093	.02	1.12
Openness					Data source	10.72	.001	.09	0.11
Male	3.64	0.59	3.98	0.66	Gender	1.69	.196	.01	2.58
Female	3.87	0.57	3.99	0.57	Data source × Gender	2.42	.123	.02	1.75
Agreeableness					Data source	22.72	< .001	.17	< .01
Male	3.82	0.54	4.27	0.87	Gender	1.28	.261	.01	2.91
Female	4.05	0.48	4.29	0.77	Data source × Gender	2.09	.151	.02	2.07
Conscientiousness					Data source	57.38	< .001	.33	< .01
Male	3.39	0.63	4.02	0.59	Gender	3.41	.067	.03	1.64
Female	3.65	0.60	4.07	0.48	Data source × Gender	2.37	.126	.02	1.61

Main effects of gender were only found for neuroticism. However, as the differences were similar for the self-reports and assessments of the LLM, these cannot be characterized as gender biases of the LLM.

Importantly, for none of the Big Five personality traits, interaction effects could be found, which would indicate gender differences of varying size for the assessments of the LLM compared to the self-reports. However, the Bayes factors suggest only anecdotal evidence in favor of the null hypothesis.

With regard to leniency biases, main effects of data source were found for each personality trait. For neuroticism, agreeableness, and conscientiousness, these differences were in the expected direction, that is, the LLM was more lenient in the direction of social desirability. For openness, the LLM also gave higher ratings, whereas for extraversion, the LLM's assessments were significantly lower than the self-reports.

Discussion

As expected, internal consistencies of the LLM-assessed Big Five personality traits were high, ranging from .77 for extraversion to .88 for neuroticism, and were similar to those in self-report data, ranging from .83 for agreeableness to .87 for neuroticism. This indicates that our model “understands” the structure of human personality and can answer related questions consistently (acceptance of hypothesis 1). Importantly, however, these assessments were not only internally consistent but also largely demonstrated convergent validity with self-reports of personality (acceptance of hypothesis 2). Correlations ranged from .15 to .40 for non-contextual assessment and from .09 to .52 for contextual assessment. With this, predictive validity showed larger variability than the effect sizes found in Wright et al. but a similar ceiling⁸, indicating that prediction quality largely depends on the prompting strategy, respective personality dimensions, and potential biases, which will be discussed further below. A substantial increase in predictive validity due to the live interaction compared to spontaneous streams of thought and daily video diaries used in Wright et al. could not be observed⁸, probably due to the limited amount and representativeness of linguistic data obtained in this short interaction, which may not sufficiently capture stable personality-related patterns²⁰. In the future, more specific prompts will need to be used to better elicit certain behavioural indicators of the Big Five personality traits during interactions with the LLM (e.g., prompting the LLM to ask about the social life of the participants).

Dependency on prompting strategy

With the exception of openness, Big Five personality traits showed convergent validity between contextual and non-contextual assessments, indicating that LLMs can assess Big Five personality traits without the need for additional context. This was also evident in the non-significant correlation comparisons between contextual and non-contextual assessments of Big Five personality traits. Thus, our model was able to assess Big Five personality traits without us defining these traits, indicating that this knowledge can be retrieved from the training data. Openness, however, may be more complex and thus often less comprehensively defined in the training literature, leading to less overlap between the non-contextual assessment and the specific items of the BFI-2 used in the contextual assessment (for a discussion, see ²¹). In addition, the overlap among the other Big Five personality traits was only moderate, ranging from .33 for conscientiousness to .56 for agreeableness, indicating that there is also substantial methodological variance stemming from the prompting technique.

While all Big Five personality traits could be validly predicted using non-contextual assessment, openness and extraversion could not be validly predicted using contextual assessment. However, this may be the case due to the smaller sample size of contextual assessment. While non-contextual assessment was mostly stable, contextual assessment resulted in over 70% of assessments failing. This might have been caused by an insufficient context window of the LLM due to submitting all BFI-2 items at once. Future research should thus employ a step-by-step routing as described in Wright et al., submitting one item after another⁸. Further evidence for the more robust assessment of non-contextual prompting is the narrower range of convergent relationships, whereas contextual assessment showed a larger range. Contextual assessment, on the other hand, provided a higher ceiling for predictive validity of agreeableness and conscientiousness, suggesting that non-contextual assessment yielded a less comprehensive understanding of the constructs or less precise gradations due to the single-item design, which may have limited its effect sizes.

