PNAS Nexus, 2024, 3, pgae231
https://doi.org/10.1093/pnasnexus/pgae231
Advance access publication 13 June 2024
Research Report

Large language models can infer psychological dispositions of social media users

Heinrich Peters* and Sandra C. Matz

Columbia Business School, Columbia University, New York, NY 10027, USA
*To whom correspondence should be addressed: Email: hp2500@columbia.edu
Edited By: Michele Gelfand

Abstract

Large language models (LLMs) demonstrate increasingly human-like abilities across a wide variety of tasks. In this paper, we investigate whether LLMs like ChatGPT can accurately infer the psychological dispositions of social media users and whether their ability to do so varies across socio-demographic groups. Specifically, we test whether GPT-3.5 and GPT-4 can derive the Big Five personality traits from users' Facebook status updates in a zero-shot learning scenario. Our results show an average correlation of $r = 0.29$ (range = [0.22, 0.33]) between LLM-inferred and self-reported trait scores—a level of accuracy that is similar to that of supervised machine learning models specifically trained to infer personality. Our findings also highlight heterogeneity in the accuracy of personality inferences across different age groups and gender categories: predictions were found to be more accurate for women and younger individuals on several traits, suggesting a potential bias stemming from the underlying training data or differences in online self-expression. The ability of LLMs to infer psychological dispositions from user-generated text has the potential to democratize access to cheap and scalable psychometric assessments for both researchers and practitioners. On the one hand, this democratization might facilitate large-scale research of high ecological validity and spark innovation in personalized services. On the other hand, it also raises ethical concerns regarding user privacy and self-determination, highlighting the need for stringent ethical frameworks and regulation.

Keywords: LLMs, ChatGPT, GPT-4, personality, Big Five

Significance Statement

The rapid adoption of large language models (LLMs) in research and practice raises concerns about the balance between their potential and perils. Our study suggests that LLMs like ChatGPT are developing nuanced representations of human psychology and capacities for psychological profiling. While these advancements offer unparalleled opportunities for understanding individual differences and personalizing services at scale, they also pose significant ethical challenges in terms of user privacy and self-determination. In demonstrating the potential of LLMs for large-scale psychological assessments based on social media data, our study can inform broader discussions about AI governance and regulatory frameworks to safeguard individual rights and prevent abuse.

Introduction

Large language models (LLMs) and other transformer-based neural networks have revolutionized text analysis in research and practice. Models such as OpenAI's GPT-4 (1) or Anthropic's Claude (2), for example, have shown a remarkable ability to represent, comprehend, and generate human-like text. Compared to prior Natural Language Processing (NLP) approaches, one of the most striking advances of LLMs is their ability to generalize their "knowledge" to novel scenarios, contexts, and tasks (3, 4).

While LLMs were not explicitly designed to capture or mimic elements of human cognition and psychology, recent research suggests that—given their training on extensive corpora of human-generated language—they might have spontaneously developed the capacity to do so. For example, LLMs display properties that are similar to the cognitive abilities and processes observed in humans, including theory of mind (i.e. the ability to understand the mental states of other agents (5)), cognitive biases in decision-making (6), and semantic priming (7). Similarly, LLMs are able to effectively generate persuasive messages tailored to specific psychological dispositions (e.g. personality traits, moral values (8)).

Here, we examine whether LLMs possess another quality that is fundamentally human: The ability to "read" people and form first impressions about their psychological dispositions in the absence of direct or prior interaction. As research under the umbrella of zero-acquaintance studies shows, people can be remarkably accurate at judging the psychological traits of strangers simply by observing traces of their behavior under certain conditions (9). While such judgments can be influenced by stereotypes and their accuracy can vary based on the traits being assessed and the context in which judgments are made (10), past work indicates that people are able to predict a stranger's personality traits by observing their offices or bedrooms (11), examining their music preferences (12), or scrolling through their social media profiles (13).

Competing Interest: The authors declare no competing interest.
Received: November 9, 2023. **Accepted:** May 28, 2024

Existing research in computational social science shows that supervised machine learning models are able to make similar predictions. That is, given a large enough dataset including both self-reported personality traits and people's digital footprints—such as Facebook Likes, music playlists, or browsing histories—machine learning models are able to statistically relate both inputs in a way that allows them to predict personality traits after observing a person's digital footprints (14, 15). This is also true for various forms of text data, including social media posts (16, 17), personal blogs (18), or short text responses collected in the context of job applications (19).

In this article, we test whether LLMs have the ability to make similar psychological inferences without having been explicitly trained to do so (known as zero-shot learning (3)). Specifically, we use Open AI's ChatGPT (GPT-3.5 and GPT-4 (1)) to explore whether LLMs can accurately infer the Big Five personality traits Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (20) of social media users from the content of their Facebook status updates in a zero-shot scenario. In addition, we test for biases in ChatGPT's judgments that might arise from its foundation in equally biased human-generated data. Building on previous work highlighting inherent stereotypes in pretrained NLP models (21, 22), we explore the extent to which the personality inferences made by ChatGPT are indicative of gender and age-related biases (e.g. potential biases in how the personality of men and women or older and younger people is judged).

Understanding the capabilities and limitations of LLMs with regard to inferring highly intimate psychological traits from digital footprints is critical, given their rapid adoption in both research and practice. On the one hand, easy access to the psychological profiles of individuals creates unprecedented opportunities to study individual differences at scale and customize products, services, or behavioral interventions to individuals' unique dispositions. On the other hand, however, automated psychological inferences pose considerable ethical and legal challenges with regard to individuals' privacy and self-determination (23). This problem is exacerbated by the fact that LLMs are also able to automatically craft persuasive messages based on users' personality profiles (8). The combination of fully automated psychological assessments and personalized interactions opens the door for manipulation and misuse at scale and with little to no human oversight. We discuss our findings in light of both the opportunities for scientists and practitioners and the challenges that will require new forms of AI governance and regulation (24–26).

