Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA37074

Analysis of Prediction Accuracy of Different Chatbots for Big Five Personality Traits in High School Students

Shengxiang Qiao

Nanjing Foreign Language School British Colombia Academic, Nanjing, China
Simon.Qiao27@nflsbca.com

Abstract. Chatbots based on large language models (LLMs) are commonly used by teenagers for daily communication and introspection. All the previous studies merely compare personality outputs created by a certain AI with no numeric assessment of differences from the self-reported personality scores of students. Questionnaire data are obtained from 33 Chinese students who study in Canada-based education programs. BFI-10 ten-question short inventory is selected for measuring self-reported personality trait scores of the participants which are compared to personality prediction scores by five major chatbots. Participants provide subjective evaluation regarding the ability of chatbots to understand their personality traits. Data cleaning, calculating errors and visualization of the results are conducted with Pandas, Numpy and Matplotlib in Python. Two research hypotheses are proposed before conducting data analysis. It is found that the AI-based personality trait scores are still relatively correlated with the results of self-assessment of participants. There are some discrepancies in conscientiousness and openness but the highest differences occur in case of neuroticism. Moderate negative correlation is found between subjective evaluation of AI understanding and prediction error. This research provides practical data for further studies of LLM personality prediction in adolescents and proposes certain directions for further investigation.

Keywords: large language model, chatbot, Big Five personality, personality prediction, adolescent self-evaluation

1. Introduction

LLM Chatbots are AI models built on the basis of huge datasets of public texts that allow conversational AI [1]. This technology enables AI to understand natural conversations and perform personality evaluation according to Big Five Personality Model [2]. Many young people use AI chatbots every day to receive help with doing homework, for regular conversation, and even for self-reflection. The Big Five Personality Model, that defines a person's stable dispositions in terms of openness, conscientiousness, extraversion, agreeableness, and neuroticism, is currently the most credible framework for personality research in academic literature [2].

The literature review reveals that previous academic papers recognize the potential of LLMs to extract personality-related information from ordinary conversation [3]. Nevertheless, there are not enough studies comparing AI-generated personality scores numerically against the self-reported

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA37074

traits of students. Additionally, there is no research that compares simultaneously the accuracy of various chatbots regarding their ability to evaluate the personality of users [4]. Besides, there is a lack of quantifiable evidence for or against the subjective opinion of teenagers whether AI understands them better than shown by differences in prediction scores [5].

Currently, the literature can be divided into two completely different directions. The first one is aimed at improvement of the algorithm of personality recognition and validation of the possibility of automated scoring [6]. Such studies usually present only one product of AI and have a small sample size, which hinders the fair comparison of different chatbots on an equal basis. The second direction explores the attitudes of users to conversational AI [7], and does not correlate user experience data with the actual objective results of personality survey deviation.

The following research project aims to fill three main gaps in previous literature: the first one is to compare the consistency of predictions of personalities by five different chatbots used frequently by international high school students [8]; the second one is to measure the accuracy of predictions for each big five trait with the use of 10-item Big Five Inventory questionnaire (BFI-10) [8]; the third one is to reveal the linear correlation between the subjective feelings of participants about AI's ability to understand them and the total difference in scores. Lower levels of error in scoring conscientiousness and openness but not in neuroticism.

Two unique research hypotheses are formed according to academic conventions to ensure that they are clear and distinct from one another.

H1. The evaluations made by the large language models are generally consistent with the evaluations provided by the high school students concerning all five dimensions of the Big Five personality structure.

H2. The large language models make small scoring errors in the dimensions of conscientiousness and openness, but inaccurate predictions about neuroticism.

2. Literature review

2.1. LLM personality inference research

The early studies on personality recognition through LLMs are largely concentrated on algorithm improvement and model validation [6]. Psychometric research done recently proves the capability of LLMs to analyze short personality tests and provide consistent results, although in practice, studies have not included Chinese teenager groups [1]. Experiments that involve cross-platform comparisons have been performed for adult populations; there is no research dedicated to international schoolchildren.

2.2. Application of BFI-10 short personality scale

BFI-10 is an extremely popular low-invasive test consisting of only ten items; it is well-supported for teenage subjects [8]. The easy 1-5 Likert scale format makes collecting big data easier and works well with small samples of students' personalities. There are very few studies where BFI-10 self-assessments are compared with AI-generated personality predictions and error margins are calculated.

DOI: 10.54254/2755-2721/2026.BA37074

2.3. Application of BFI-10 short personality scale

Existing studies have shown that there is a relationship between the perceived feeling of understanding from the users' side and the accuracy of the output results generated by the AI [5]. However, almost all of the existing literature on this topic documents the user's attitude towards AI without providing the actual measure of the relationship between both variables.