Dependency on personality traits

Beyond the effects of the prompting strategy, differences in the predictability of the Big Five personality traits could be observed. For the non-contextual assessment, neuroticism and agreeableness could be predicted most accurately, whereas extraversion and conscientiousness were more difficult for the LLM to predict. Importantly, for neuroticism, extraversion, and conscientiousness, these results were reflected across all subfacets, whereas for openness, the creative imagination subfacet contributed the least to predictability. For agreeableness, the predictability was mainly driven by compassion, at least numerically.

The differences in the predictability are most likely to be explained by their linguistic observability. Neuroticism and agreeableness are constructs with strong linguistic markers such as the expression of negative emotions and prosocial or affiliative language. In particular, the compassion facet of agreeableness may play a central role, as it is closely linked to empathic and supportive language¹³. In contrast, extraversion and conscientiousness are more strongly expressed in overt behaviour (e.g., social activity, goal-directed behaviour) and are therefore less directly observable in linguistic data. Sociability, for example, is expressed in the “desire to socially approach and engage with others”¹, a behavior that cannot be observed in non-human interaction. The same is true for the openness subfacet of creative imagination, which is rather unlikely to be elicited in a non-guided interaction with an LLM, or to be masked as belonging to the participants’ inner world rather than their linguistic expression²².

For the contextual assessment, the LLM was also struggling with extraversion and openness, supporting the idea that these constructs are less observable in linguistic data. This time, however, assertiveness fell behind sociability. This indicates that the BFI-2 items were indeed supporting the model to explicitly retrieve information about the sociability of the participants, once again highlighting the importance of contextual guidance. As with non-contextual assessment, neuroticism and agreeableness could be assessed precisely, using contextual prompts, especially in the subfacets depression, compassion and trust. This indicates that, for these subfacets, there was the closest alignment between the BFI-2 items and the behaviors observable in the interaction, whereas other subfacets may need to be elicited more directly in future research. Surprisingly, however, in contrast to non-contextual assessment, conscientiousness could be predicted most accurately in the contextual assessment, especially the subfacets responsibility, indicating that the LLM’s

natural understanding of conscientiousness may be underdeveloped, explaining its need to rely on the specific behavioral indicators of the BFI-2.

Differential effects between the Big Five personality traits were also observed in the LLM's leniency biases. All effects were in the expected direction (acceptance of hypothesis 3). In line with social desirability¹⁷, the LLM showed significantly higher scores in agreeableness and conscientiousness and lower scores in neuroticism, indicating a tendency toward benevolent assessments (see ¹⁶).

Importantly, although LLM leniency biases affect the absolute scores on Big Five personality assessments, this should not alter the relative ranking of participants and therefore should not impair the predictive validity of our results. The expression of gender stereotypes, on the other hand, would only affect subsamples of our participants (i.e., females and males differently), which could compromise the validity of the LLM's predictions.

Dependency on gender

In line with our expectations, our model did not amplify gender stereotypes (acceptance of hypothesis 4), indicating that our participants' gender could be successfully concealed from the LLM, thereby preventing stereotypical responses typical of LLMs¹⁴. While differences in neuroticism were observed between male and female participants, these differences were consistent across self-reports and LLM-based assessments, suggesting real differences between female and male participants¹⁵ rather than systematic biases. Importantly, however, the Bayes factors suggest only anecdotal evidence in favor of the null hypothesis. Nevertheless, even if a larger sample size would yield significant effects, these would likely be small, as shown in Table 3.