Methods

Data and sampling

Our analyses are based on text data obtained from MyPersonality (27), a Facebook application that allowed users to take real psychometric tests—including a validated measure of the Big Five personality traits (IPIP (28))—and receive immediate feedback on their responses. Users also had the opportunity to donate their Facebook profile information—including their public profiles, Facebook Likes, and status updates—to research. For the purpose of this study, we randomly subsampled 1,000 adult users (24.2 ± 8.8 years old, 63.1% female) who completed the full 100-item IPIP personality questionnaire and had at least 200 Facebook status updates (if they had more, we used the most recent 200). The study received IRB approval from Columbia University's ethics review board (Protocol #AAAU8559).

Measures

MyPersonality measured users' personality traits using the International Personality Item Pool (IPIP (28)), a widely established self-report questionnaire that captures the Big Five personality traits of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (20). We only included users who had completed the full questionnaire with all 100 items.

To obtain inferred personality traits from ChatGPT, we used the last 200 Facebook status updates generated by each user without additional preprocessing. The average length of status updates in our sample was 17.10 words (SD = 15.03). Status updates were scored using the ChatGPT API with GPT-3.5 (version gpt-3.5-turbo-0301) and GPT-4 (version gpt-4-0314) (1) as underlying models. For this purpose, the status updates were first concatenated into chunks and then fed into the GPT model, using a set of simple prompts to guide the behavior of the model. The system prompt was the default for GPT-3.5 and GPT-4, respectively: "You are a helpful assistant". Additionally, we prompted the model to infer Big Five traits using the inference prompt: "Rate the text on the Big Five personality dimensions. Pay attention to how people's personalities might be reflected in the content they post online. Provide your response on a scale from 1 to 5 for the traits Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Provide only the numbers." We then used a simple text-parsing script to transform the responses into numerical scores. In order to avoid exceeding the ChatGPT token limit, status update histories were processed in chunks of 20 messages, and the inferred personality scores were then averaged to derive overall scores.

To boost the reliability of the inferred personality estimates, we queried ChatGPT three times for each inference. Agreement across ratings across rating rounds was high for all traits (Openness: $r_{GPT3.5} = 0.88$, $r_{GPT4} = 0.73$; Conscientiousness: $r_{GPT3.5} = 0.88$, $r_{GPT4} = 0.91$; Extraversion: $r_{GPT3.5} = 0.92$, $r_{GPT4} = 0.87$; Agreeableness: $r_{GPT3.5} = 0.96$, $r_{GPT4} = 0.94$; Neuroticism: $r_{GPT3.5} = 0.91$, $r_{GPT4} = 0.93$), and all P-values were smaller than 0.001 with Bonferroni correction for multiple comparisons. Given the high level of agreement, we computed aggregate inferred scores by averaging scores across the three rounds of rating. We used the aggregate scores for all further analyses.

Results

Can LLMs infer personality traits from social media posts?

In order to assess the capacity of LLMs to infer psychological traits from social media data, we compared the inferred Big Five personality scores with self-reported scores. A comparison of the distributions suggests that both versions of ChatGPT tended to underestimate Conscientiousness and Agreeableness while overestimating Neuroticism. For Openness and Extraversion, the deviations were inconsistent across ChatGPT versions: While GPT-3.5 tended to underestimate Openness and Extraversion, GPT-4 tended to overestimate Extraversion. Overall, the distributions of inferred scores were more closely aligned with self-reported scores for GPT-4 compared to GPT-3.5, suggesting a potential improvement across versions (see Fig. 1). Detailed descriptive statistics can be found in Supplementary Material S1.

Importantly, the mere comparison of distributions does not provide insights into the strength and directionality of the relationships between inferred and self-reported scores. For this purpose, we conducted correlation analyses. The average Pearson

