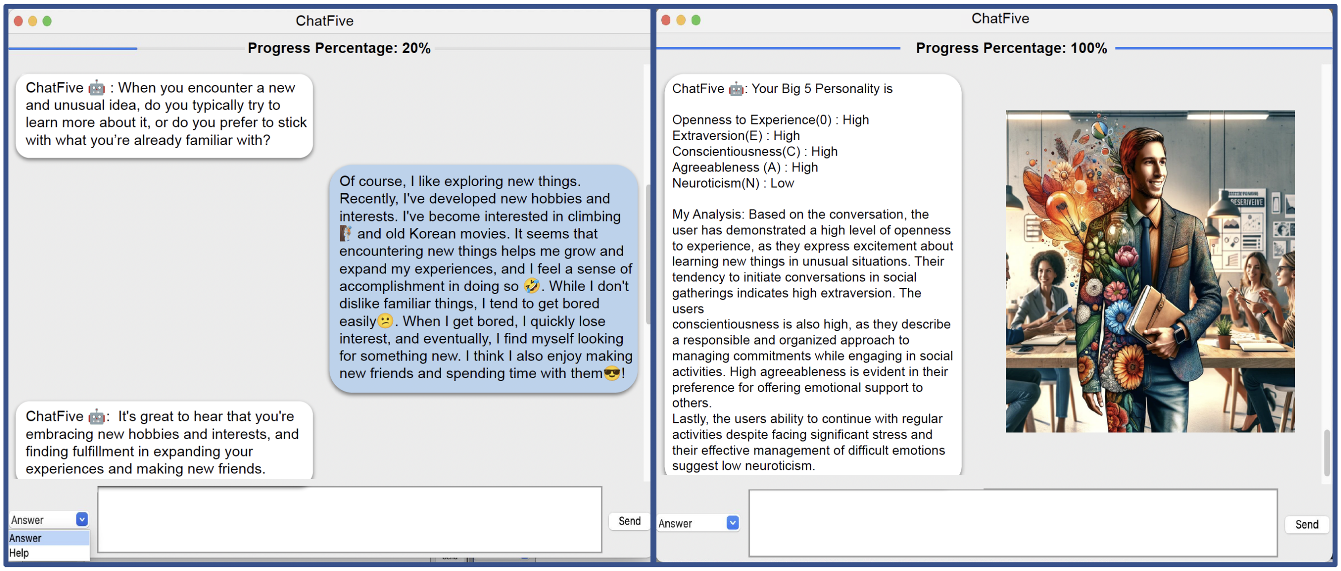
ChatFive: Enhancing User Experience in Likert Scale Personality Test through Interactive Conversation with LLM Agents

Jungjae Lee*
School of Computing, KAIST
Daejeon, Republic of Korea
dlwjdwo00701@kaist.ac.kr

Yubin Choi*
School of Computing, KAIST
Data Science Group, IBS
Daejeon, Republic of Korea
yubin.choi@kaist.ac.kr

Minhyuk Song
School of Computing, KAIST
Data Science Group, IBS
Daejeon, Republic of Korea
smh0706@kaist.ac.kr

Sanghyun Park
School of Computing, KAIST
Daejeon, Republic of Korea
sanghyun.park@kaist.ac.kr



ChatFive 🤖 : When you encounter a new and unusual idea, do you typically try to learn more about it, or do you prefer to stick with what you're already familiar with?

Of course, I like exploring new things. Recently, I've developed new hobbies and interests. I've become interested in climbing 🧗 and old Korean movies. It seems that encountering new things helps me grow and expand my experiences, and I feel a sense of accomplishment in doing so 🤣. While I don't dislike familiar things, I tend to get bored easily😕. When I get bored, I quickly lose interest, and eventually, I find myself looking for something new. I think I also enjoy making new friends and spending time with them😎!

ChatFive 🤖: It's great to hear that you're embracing new hobbies and interests, and finding fulfillment in expanding your experiences and making new friends.

ChatFive 🤖: Your Big 5 Personality is

Openness to Experience(0) : High
Extraversion(E) : High
Conscientiousness(C) : High
Agreeableness (A) : High
Neuroticism(N) : Low

My Analysis: Based on the conversation, the user has demonstrated a high level of openness to experience, as they express excitement about learning new things in unusual situations. Their tendency to initiate conversations in social gatherings indicates high extraversion. The users
conscientiousness is also high, as they describe a responsible and organized approach to managing commitments while engaging in social activities. High agreeableness is evident in their preference for offering emotional support to others.
Lastly, the users ability to continue with regular activities despite facing significant stress and their effective management of difficult emotions suggest low neuroticism.