Limitations and ethical considerations

The use of LLMs for personality inference introduces concerns regarding algorithmic bias and fairness. Prior research suggests that language models may reproduce and amplify societal stereotypes present in their training data, potentially leading to systematic distortions in personality assessment (e.g., gender stereotypes)¹⁴. Such biases may result in unfair or inaccurate evaluations of individuals and must therefore be critically monitored and transparently reported. Blinding identity-related information, as was done for gender in the present study, appears to be crucial. Nevertheless, due to limitations in validity, LLM personality predictions should not yet be treated as diagnostic or definitive. This is particularly important in applied contexts such as personnel selection or performance evaluation, where inappropriate use could lead to harmful consequences for individuals and whole organizations. In addition, transparency in model usage and limitations is crucial. Participants and stakeholders should be made aware that LLM-based assessments depend on model architecture, prompting strategies, and interaction context, which may influence the results. Researchers should therefore disclose methodological choices and ensure that findings are interpreted within their limits.

In the present study, only one model, Meta Llama 3.3 80B Instruct, was tested, so the results may not be generalizable to other LLMs. Our sample size was larger than in previous comparable studies, but due to technical limitations, it was significantly smaller for contextual assessments. Nevertheless, the sample size

for the contextual assessment was still larger than what a power analysis using the estimates from Wright et al. would require⁸. Importantly, the sample consisted mostly of psychology students of the University of Bonn. Given that psychology students are generally more homogeneous in their Big Five personality traits than a general population sample (e.g., showing more neuroticism)²³, it is impressive that significant correlations were found despite these variance restrictions.

Conclusion

Our findings demonstrate that LLMs can recover stable, meaningful personality signals from brief conversational interactions, achieving high internal consistency and moderate convergent validity, though these remain to be improved. At the same time, performance depends strongly on the interplay between prompting strategy and the linguistic observability of specific traits. While biases such as leniency exist, they should not impact the relative ranking of participants. Moreover, concealing participants' gender from the LLM helped prevent gender bias, underscoring the potential of carefully designed interaction settings. Together, our results highlight both the promise and current limitations of LLM-based personality inference and suggest that future progress will depend on more targeted prompting, richer interaction data, and continued scrutiny of model biases before application in real-world contexts.

Method

The project was preregistered on the Open Science Framework: <https://osf.io/6zucs>. The preregistration included information about the hypothesis 2, data collection procedure, outlier exclusion, and statistical analyses. The present report focusses on the live interaction with the LLM and the assessment of the Big Five personality traits.

Ethics approval

was obtained from the Ethics Committee of the Institute of Psychology of the University of Bonn. Informed consent to participate and informed consent for publication was obtained from all participants on the day of participation. The research was performed in accordance with the Declaration of Helsinki.

Sample

An a-priori power analysis was conducted using G*Power (Correlation: Bivariate normal model) to determine the minimum required sample size. Based on an expected effect size of $r = .24$, a significance level of $\alpha = .05$ (one-tailed), and a desired statistical power of $1 - \beta = .80$, the analysis indicated a minimum sample size of 106 participants²⁴. The expected effect size was derived from the average predictive validity of the Big Five personality traits reported by Grunenberg et al. who used CVs instead of a live interactions as source data²⁵. Using the recently published results by Wright et al., who provided an expected average effect size of $r = .48$, even a sample size of 25 participants would be sufficient⁸. By the end of the pre-specified data collection period, data had been obtained from 117 participants, thereby exceeding the required minimum sample size of both estimations. Participants were mainly recruited from the student body of the Institute of Psychology

of the University of Bonn. 62.4% of our sample identified as female, 37.6% as male. Age ranged between 18 and 67 ($M = 26.8$, $SD = 10.7$).

Questionnaires

For the self-assessment of the Big Five personality traits, the Big Five Inventory 2 (BFI-2) was used¹. The BFI-2 is an openly available self-report instrument comprising 60 items on a five-point Likert scale ranging from “strongly disagree” to “strongly agree”. Each Big Five personality trait is measured by twelve items, divided into three subfacets comprising four items each. Neuroticism (or negative emotionality) is subdivided into anxiety, depression and emotional volatility; extraversion is subdivided into sociability, assertiveness and energy level; openness (or open-mindedness) is subdivided in intellectual curiosity, aesthetic sensitivity and creative imagination; agreeableness is subdivided into compassion, respectfulness and trust; and conscientiousness is subdivided into organization, productiveness and responsibility.