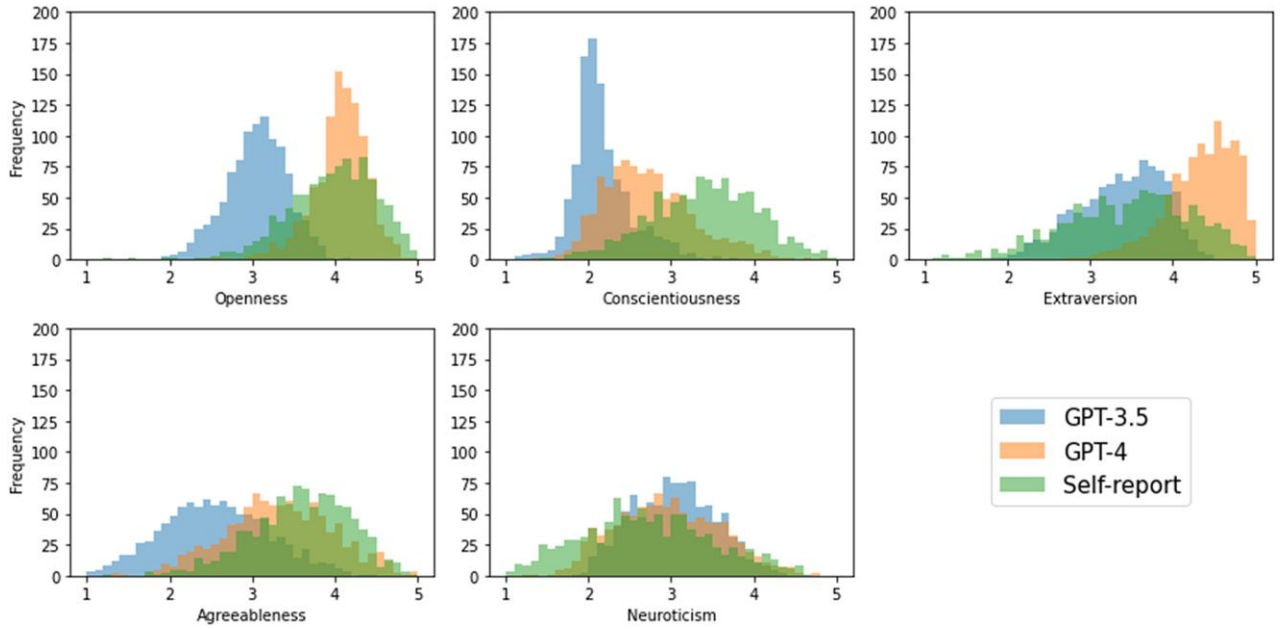

Fig. 1. Distributions of self-reported and inferred personality scores for GPT-3.5 and GPT-4. Histograms show absolute frequencies for an overall sample size of $n = 1{,}000$. GPT-3.5 underestimates Openness. Both models underestimate Conscientiousness and Agreeableness but overestimate Neuroticism. For Extraversion, the two models diverge with GPT-3.5 underestimating and GPT-4 overestimating the true scores. Overall, GPT-4 inferred scores were more aligned with self-reported scores, indicating a potential improvement over GPT-3.5.

correlation coefficient of inferred and self-reported scores across all personality traits was $r_{GPT3.5} = 0.27$ and $r_{GPT4} = 0.31$. The correlations were highest for the traits of Openness ($r_{GPT3.5} = 0.28$; $r_{GPT4} = 0.33$), Extraversion ($r_{GPT3.5} = 0.29$; $r_{GPT4} = 0.32$) and and Agreeableness ($r_{GPT3.5} = 0.30$, $r_{GPT4} = 0.32$), and were slightly lower for Conscientiousness ($r_{GPT3.5} = 0.22$; $r_{GPT4} = 0.26$) and Neuroticism ($r_{GPT3.5} = 0.26$; $r_{GPT4} = 0.29$). All correlation coefficients were significantly different from 0 at $P < 0.001$ with Bonferroni correction for multiple comparisons. Similar to the comparison of distributions, GPT-4 showed higher levels of accuracy across all five personality traits, although none of the individual comparisons reached statistical significance (see Fig. 2). Detailed results, including confidence intervals and significance levels, can be found in Supplementary Material S2.

In addition to exploring the capacity of ChatGPT to infer personality traits from social media user data, we also tested the extent to which this capacity is sensitive to changes in the amount of data that was available for inference. Specifically, we computed correlations between self-reported and inferred personality scores based on different numbers of status messages. Specifically, we computed correlations obtained from inferences for a single chunk of status messages (20 status messages) all the way up to ten chunks (200 status messages). As expected, having access to more status messages resulted in more accurate inferences. Notably, however, most correlations are close to their maximum level after observing far less than the ultimate number of 200 status messages. In addition, the inference of certain traits seems to be particularly susceptible to the volume of input data. For example, the models' accuracy kept increasing with higher levels in input volume for Openness, Extraversion, Agreeableness, and Neuroticism, while the benefits of additional status messages leveled off earlier for Conscientiousness. See Fig. 2 for a graphical representation and Supplementary Material S3 for detailed statistics.

Does the quality of LLM inferences vary across demographic groups?

In order to uncover potential gender and age-related biases, we analyzed group differences in inferred Big Five scores, as well as their residuals with respect to self-reported scores. Notably, such gender and age differences might not only emerge in inferred personality scores but are also known to exist in self-reports (29, 30). Consequently, we test for both overall group differences and differences in the residuals between the self-reported and inferred personality scores of each individual.

Gender differences

We first explored the extent to which any observed group differences in inferred personality traits across men and women aligned with those observed in self-reports. As Fig. 3 shows, women tend to score significantly higher in Agreeableness ($t = 2.31$; $P = 0.021$) and Neuroticism ($t = 6.53$; $P < 0.001$) when these traits are measured using questionnaires. In contrast, women scored significantly higher in Openness ($t_{GPT3.5} = 3.42$, $p_{GPT3.5} < 0.001$; $t_{GPT4} = 2.72$, $p_{GPT4} = 0.007$), Conscientiousness ($t_{GPT3.5} = 5.28$, $p_{GPT3.5} < 0.001$; $t_{GPT4} = 5.73$, $p_{GPT4} < 0.001$), Extraversion ($t_{GPT3.5} = 5.21$, $p_{GPT3.5} < 0.001$; $t_{GPT4} = 7.25$, $p_{GPT4} < 0.001$), and Agreeableness ($t_{GPT3.5} = 13.53$, $p_{GPT3.5} < 0.001$; $t_{GPT4} = 13.63$, $p_{GPT4} < 0.001$) when these traits were inferred by ChatGPT models, with no significant differences found for Neuroticism. This finding offers initial evidence for potential gender biases in the personality inferences made by LLMs (see Fig. 3).

To further explore these potential biases, we analyzed the residuals between inferred scores and self-reported scores as an indication of how well GPT is able to represent the personality traits of male and female users. The findings suggest that GPT's personality inferences are less accurate for men than women. First, we observed larger absolute residuals for male users in Conscientiousness