3. Methodology

Thirty-three surveys which have gone through all kinds of screening of data qualify to be the valid sample. The participants are all Chinese high school students who are enrolled in Canadian curriculum courses but do not include Canadian locals. The five common chatbot apps used by teenagers are chosen to be the research objects, including Doubao, ChatGPT, Qianwen, DeepSeek, and Gemini [4]. All descriptions of the characteristics of participants use objective third-person narration. The online survey contains four parts for unified standardized data collection, and all raw response data will be archived in accompanying Excel files. All personality measures are taken straight from the BFI-10 personality test, which is already verified, using a 1 to 5 Likert scale of response (1 = Strongly Disagree to 5 = Strongly Agree) [8]. Section One obtains the base information about AI use (main platform and its function daily). In Section Two, participants will fill out the ten personality questions from BFI-10. In Section Three, participants will ask their daily chatbot to evaluate their personality using the same ten questions from Section Two and provide numerical results of such evaluation by AI. In Section Four, participants will answer four attitude questions concerning AI evaluation of personality, emotional similarity, AI understanding, and AI prediction. Participants fill out questionnaires individually without any external monitoring based on the set pattern of answering: self-reporting personality first, entering the same prompts in chatbot and recording scores generated by AI, and filling out attitude-related questions lastly. All valid answers are collected in a single Excel file. Data cleansing and calculations are done using Python packages such as Pandas, NumPy, and Matplotlib [6].

3.1. Core calculation indicators

There are four indicators used in this paper, as follows:

1.Total prediction error: The sum of numerical differences between AI scores and self-reported scores for all ten BFI-10 questions [8]; the smaller the value, the greater prediction accuracy of the chatbot.

2.Comparative error indicators: The set of criteria used for assessing overall performance of the five chatbots under test [4].

3.Mean error of dimension: Accuracy criterion evaluated independently for each Big Five personality dimension.

4.Pearson's correlation coefficient: Statistical criterion for evaluating the correlation between attitude questionnaire variables and total prediction error.

3.2. Ethical compliance statement

All research conducted adheres to ethical standards for experiments with human subjects. Participation in the study is entirely voluntary, and there are no incentives offered to the participants [9]. Any data provided by the participants is strictly confidential. The participants have the freedom

DOI: 10.54254/2755-2721/2026.BA37074

to discontinue participation at any stage of the survey. All raw data used in the research are locally stored and completely deleted when the research process ends.

4. Results

Thirty-three fully filled questionnaires are validated as valid samples. The sample population consists of Chinese students taking Canadian syllabus courses. There will be no visualization on the demographic data collected and all descriptions of the demographic data are in textual format.

4.1. Cross-chatbot overall prediction accuracy

Table 1 shows the five chatbots are arranged in ascending order of the average total prediction error. DeepSeek has the lowest average total prediction error of 10.60 and is the most consistent among the five. Qianwen follows with an average error of 11.00, then Gemini 11.33, ChatGPT 13.80, and Doubao with the largest deviation of 15.38. Sample sizes vary from one platform to another, with DeepSeek (5), Qianwen (4), Gemini (6), ChatGPT (10), and Doubao (8). There is a noticeable gap between the average error values of the different AI tools [4]. The chart produced using Python code is illustrated below in Figure 1.