LLM assessment

Participants were instructed to interact with our model (Meta Llama 3.3 70B Instruct) for full 10-minutes. Temperature and top_p were set to 0 and 0.05, respectively, to ensure standardized communication between our participants and our model. Temperature controls the randomness of token sampling, where tokens represent the basic text units processed by the language model. The top_p parameter restricts token selection by considering only candidates within a predefined cumulative probability mass. Setting these parameters to their minimum values increases the likelihood that the same input yields the same output, making the assessment more reliable (see ⁹). The conversation was started with the following prompt:

Have a 10-minute conversation with me about the past week so that, afterward, you can assess my personality [...] as accurately as possible. Ask open-ended questions and keep your own contributions as brief as possible. Let me know when the 10 minutes are up.

After the conversation, the following two assessment prompts were submitted to the LLM:

Non-contextual assessment: *“Based on our previous conversation, excluding the initial prompt and this message: Estimate my personality on the Big Five personality traits (i.e., Extraversion, Conscientiousness, Agreeableness, Neuroticism, and Openness to Experience). Use a scale from 1 to 5 for each trait. Respond only with the numbers and provide no text or explanations. Separate the numbers with commas. Give the five numbers in the following order: Extraversion, Conscientiousness, Agreeableness, Neuroticism, Openness to Experience.”*

Contextual assessment: *“Based on our previous conversation, excluding the initial prompt and this message: Estimate how strongly I would agree with each of the following 60 statements on a scale from 1 to 5, where 1 means “strongly disagree,” 2 means “somewhat disagree,” 3 means “neither agree nor disagree,” 4 means “somewhat agree,” and 5 means “strongly agree.” Respond only with the numbers and provide no text or explanations. Separate the numbers with commas. For statements 1–60, give the corresponding rating (1–5) in the same order as the questions. The 60 statements are: {full list of the BFI-2 items}”*

The values were then transferred to our survey software to be stored in the dataset.

Procedure

All questionnaires as well as demographic questions (age, gender) were applied using the online software SosciSurvey²⁶. Our LLM was deployed through Chat AI and hosted on the servers of the Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) via Academic Cloud. In contrast to commercial chatbot services such as ChatGPT, all data displayed in the browser is stored exclusively on the client side and remains within the user's browser. Data are transmitted to the server only when a user submits a request that requires processing by the backend models. Once processing is complete and the browser session ends, no user inputs or session data are retained.

Declarations

Acknowledgements

I would like to express my sincere gratitude to the students who contributed to the data collection for this project. Their valuable assistance, dedication, and careful work were essential to the successful completion of this research. I especially thank Kseniia Akimbekova, Jennifer Eiring, Linn Hartmann, Sophia Myrianthopoulou, and Philippa Tuszik for their support in collecting the data, as well as for their commitment and reliability throughout the research process. I greatly appreciate their contributions and the time and effort they invested in this project.

CRedit Author Statement

Merlin Monzel: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – Original draft, Writing – Review & Editing. **Sebastian Jung:** Writing – Original draft. **Martin Reuter:** Resources, Supervision, Writing – Review & Editing.

Data availability statement

The data and code for all analyses are available at https://osf.io/dujeg/overview?view_only=3760f361b3784856a28f846da78e34f4.

Ethics statement

Ethics approval was obtained from the Ethics Committee of the Institute of Psychology of the University of Bonn. Informed consent to participate and informed consent for publication was obtained from all participants on the day of participation. The research was performed in accordance with the Declaration of Helsinki.

Competing interest statement

No competing interests to declare.

Funding declaration

No funding was received for this work.