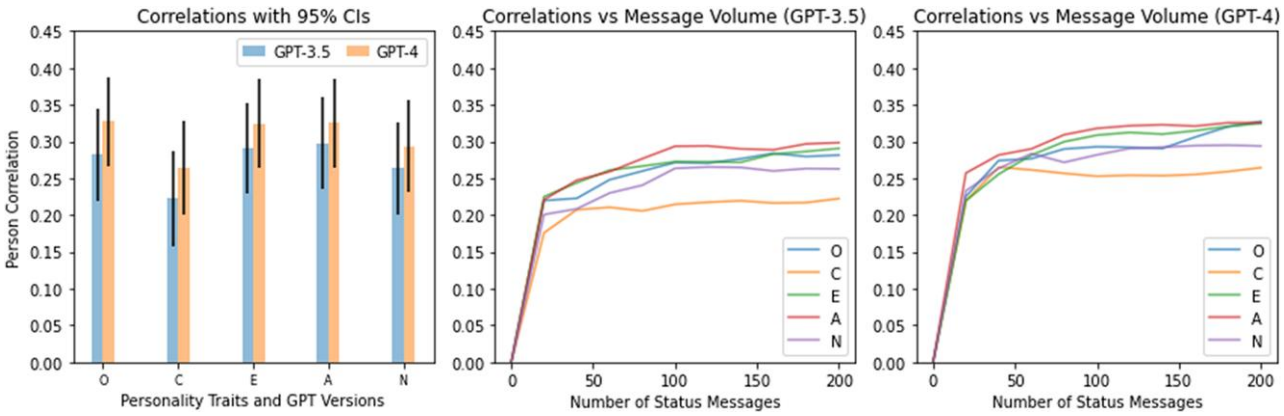

Fig. 2. Pearson's correlation coefficients between inferred and self-reported scores with 95% confidence intervals (left), and Pearson's correlation coefficients for GPT-3.5 (mid) and GPT-4 as a function of message volume (right). O, Openness; C, Conscientiousness; E, Extraversion; A, Agreeableness; N, Neuroticism. Inferences for Openness, Extraversion, and Agreeableness were more accurate than those for Conscientiousness and Neuroticism, but the differences remained nonsignificant. Higher message volume was associated with higher levels of predictive accuracy, but a substantial share of variance was captured in as little as 20 status messages.

($t_{GPT3.5} = -3.53$, $p_{GPT3.5} < 0.001$; $t_{GPT4} = -4.48$, $p_{GPT4} < 0.001$), Agreeableness ($t_{GPT3.5} = -9.22$, $p_{GPT3.5} < 0.001$; $t_{GPT4} = -5.22$, $p_{GPT4} < 0.001$), and Neuroticism ($t_{GPT3.5} = -4.55$, $p_{GPT3.5} < 0.001$; $t_{GPT4} = -2.39$, $p_{GPT4} = 0.017$) across both GPT models, indicating lower accuracy on these traits for men. Additionally, we found larger residuals for male users for GPT-3.5 in Openness ($t_{GPT3.5} = -3.84$, $p_{GPT3.5} < 0.001$; $t_{GPT4} = -0.92$, $p_{GPT4} = 0.357$) and larger residuals for female users in Extraversion for GPT-4 ($t_{GPT3.5} = -1.36$, $p_{GPT3.5} < 0.173$; $t_{GPT4} = 3.12$, $p_{GPT4} = 0.002$). For a visual representation, please refer to Fig. 4). Detailed statistics can be found in Supplementary Material S4.

Taken together, the findings suggest that ChatGPT's personality inferences are less accurate for men than women. Notably, however, these biases seem to be limited to the absolute measures of accuracy and do not necessarily translate to ChatGPT's ability to make inferences about men's relative personality levels. That is, when computing Pearson correlations within gender groups, we did not observe any significant difference in the magnitude of these correlations. Similarly, controlling for gender in the overall correlations between self-reported and inferred personality scores by z-standardizing inferred scores within each gender group did not yield correlations significantly different from those obtained before.

Age differences

As for gender, we first explored the extent to which any observed group differences in inferred personality traits across younger and older adults (classified using a median split) were aligned with those observed in self-reports. As Fig. 3 shows, older users displayed significantly higher self-reported scores in Openness ($t = 2.96$; $P = 0.003$) and Conscientiousness ($t = 7.27$; $P < 0.001$) and significantly lower self-reported scores in Neuroticism ($t = -3.28$; $P = 0.001$) compared to younger users. Partially mimicking these differences in self-reported personality traits, inferred scores were significantly higher in Conscientiousness ($t_{GPT3.5} = 9.23$, $p_{GPT3.5} < 0.001$; $t_{GPT4} = 10.41$, $p_{GPT4} < 0.001$) and Agreeableness ($t_{GPT3.5} = 4.87$, $p_{GPT3.5} < 0.001$; $t_{GPT4} = 4.39$, $p_{GPT4} < 0.001$), and lower in Neuroticism ($t_{GPT3.5} = -3.43$, $p_{GPT3.5} < 0.001$; $t_{GPT4} = -4.37$, $p_{GPT4} < 0.001$) for older compared to younger users. For Openness ($t_{GPT3.5} = -2.86$, $p_{GPT3.5} = 0.004$; $t_{GPT4} = -0.72$, $p_{GPT4} = 0.472$), and Extraversion ($t_{GPT3.5} = -3.55$, $P < 0.001$; $t_{GPT4} = 0.36$, $p_{GPT4} = 0.717$), older individuals scored significantly lower on inferred scores for GPT-3.5 but not GPT-4 (see Fig. 3).

As before, we further explore these differences by analyzing age differences in the residuals between self-reported and inferred scores. Unlike in the analyses of gender, we found substantial inconsistency in the group differences between GPT-3.5 and GPT-4. While the inferences made by GPT-3.5 showed significantly larger absolute residuals for older users in Openness ($t_{GPT3.5} = 4.78$, $p_{GPT3.5} < 0.001$), Conscientiousness ($t_{GPT3.5} = 2.64$, $p_{GPT3.5} = 0.008$), and smaller residuals for Agreeableness ($t_{GPT3.5} = -2.64$, $p_{GPT3.5} = 0.009$), no differences in absolute residuals were found for GPT-4. For a visual representation, please refer to Fig. 4) Detailed statistics can be found in Supplementary Material S5.