Figure 1: ChatFive Interface: Right box shows Big Five Personality prediction explanations for Five traits (Low, Moderate, High) and a DALL·E generated personalized profile based on these results.

ABSTRACT

Personality assessments provide insights into understanding individual differences. In HCI, personality assessments are used to model user behavior or tailor user interfaces. However, conventional Likert-scale personality tests face issues in user engagement and capturing comprehensive personality nuances. Building upon prior work using conversational user interfaces for personality prediction, we delve deeper into personalized personality tests. Through a formative study (n=4), we identified three design goals for user engagement. Informed by these goals, we propose a novel architecture integrating multiple large language model agents to support free-form conversation-based personality assessment. Our system, ChatFive, predicts users' Big Five traits through real-time personalized dialogue. Evaluations from our user study (n=20) revealed that ChatFive significantly improved conveying true responses and felt more engaged, though requiring longer response times and different validation. We discuss the limitations on the validity of ChatFive and its implications.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Natural language interfaces**.

[*] Both authors contributed equally to this research.

CUI '24, July 08–10, 2024, Luxembourg, Luxembourg
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665572

KEYWORDS

Personality Test, Large Language Models(LLM), Conversational User Interface(CUI), Conversational Agents

ACM Reference Format:

Jungjae Lee, Yubin Choi, Minhyuk Song, and Sanghyun Park. 2024. ChatFive: Enhancing User Experience in Likert Scale Personality Test through Interactive Conversation with LLM Agents. In *ACM Conversational User Interfaces 2024 (CUI '24), July 08–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3640794.3665572

1 INTRODUCTION

The assessment of personality traits holds significant importance in comprehending individual differences, understanding others, being used as a predictive indicator of life outcomes [6, 29]. Conventionally, asessments have relied upon structured Likert-scale inventories for scalability such as the Big Five Inventory(BFI [13]). However, these approach is beset by several challenges, fake response [37] and limitations to fully encapsulate the intricate nuances and diversity inherent to individual characteristics [14]. While prior work have explored conversational interfaces for personality prediction to address previous challenges [10, 24, 26, 34], we exploration delves into the potential benefits of tailored questioning and natural language user responses to express true-self. We hypothesize that this approach could strike a balance between engagement and comprehensiveness, allowing diverse, authentic responses. Hence, we built ChatFive, predicting Big Five personality based on real-time personalized conversation using large language model(LLM) multiple-agents. We conducted a user study (n=20) comparing ChatFive to the online IPIP Big Five inventory baseline [1]. Our findings revealed that ChatFive notably enhanced engagement and provided a personalized experience. We discuss the limitations of assessing validity and reliability and the potential of using LLM into convert Likert-scale questionnaires to conversational ones.

Our main contributions are:

- **C1.** ChatFive: A novel system that supports personalized questioning and natural language based user replies for assessing the Big Five personality traits, powered by LLMs.
- **C2.** Empirical insights: Analysis of how users interact with ChatFive compared to the Likert-scale inventory, highlighting the differences in user experience.
- **C3.** Architectural framework: A proposed architecture for converting traditional Likert-scale personality inventory into engaging, personalized conversation.

2 RELATED WORK

ML based Personality Predictions. Several papers have tackled machine learning based personality detection to address aforementioned challenges of Likert-scale personality test. A large branch of work focuses on using the digital footprints such as Facebook profile features [12], Reddit [11], video interviews [15, 24, 31], smartphone data [5], or behavioral data [18] to predict Big Five, the most well researched personality measure. Another branch focuses on devising better algorithms to predict personality based on texts, as the original test was invented based on English lexicon analysis [7], using LSTM, Adaboost etc [8, 21]. However, such methods fundamentally rely on LIWC category of pre-defined words without using rich linguistic cues like words, nuance, etc.

Conversational User Interfaces. Conversational user interfaces (CUIs) offer a alternative to enhance engagement [19]. On this note, Celino et. al [4] explored conversational UX surveys leading to higher user response than quantified surveys. Furthermore, allowing users to express themselves intuitively through natural conversations mitigate issues like careless responding. Thus, a few recent research focuses on conversation-based personality prediction [24] and showed validity of using CUI to infer personality [26]. Recently, personality prediction using large language models [10, 34], known to reason with zero-shot [17, 20] when using appropriate prompts and leveraging multiple-agents, have been shed light. We concur with these direction but note the particularly pertinent to our direction on personalized questions enhancing user engagement [28, 30]. We take these ideas one step further and propose the concept of finding personality from personalized question tailored to each individual's natural language responses simultaneously to account for unique personality nuances and provide enhanced engagement.