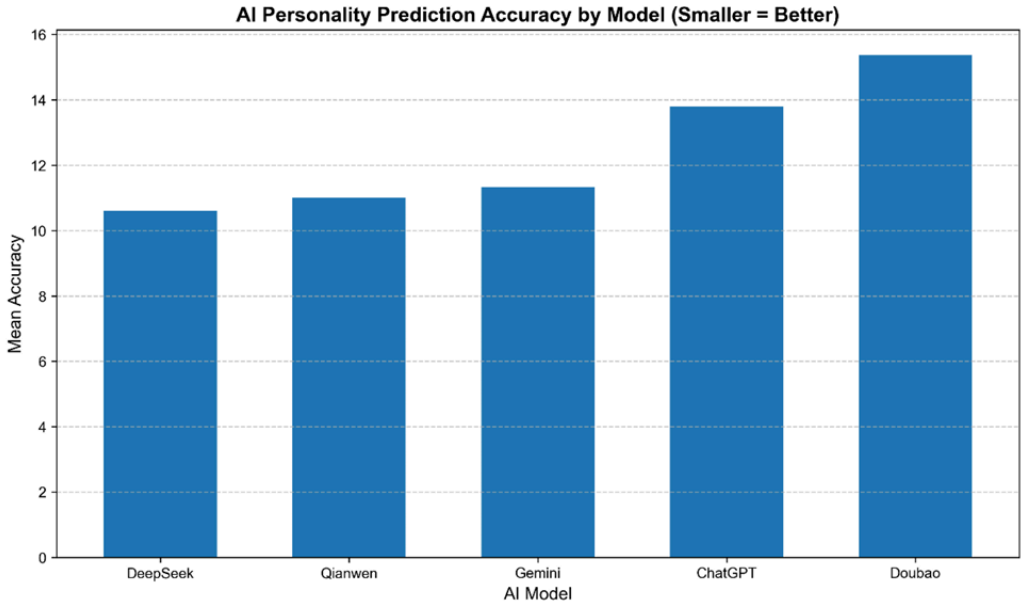


Figure 1. AI personality prediction accuracy by model

Table 1. Average total prediction error and sample size of each chatbot

Chatbot Platform	Number of Participants	Average Total Prediction Error
DeepSeek	5	10.60
Qianwen	4	11.00
Gemini	6	11.33
ChatGPT	10	13.80
Doubao	8	15.38

DOI: 10.54254/2755-2721/2026.BA37074

4.2. Dimension-wise average prediction error

The two traits with the lowest prediction error on average are conscientiousness and openness, which indicates that LLMs retain the greatest consistency in predicting those two traits, as shown in Table 2. The average error for extraversion and agreeableness is moderate, while the highest average prediction error is for neuroticism, as shown in Figure 2, which reveals significant numerical differences between AI predictions and students' self-reports [10].

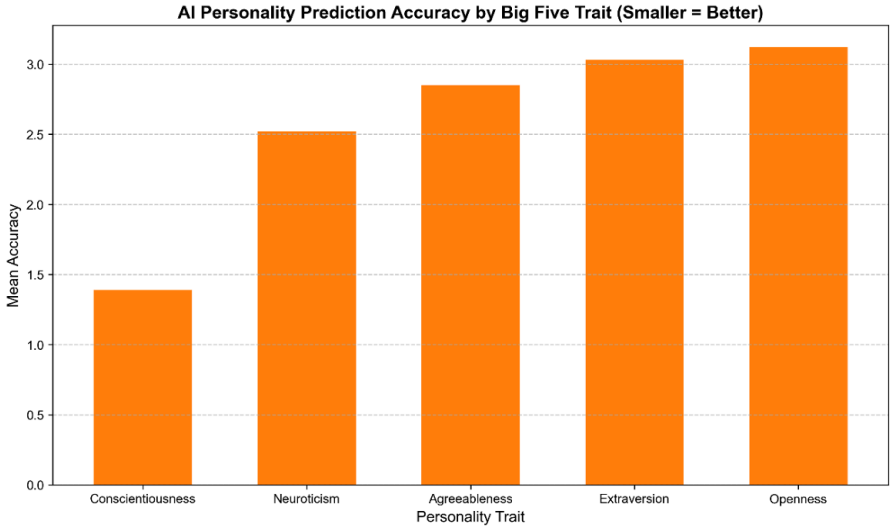


Figure 2. AI personality prediction accuracy by five traits

Table 2. Average prediction error of five Big Five personality dimensions

Personality Dimension	Average Error Level	Prediction Reliability Tier
Conscientiousness	Lowest	High
Openness	Lowest	High
Extraversion	Medium	Medium
Agreeableness	Medium	Medium
Neuroticism	Highest	Low

4.3. Correlation analysis outcomes

Most score deviations between self-reported and AI-generated results fall within the 0-2 range of the Likert scale; only a small number of gaps exceed 3 points, and no extreme outliers interfere with the overall data distribution. Neuroticism still carries the highest average prediction error, while conscientiousness and openness produce minimal scoring deviations [10]. Table 3 records the Pearson correlation coefficients between subjective attitude indicators and total prediction error.

Table 3. Pearson correlation coefficients between subjective attitude indicators and total prediction error

Attitude Measurement Index	Correlation Coefficient with Total Prediction Error	Correlation Strength
Perceived AI personality understanding	-0.44	Moderate negative correlation

DOI: 10.54254/2755-2721/2026.BA37074

Table 3. (continued)

Trust in AI personality judgments	Close to 0	No obvious linear correlation
Perceived similarity between AI and human emotion	Close to 0	No obvious linear correlation
Overall subjective opinion of AI accuracy	Close to 0	No obvious linear correlation
Attitude Measurement Index	Correlation Coefficient with Total Prediction Error	Correlation Strength

5. Results

5.1. Hypothesis validation analysis

Both the hypotheses set at the beginning of the research have been proved statistically accurate. H1 has been confirmed, as the personality scores of all chatbots were estimated within the tolerable error rate, thus confirming the ability of LLMs to estimate personality traits accurately enough for adolescents [3]. H2 has been confirmed as well, as the prediction errors of conscientiousness and openness are much lower than that of neuroticism, which corresponds to the expected unequal measurement accuracy of personality aspects [10].