Taken together, the findings suggest that ChatGPT's personality inferences might be less accurate for older adults. However, as before, these biases did not translate to ChatGPT's ability to make inferences about people's relative personality levels. We did not find significant differences between within-group correlation coefficients, and z-standardizing personality scores within age groups did not yield correlation coefficients significantly different from those reported before.

Agreement with third-person observer ratings

We conducted a preliminary analysis examining the correlations between self-reported personality scores and third-person observer ratings, as well as between LLM-inferred scores and third-person observer ratings. This allowed us to (i) compare the quality of LLM inferences against a strong human benchmark (i.e. ratings from people who have access to more identity cues than just social media profiles) and (ii) examine the level of agreement between LLM inferences and human judgments. Third-person ratings were collected by letting users' Facebook friends complete a 10-item version of the IPIP personality questionnaire (28, 31) about them. The analysis includes a subset of 68 individuals for whom third-person ratings were available.

The results show that correlations between self-reported scores and observer ratings ranged from $r = 0.198$ to $r = 0.378$ (mean = 0.304), while the correlations between ChatGPT-inferred scores and observer ratings ranged from $r = 0.057$ to $r = 0.457$ (mean = 0.269) for GPT-3.5 and $r = 0.152$ to $r = 0.400$ (mean = 0.276) for GPT-4 (see Supplementary Material S6

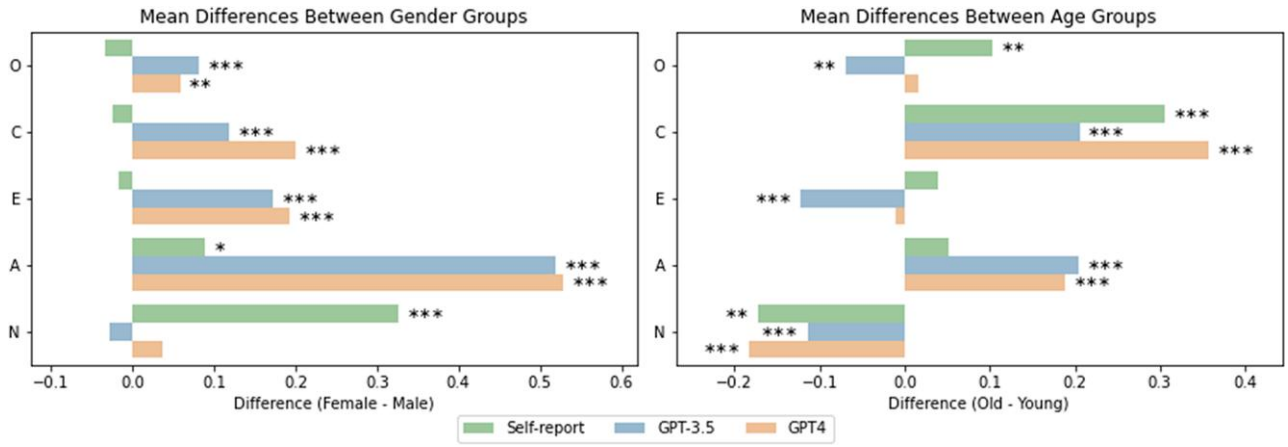

Fig. 3. Mean differences in personality scores between gender groups (left) and age groups (right) for self-reported scores as well as inferences by GPT-3.5 and GPT-4. Positive values indicate higher scores for female users compared to male users and older users compared to younger users. O, Openness; C, Conscientiousness; E, Extraversion; A, Agreeableness; N, Neuroticism. *** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$. The results show significant gender and age differences across all personality traits.

for detailed results). Overall, the correlation coefficients were largely in the same range as those between self-reported and ChatGPT-inferred scores. The analyses thus suggest that the accuracy of ChatGPT's inferences is on par with that of human observers. They also suggest that the ChatGPT is using similar cues to human judges. This is true even though these cues may not always be valid predictors of people's self-perceptions. For instance, in the case of Conscientiousness, the agreement between ChatGPT's inferences and observer ratings was higher than the agreement of either ChatGPT or human observers with participants' self-reports.

Discussion

Interpretation of results

Our findings suggest that LLMs, such as ChatGPT, can infer psychological dispositions from people's social media posts without having been explicitly trained to do so. They also offer preliminary evidence that LLMs might generate more accurate inferences for women and younger individuals (compared to men and older adults). Notably, the overall accuracy of the observed inferences (Pearson correlations between self-reported and inferred personality traits ranging between $r = 0.22$ and 0.33, average = 0.29) is slightly lower than that accomplished by supervised models which have been trained or fine-tuned specifically for this purpose and with the same textual data source as used in testing (e.g. Park et al. (16), who reported correlations between $r = 0.26$ and $r = 0.41$, average $r = 0.37$). Yet, the ability of LLMs to produce inferences of reasonably high accuracy in zero-shot learning scenarios has both important theoretical and practical implications.

Our study contributes to a growing body of research comparing the abilities of LLMs to those observed in humans (5, 7, 8). As our findings suggest, LLMs might have the human-like ability to "profile" people based on their behavioral traces, without ever having had direct interactions with them. Although most social media posts do not contain explicit references to a person's character, ChatGPT—just like human judges (13, 32) or supervised models (31)—is able to translate people's accounts of their daily activities and preferences into a holistic picture of their psychological dispositions. Our results are aligned with previous work suggesting that Openness and Extraversion are more easily inferred than other traits (13, 31). At the same time, LLM inferences were more congruent with observer ratings than self-reports in the case of Conscientiousness, indicating that LLMs may also replicate biases in human judgment for certain traits.