References

1. Soto, C. J. & John, O. P. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *J. Pers. Soc. Psychol.* **113**, 117–143 (2017).
2. Nisbett, R. E. & Wilson, T. D. Telling more than we can know: Verbal reports on mental processes. *Psychol. Rev.* **84**, 231–259 (1977).
3. Paulhus, D. L. Measurement and Control of Response Bias. in *Measures of Personality and Social Psychological Attitudes* 17–59 Elsevier, (1991). 10.1016/B978-0-12-590241-0.50006-X
4. Park, G. et al. Automatic personality assessment through social media language. *J. Pers. Soc. Psychol.* **108**, 934–952 (2015).
5. Grimmer, J. & Stewart, B. M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Polit Anal.* **21**, 267–297 (2013).
6. Lombard, M., Snyder-Duch, J. & Bracken, C. C. Content Analysis in Mass Communication: Assessment and Reporting of Inter-coder Reliability. *Hum. Commun. Res.* **28**, 587–604 (2002).
7. Tausczik, Y. R. & Pennebaker, J. W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *J. Lang. Soc. Psychol.* **29**, 24–54 (2010).
8. Wright, A. G. C. et al. Assessing personality using zero-shot generative AI scoring of brief open-ended text. *Nat. Hum. Behav.* **10**, 541–555 (2026).
9. Monzel, M., Hogeterp, P. & Reuter, M. LLMs in natural language analysis: Inferring political orientation from political speeches and CVs. (2026).
10. Matšenas, A., Lello, A., Lees, T., Peep, H. & Tamm, K. L. Can LLMs Assess Personality? Validating Conversational AI for Trait Profiling. Preprint at (2026). <https://doi.org/10.48550/arXiv.2602.15848>
11. Goldberg, L. R. An alternative ‘description of personality’: The Big-Five factor structure. *J. Pers. Soc. Psychol.* **59**, 1216–1229 (1990).
12. Pennebaker, J. W. & King, L. A. Linguistic styles: Language use as an individual difference. *J. Pers. Soc. Psychol.* **77**, 1296–1312 (1999).
13. Yarkoni, T. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *J. Res. Personal.* **44**, 363–373 (2010).
14. Zhao, J. et al. Gender Bias in Contextualized Word Embeddings. in *Proceedings of the Conference of the North* 629–634 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). 629–634 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). (2019). 10.18653/v1/N19-1064
15. Costa, P. T., Terracciano, A. & McCrae, R. R. Gender differences in personality traits across cultures: Robust and surprising findings. *J. Pers. Soc. Psychol.* **81**, 322–331 (2001).
16. Ouyang, L. et al. Training language models to follow instructions with human feedback. Preprint at. <https://doi.org/10.48550/ARXIV.2203.02155> (2022).
17. Ones, D. S., Viswesvaran, C. & Reiss, A. D. Role of social desirability in personality testing for personnel selection: The red herring. *J. Appl. Psychol.* **81**, 660–679 (1996).

18. Ames, D. R. & Flynn, F. J. What breaks a leader: The curvilinear relation between assertiveness and leadership. *J. Pers. Soc. Psychol.* **92**, 307–324 (2007).
19. Le, H. et al. Too much of a good thing: curvilinear relationships between personality traits and job performance. *J. Appl. Psychol.* **96**, 113 (2011).
20. Grimmer, J. & Stewart, B. M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Polit Anal.* **21**, 267–297 (2013).
21. DeYoung, C. G. Openness/intellect: A dimension of personality reflecting cognitive exploration. in *APA handbook of personality and social psychology, Volume 4: Personality processes and individual differences*. (eds Mikulincer, M., Shaver, P. R., Cooper, M. L. & Larsen, R. J.) 369–399 American Psychological Association, Washington, (2015). 10.1037/14343-017
22. Monzel, M., Digard, B. G. & Zeman, A. The diversity of consciousness: Interindividual differences in mental imagery. in Reference Module in Neuroscience and Biobehavioral Psychology (Elsevier, doi:10.1016/B978-0-443-29258-3.00044-6. (2026).
23. Vedel, A. Big Five personality group differences across academic majors: A systematic review. *Personal Individ Differ.* **92**, 1–10 (2016).
24. Faul, F., Erdfelder, E., Lang, A. G., Buchner, A. & G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods.* **39**, 175–191 (2007).
25. Grunenberg, E., Peters, H., Francis, M. J., Back, M. D. & Matz, S. C. Machine learning in recruiting: predicting personality from CVs and short text responses. *Front. Soc. Psychol.* **1**, 1290295 (2024).
26. Leiner, D. J. SoSci Survey (Version 3.2.30) [Computer software]. (2021).