3 FORMATIVE STUDY

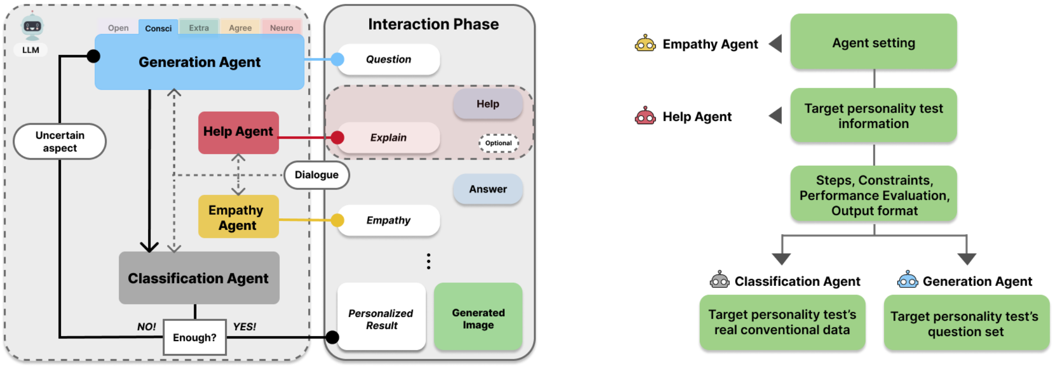
To gain insights into the UX challenges of Likert-scale personality tests and explore potential design considerations for conversational personality tests, we conducted a focus group interview. The group comprised 4 participants (M_{age}=25, 2 Males) recruited through a posting at university. Participants completed a worksheet, which involved: A. sharing their positive and negative experiences with previous personality tests, B. brainstorming what a conversational personality test might entail. The entire 57 minutes session was recorded and analyzed through the thematic approach [3]. The analysis focused on 1) the UX of personality tests and 2) expectations towards a conversational personality test.

Participants claimed utilizing personality assessments as tools for sincere self-reflection to monitor their psychology over time or for deriving enjoyment. Though appreciating the profile summaries and explanations, a consistent point of dissatisfaction emerged one-size-fits-all assessments. Furthermore, the ambiguity of Likert-scales posed difficulties across the board, as exemplified by FP2's struggle to differentiate *agree or slightly agree* on complex topics. Also, We found that people expect a personality test counselor that was completely personalized for them. Informed by these feedback, we check the potential of conversational test over Likert-scales. So, we aligned ChatFive's functionality and devised additional three key design goals (DGs):

- **DG1.** Understand users as they share their thoughts and experiences in conversation.
- **DG2.** Allow users to inquire about ChatFive's generated question (e.g., clarify term, meaning).
- **DG3.** Give feedback on users' answers to show understanding and personalized profile.

4 CHATFIVE ARCHITECTURE

Prior studies and formative study showed conversing with chatbots could enhances engagement over Likert-scale personality tests. To fulfill a comprehensive counselor's role participants envisioned, a system is needed that can maintain personalized conversation



(a) System Framework

(b) Prompt Framework

Figure 2: ChatFive Architecture: (a) is a framework that shows the phase of user interaction by each system agent. The agents operate based on a past dialogue with the user. (b) is the framework for the needed information that construct each agent.

with user while simultaneously analyzing their response and giving appropriate follow-up questions. So, we leveraged multiple LLM agents that are capable of actively high-performing individually with autonomy and rich contextual understanding enables personalized responses. Hence, we propose an architecture designed to convert targeted Likert-scale personality tests into conventional formats, as illustrated in Fig. 2. Additionally, we incorporated features inspired by the design goals (DGs) in Section 3.

4.1 System Framework

Our proposed system framework comprises four key agents: *Classification*, *Generation*, *Empathy*, and *Help*. We expect to maintain the user's interest and engagement throughout the test by effectively organizing these personalized agents.

4.1.1 Classification Agent (CA). *CA* decides when to end the interaction with the user. *CA* analyzes the past dialogue with the user to determine whether the test should 'complete' or 'continue'. The dialogue comprises a sequence of the questions asked by *GA* and the user's answers.