Notably, the specific pathways by which LLMs such as ChatGPT arrive at their judgments and the reasons for why certain biases are introduced into the predictions (e.g. systematic gender and age differences) remain unknown. That is, we cannot speak to the question of whether LLMs use the same behavioral cues as humans or supervised machine learning models when translating behavioral residues into psychological profiles or offer an in-depth explanation for the observed differences in accuracy across age and gender categories. For example, the fact that ChatGPT shows systematic biases in its estimation of certain personality traits and is more accurate for women and younger adults could either be indicative of a bias introduced in the training of the models and/or the corpora of text data the models have been trained on, or reflective of differences in people's general self-expression on social media.

Specifically, past work indicates that LLMs are susceptible to stereotyping and bias with regard to demographic and geographic groups (21, 33–37), likely reflecting groups' representation in the underlying training data. At the same time, past work has shown differences in social media use and online self-expression across demographic groups, including age and gender (38–42). While the past literature does not directly speak to differences in personality expression, the observed pattern of results would indicate that women and younger individuals tend to reveal more accurate information about their personalities online.

Limitations and future research

Our study has several limitations that should be addressed by future research. Firstly, as mentioned above, the black box character of LLMs prevents us from examining the precise mechanisms by which personality inferences are derived. As a first step in this direction, future research should analyze cue utilization by investigating which language features are highly correlated with inferred trait scores. Similarly, it would be useful to examine which language features are predictive of inference errors in order to better understand the origins of the observed gender and age biases.

Second, the text data used in our analysis was obtained from the MyPersonality Facebook application (27), which was active between 2007 and 2012. Linguistic conventions from this period

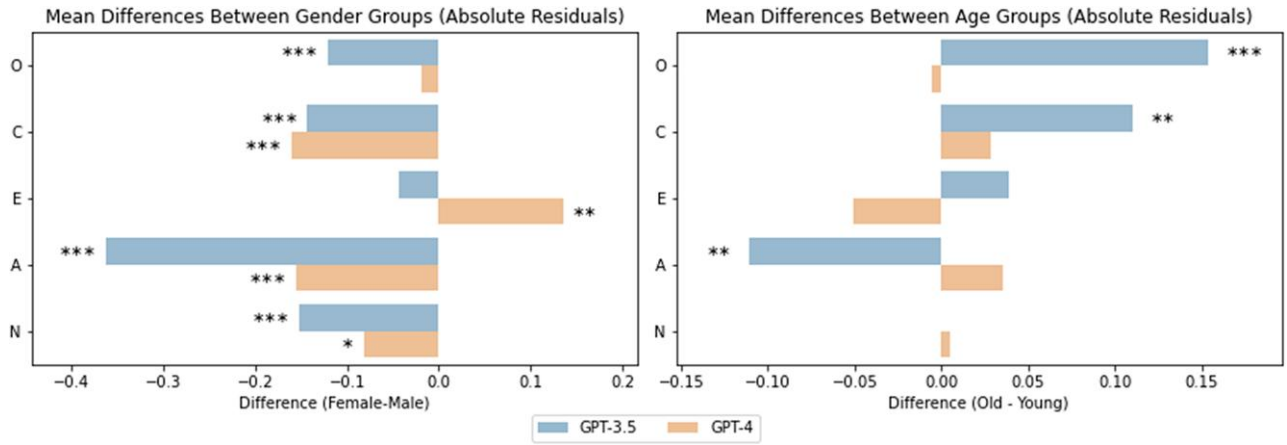

Fig. 4. Mean differences in absolute residuals between gender groups (left) and age groups (right) for inferences by GPT-3.5 and GPT-4. Positive values indicate higher residuals for female users compared to male users and older users compared to younger users. O, Openness; C, Conscientiousness; E, Extraversion; A, Agreeableness; N, Neuroticism. *** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$. The results indicate lower residuals for female users in all personality traits except Extraversion. Age-related biases were observed for Openness, Conscientiousness, and Agreeableness in inferences by GPT-3.5 but not GPT-4.

might differ from contemporary online language, potentially limiting the zero-shot performance of LLMs, which have been trained on newer data. As a result, we would expect the personality inferences of LLMs to be even more accurate when applied to more contemporary data.

Third, our data were sourced from Facebook users who interacted with the MyPersonality application. As such, our sample might not be representative of the broader population of social media users (or people more generally), which could limit the external validity of our findings. For example, the general underestimation of personality traits such as Openness might be due to the fact that myPersonality users were particularly curious and open-minded.

Fourth, while our study probed how sensitive the accuracy of LLM-based inferences is to the volume of text input, we limited our data to the 200 most recent status updates. In practice, predictive performance might vary for users with fewer or more status updates. Relatedly, due to the inherent token limit in models like ChatGPT, all input data were processed in chunks. It is possible that the accuracy of future models with the ability to process larger amounts of input data at once might be higher.

Fifth, our study did not encompass the dynamics of live interactions between LLMs and users. Real-time interactions might yield different insights and highlight additional complexities not captured in our static textual data set (43). Relatedly, while our research underscores the potential for LLMs in personalizing interactions and enhancing social computing, it does not examine the specifics of how these personalizations can be effectively implemented.

Sixth, the current research demonstrates the potential of out-of-the-box LLMs for inferring psychological variables using simple techniques such as zero-shot learning and commercially available models. It is likely that the predictive performance of LLMs could be improved through more sophisticated prompting strategies, such as chain-of-thought prompting (44) and a combination of in-context learning and supervised fine-tuning (45). While we purposefully focused on zero-shot learning in order to establish a lower bound of predictive accuracy and investigate LLMs' inherent ability to make such predictions, future research could focus on identifying levers that maximize predictive accuracy. Aside from more sophisticated prompting paradigms, this could include giving LLMs access to users' demographic information which is typically available to human perception and could moderate the interpretation of personality-related signals. For example, the content of status messages may be interpreted differently depending on whether the user is an 18-year-old man or a 55-year-old woman. Being able to interpret message content in the context of sender identity could lead to improved inferences but could also amplify implicit biases that are known to persist in language models (21, 33, 37).