Conscientiousness and openness are characterized by stable linguistic patterns in the dataset used to train LLMs, thus making automatic evaluation of these personality traits stable and consistent. In contrast, neuroticism is based on private emotional states of individuals that lack a common linguistic framework and continue to produce huge discrepancies between the AI predictions and students' self-reports [10].

Regular interaction of teenagers with different chatbot systems on a day-to-day basis supplies enough context clues for the development of more precise personality profiling by LLMs. The given aspect serves as an explanation for the moderate negative correlation coefficient of -0.44 between subjects' perceptions of AI understanding and prediction errors in total [3]. With greater integration of AI chatbots into the daily life of teenagers, LLM personality testing can become a useful auxiliary measure.

5.2. Research limitations

There are two main limitations that hinder the generalization of the research findings. One is that the number of valid subjects used, which is 33 people, is quite low; thus, one cannot generalize the findings to the entire adolescent group. The second is that all the subjects belong to the same nationality – Chinese high school students studying according to Canadian educational curricula. Thus, this study is not applicable to other groups of people, education system, and other nationalities. Moreover, this research only uses subjective data obtained from BFI-10 test.

5.3. Suggestions for subsequent research

Future research on this topic could increase the sample size and use people from various ages and educational levels to ensure the generalizability of research findings. Integrating objective behavior data with self-reporting using questionnaires will serve to validate the actual prediction performance of LLMs. Additionally, future studies may delve into algorithmic reasons behind prediction

DOI: 10.54254/2755-2721/2026.BA37074

differences among various personality traits and develop personalized prompt templates specifically for personality assessment [6]. The Python program used for data analysis in this study can be adapted to bigger sample sizes in future studies.

6. Conclusion

This study analyzes questionnaire data from 33 Chinese high school students enrolled in the Canadian curriculum to examine the accuracy of personality prediction by mainstream large language model (LLM) chatbots. It comprehensively evaluates cross-platform performance differences, variations in prediction accuracy across the five major personality dimensions, and the linear relationship between participants' subjective attitudes and scoring biases, statistically confirming two pre-established hypotheses.

LLM chatbots demonstrate relatively consistent personality predictions for adolescent participants, with higher accuracy in conscientiousness and openness compared to neuroticism. Significant differences in prediction accuracy are evident among various mainstream chatbot tools. Adolescents who subjectively believe that artificial intelligence understands their personalities exhibit smaller scoring biases.

The paper provides original quantitative baseline data for future research on LLM-based personality prediction among youth. However, due to its small sample size and focus on a single demographic characteristic, the findings lack generalizability. Future studies could increase participant numbers and collect data from diverse sources to enhance the consistency of automated AI personality assessments. Meanwhile, the optimized Python analysis script may also serve as a useful tool for researchers.

References

[1] Jiang, L., Pellert, M., & Safdari, R. (2024). Psychometric validation of big five questionnaires for large language models. *arXiv Preprint*, arXiv:2408.04826.
[2] McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications.
[3] Wright, A., Chen, L., & Thompson, K. (2025). Generative AI personality prediction from personal narrative text. *Nature Human Behaviour, 9*(7), 1123–1132.
[4] Nguyen, H., & Park, S. (2024). Perceived AI understanding and objective personality prediction error. *arXiv Preprint*, arXiv:2405.16355.
[5] Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *Journal of Research in Personality, 41*(1), 203–212.
[6] Gupta, A., Song, Y., & Anumanchipalli, G. (2024). Prompt sensitivity in LLM personality measurement tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1950–1976.
[7] Kovačević, M., Horvat, T., & Šimunić, D. (2024). User expectation formation for conversational AI personality profiles. *arXiv Preprint*, arXiv:2401.08194.
[8] Peng, X., Qian, Y., & He, Y. (2026). When AI chatbots understand emotions: Exploring the mechanisms of self-disclosure from emotional arousal to psychological acceptance. *Frontiers in Psychology*.
[9] Binnquist, A. L., Dolbier, S. Y., Dieffenbach, M. C., & Lieberman, M. D. (2022). The Zoom solution: Promoting effective cross-ideological communication online. *PLOS ONE, 17*(7), e0270355.
[10] Krpan, D. (2025). Comparing chatbot personality inference to traditional psychometric testing. *Frontiers in Psychology, 16*, 1564979.