Figures

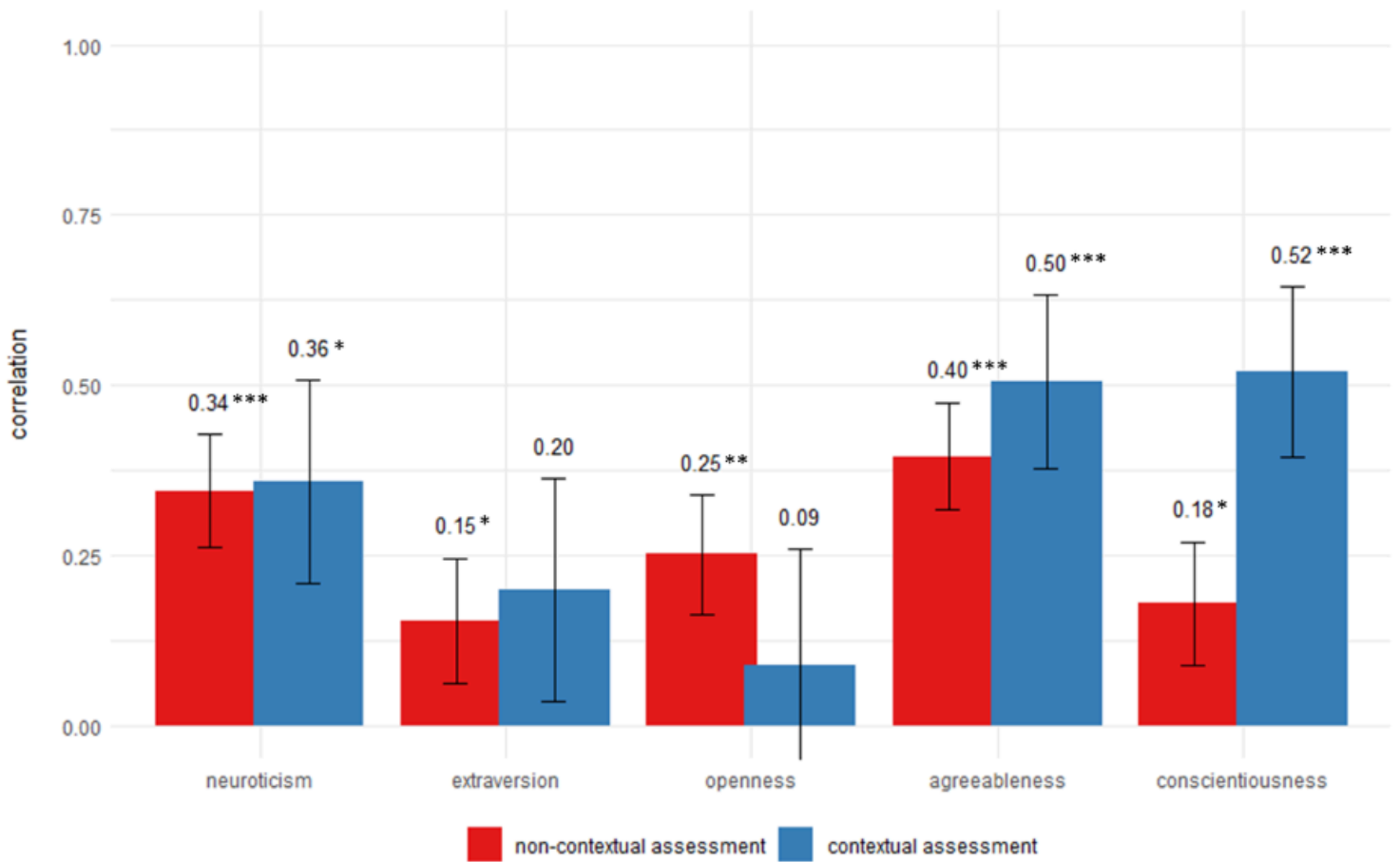


Figure 1

Depicted are correlations (± 1 SE) between self-reported personality traits and our model’s non-contextual (red) or contextual assessments (blue). * $p < .05$, ** $p < .01$, *** $p < .001$.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementalMaterial.docx](#)