Finally, while we make an effort to discuss the societal implications of our findings (see below), detailed recommendations regarding privacy concerns and the potential for misuse should be addressed in future research.

Implications

Our findings also have important practical implications for the application of automated psychological profiling in research and industry. Specifically, the ability of LLMs to infer psychological traits from social media data could foreshadow a remarkable shift in the accessibility—and therefore potential use—of scalable psychometric assessments. For decades, the assessment of psychological traits relied on the use of self-report questionnaires, which are known to be prone to self-report biases and difficult to scale due to their costly and time-consuming nature (46). With the introduction of automated psychological assessments driven by supervised machine learning models (15, 19), scientists and practitioners were afforded an alternative approach that promised to expand the study and application of individual differences to research questions and domains that were previously impractical if not impossible (e.g. the use of personality traits in targeted advertising (47); or the investigation of individual differences in large scale, ecologically valid observational studies (48)). However, the widespread application of such automated personality predictions from digital footprints among scientists and practitioners was hindered by the need to collect large amounts of self-report surveys in combination with textual data (see e.g. the myPersonality dataset (27)) to train and validate the predictive models. With the ability to make similar inferences with models that are available to the broader public, LLMs could democratize access to cheap and scalable psychometric assessments.

While this democratization holds remarkable opportunities for scientific discovery and personalized services, it also introduces considerable ethical challenges. Specifically, the ability to predict people's intimate psychological needs and preferences without their knowledge or consent poses a threat to people's privacy and self-determination (23). For instance, users often share information online without considering how this information can be used by third parties and the use of LLMs for psychological profiling may not align with their original intentions. As the case of Cambridge Analytica (49) alongside a growing body of research on personalized persuasion and psychological targeting (47, 50, 51) has highlighted, insights into people's psychological dispositions can easily be weaponized to sway opinions and change behavior. Consequently, it might be necessary to introduce guardrails into systems like LLMs that prevent actors from obtaining psychological profiles of thousands or millions of users. Notably, the outlined concerns are aligned with recent calls for regulation (24–26) and the fact that the EU AI Act (52) explicitly bans emotion recognition in the workplace and educational institutions, as well as social scoring based on social behavior or personal characteristics.

Conclusion

Taken together, our research demonstrates the capacity of LLMs to derive psychological profiles from social media data, even without specific training. This zero-shot capability underscores the remarkable advancement LLMs represent in the domain of text analysis. While this "intuitive" understanding mirrors distinctly human abilities, the mechanisms and inherent biases associated with LLM-based personality judgments remain elusive and warrant further research. From a practical perspective, the potential of LLMs to effectively infer psychological traits from digital footprints presents a shift in psychometric evaluations, paving the way for large-scale AI-driven assessments. The prospect of democratized, scalable psychometric tools will enable breakthroughs in personalized services and large-scale research. Nevertheless, these advancements bring forth ethical challenges. The potential for nonconsensual psychological predictions and other misuses highlights the necessity for stringent ethical frameworks.

Acknowledgments

We thank the Digital Future Initiative and Columbia Business School for their generous support. We thank Michal Kosinski for fruitful conversations and advice.

This manuscript was posted as a preprint (https://doi.org/10.48550/arXiv.2309.08631).

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Author Contributions

H.P.: conceptualization, methodology, software, formal analysis, investigation, writing - original draft, visualization; S.C.M.: conceptualization, methodology, writing - original draft, visualization.

Data Availability

The data and code needed to reproduce the results, along with other supplementary materials, have been made available on the project's OSF page (53).

References

1 OpenAI. 2023. GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf.
2 Anthropic. 2023. Model card and evaluations for claude models. https://www-cdn.anthropic.com/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226/Model-Card-Claude-2.pdf.
3 Brown T, *et al.* 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*. 33:1877–1901.
4 Radford A, *et al.* 2019. Language models are unsupervised multitask learners. https://paperswithcode.com/paper/language-models-are-unsupervised-multitask.
5 Kosinski M. 2023. Theory of mind might have spontaneously emerged in large language models, arXiv, arXiv:2302.02083, preprint: not peer reviewed.
6 Hagendorff T, Fabi S, Kosinski M. 2023. Thinking fast and slow in large language models, arXiv, arXiv:2212.05206, preprint: not peer reviewed.
7 Digutsch J, Kosinski M. 2023. Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. *Sci Rep*. 13(1):5035.
8 Matz SC, *et al.* 2024. The potential of generative AI for personalized persuasion at scale. *Sci Rep*. 14(1):4692.
9 Albright L, Kenny DA, Malloy TE. 1988. Consensus in personality judgments at zero acquaintance. *J Pers Soc Psychol*. 55(3):387–395.
10 Kenny DA, Albright L, Malloy TE, Kashy DA. 1994. Consensus in interpersonal perception: acquaintance and the big five. *Psychol Bull*. 116(2):245–258.
11 Gosling SD, Ko SJ, Mannarelli T, Morris ME. 2002. A room with a cue: personality judgments based on offices and bedrooms. *J Pers Soc Psychol*. 82(3):379–398.
12 Rentfrow PJ, Gosling SD. 2006. Message in a ballad: the role of music preferences in interpersonal perception. *Psychol Sci*. 17(3):236–242.
13 Back MD, *et al.* 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychol Sci*. 21(3):372–374.
14 Azucar D, Marengo D, Settanni M. 2018. Predicting the Big 5 personality traits from digital footprints on social media: a meta-analysis. *Pers Individ Dif*. 124:150–159.
15 Kosinski M, Stillwell D, Graepel T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proc Natl Acad Sci USA*. 110(15):5802–5805.
16 Park G, *et al.* 2015. Automatic personality assessment through social media language. *J Pers Soc Psychol*. 108(6):934–952.
17 Schwartz HA, *et al.* 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One*. 8(9):e73791.
18 Yarkoni T. 2010. Personality in 100,000 words: a large-scale analysis of personality and word use among bloggers. *J Res Pers*. 44(3):363–373.
19 Grunenberg E, Peters H, Francis MJ, Back MD, Matz SC. 2024. Machine learning in recruiting: predicting personality from CVs and short text responses. *Front Soc Psychol*. 1:1–14.
20 McCrae RR, Costa Jr PT. 2008. The five-factor theory of personality. In: *Handbook of personality: theory and research*. 3rd ed. New York (NY): The Guilford Press. p. 159–181.
21 Bolukbasi T, Chang K-W, Zou J, Saligrama V, Kalai A. 2016. *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates. p. 4356–4364.
22 Wan Y, *et al.* 2023. BiasAsker: measuring the bias in conversational AI system, arXiv, arXiv:2305.12434, preprint: not peer reviewed.

23 Matz SC, Appel RE, Kosinski M. 2020. Privacy in the age of psychological targeting. *Curr Opin Psychol*. 31:116–121.
24 Chan A. 2023. GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. *AI Ethics*. 3(1):53–64.
25 Hacker P, Engel A, Mauer M. 2023. Regulating ChatGPT and other large generative AI models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY: ACM. p. 1112–1123.
26 Perc M, Ozer M, Hojnik J. 2019. Social and juristic challenges of artificial intelligence. *Palgrave Commun*. 5(1):61.
27 Kosinski M, Matz SC, Gosling SD, Popov V, Stillwell D. 2015. Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. *Am Psychol*. 70(6):543–556.
28 Goldberg LR, *et al.* 2006. The international personality item pool and the future of public-domain personality measures. *J Res Pers*. 40(1):84–96.
29 Costa Jr PT, Terracciano A, McCrae RR. 2001. Gender differences in personality traits across cultures: robust and surprising findings. *J Pers Soc Psychol*. 81(2):322–331.
30 Feingold A. 1994. Gender differences in personality: a meta-analysis. *Psychol Bull*. 116(3):429–456.
31 Youyou W, Kosinski M, Stillwell D. 2015. Computer-based personality judgments are more accurate than those made by humans. *Proc Natl Acad Sci USA*. 112(4):1036–1040.
32 Vazire S, Gosling SD. 2004. e-Perceptions: personality impressions based on personal websites. *J Pers Soc Psychol*. 87(1):123–132.
33 Abdurahman S, *et al.* 2023. Perils and opportunities in using large language models in psychological research. https://osf.io/tg79n.
34 Atari M, Xue MJ, Park PS, Blasi D, Henrich J. 2023. Which humans? https://osf.io/5b26t.
35 Durmus E, *et al.* 2024. Towards measuring the representation of subjective global opinions in language models, arXiv, arXiv:2306.16388 [cs], preprint: not peer reviewed.
36 Rathje S, *et al.* 2023. GPT is an effective tool for multilingual psychological text analysis. https://osf.io/sekf5.
37 Santurkar S, *et al.* 2023. Whose opinions do language models reflect? *Proceedings of the 40th International Conference on Machine Learning*. Cambridge, MA: PMLR. p. 29971–30004.
38 Kondakciu K, Souto M, Zayer LT. 2021. Self-presentation and gender on social media: an exploration of the expression of "authentic selves". *Qual Mark Res Int J*. 25(1):80–99.
39 Oberst U, Renau V, Chamarro A, Carbonell X. 2016. Gender stereotypes in Facebook profiles: are women more female online?. *Comput Hum Behav*. 60:559–564.
40 Roberti G. 2022. Female influencers: analyzing the social media representation of female subjectivity in Italy. *Front Soc*. 7:1–11.
41 Thayer SE, Ray S. 2006. Online communication preferences across age, gender, and duration of internet use. *CyberPsychology Behav*. 9(4):432–440.
42 Tifferet S, Vilnai-Yavetz I. 2014. Gender differences in Facebook self-presentation: an international randomized study. *Comput Human Behav*. 35:388–399.
43 Peters H, Cerf M, Matz S. 2024. Large language models can infer personality from free-form user interactions. Publisher: OSF Preprints.
44 Yang T, *et al.* 2023. PsyCoT: psychological questionnaire as powerful chain-of-thought for personality detection. *Findings of the Association for Computational Linguistics: EMNLP 2023*. Stroudsburg, Pennsylvania: ACL. p. 3305–3320.
45 Karra SR, Nguyen ST, Tulabandhula T. 2023. Estimating the personality of white-box language models, arXiv, arXiv:2204.12000 [cs], preprint: not peer reviewed.
46 Podsakoff PM, MacKenzie SB, Podsakoff NP. 2012. Sources of method bias in social science research and recommendations on how to control it. *Annu Rev Psychol*. 63(1):539–569.
47 Matz SC, Kosinski M, Nave G, Stillwell DJ. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proc Natl Acad Sci USA*. 114(48):12714–12719.
48 Freiberg B, Matz SC. 2023. Founder personality and entrepreneurial outcomes: a large-scale field study of technology startups. *Proc Natl Acad Sci USA*. 120(19):e2215829120.
49 Hu M. 2020. Cambridge Analytica's black box. *Big Data Soc*. 7(2):1–6.
50 Feinberg M, Willer R. 2019. Moral reframing: a technique for effective and persuasive communication across political divides. *Soc Personal Psychol Compass*. 13(12):e12501.
51 Teeny JD, Siev JJ, Briñol P, Petty RE. 2021. A review and conceptual framework for understanding personalized matching effects in persuasion. *J Consum Psychol*. 31(2):382–414.
52 European Parliament. 2023. Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI. https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699.
53 Peters H. 2023. [dataset]* Large language models can infer psychological dispositions of social media users. https://osf.io/hq2ra/.