CA concludes the conversation if the dialogue reaches a stage where the user's test result can be determined. It then analyzes the dialogue with self-reasoning, identifying which questions and answers influenced it, and presents a detailed interpretation. However, *CA* should not determine the outcome with just a few questions. It should confidently do so after conducting multiple checks. This personalized analysis is based on the user's conversation with the system. Users who encounter these results, which vary from user to user, will feel more personalized. In addition, the system can perform additional functions based on these personalized results (image generation, ideal matching, etc.). These additional functionalities can increase personalization.

If the dialogue is insufficient to judge the result, *CA* will determine which trait is uncertain and pass that factor to *GA* to continue the dialogue. Together, *CA* generate and deliver feedback on which direction to ask more.

4.1.2 Generation Agent (GA). When requested from*CA*, *GA* generates a question. The question is presented to the user, who then responds. If the targeted personality test consists of multiple traits, the agent is divided into each type and invoked according to the uncertain trait determined by *CA*. During generation, *GA* analyzes previous dialogue and itself to understand the user's experiences and thoughts. This enables the generation of more personalized questions with different for each user. However, in order to maintain the validity of the questions in the targeted test, *GA* self-analyze the intent of the original questions and processed to personalize them while maintaining the intent.

4.1.3 Empathy Agent (EA). During our formative study, we discovered that a one-directional [question, answer] interaction between the user and the system can make users feels less like they're being consulted, which tends to be less engaging and can result in diminished interest. This insight guided us to DG3, which was solved with *EA*. *EA* generates a summary of the user's answer and a sentence that implies some empathy from the counselor's perspective within the targeted test.

4.1.4 Help Agent (HA). In our formative study, we discovered that the fixed questions in many existing Likert-scale personality tests are frequently unclear to individuals. This situation results in users facing challenges in providing accurate answer and It reduces the feeling of being counseled, leading to the idea of DG3. We implemented *HA*, which allows users to ask for additional explanations of the generated questions. The dialogue in this agent is structured differently from the other agents. A separate discussion dialogue is created for the question that the user has difficulty understanding, and it expands. As the discussion advances, the user gains a deeper understanding of the generated questions, enabling them to provide more accurate responses.

4.1.5 Interaction Phase. As shown in Fig. 2(a), *GA* and *CA* repeatedly alternate until the test concludes. The question generated by *GA* is presented to the user. As the dialogue progresses, *CA* assesses and guides the direction of the test. However, if the user encounters difficulty comprehending the question, the system pauses the loop to interact with *HA* to explain more. Moreover, following the user's answer, *EA* generates a brief empathetic sentence.

4.2 Implementation

We implemented the prototype ChatFive through our proposed architecture sing the BFI-50 [1]. The prototype was implemented using GPT-4 [27]. In particular, the *CA* analyzes with CoT [32] and is prompted with a variation of React [35]. Additionally, the final scores that ChatFive returns for each type are within the range of 0 to 100.

About the Fig. 2(b)'s prompt framework, 'Target personality test information' entered in the prompt was summarized from the wiki and the paper [7] about the Big Five. In the case of 'Target questionnaire real conventional data', the data was collected from authors, and the prompt element was filled in with a summary of what was different or noteworthy from what the system already knew about the target test. For the 'Target questionnaire's question set', we entered ten questions of each type into each *GA*.

5 USER STUDY

We evaluated ChatFive in a within-subject with 20 volunteer participants. Our purpose of the user study was to evaluate ChatFive compared to the baseline [1] based on the following RQs:

- **RQ1.** How does ChatFive influence overall UX, such as user engagement, perspicuity, and usability? (UX)
- **RQ2.** How effectively do ChatFive's conversational agents interact with users? (Agents quality)
- **RQ3.** How accurate is ChatFive in predicting Big Five personality? (Accuracy)

To address RQ1, we used the User Experience Questionnaire (UEQ) [22], and a System Usability questionnaire (UMUX-LITE) [23]. As for RQ2, we used 12 questions based on the conversational agents' evaluation review paper [36]. The questions focused on UX and perceptions towards conversational agents. There was an optional open-ended text question to explain the reason behind their Likert-scale choice.

Participants. We posted recruitment postings on our class Teams page and snowball sampled along the way. Twenty participants were recruited (9 male, M_{age}=24.4, std=1.4). All participants had experience taking the MBTI test, but only one had taken the Big Five (P1). Based on the pre-survey results, four participants were very interested in personality tests, nine participants were somewhat interested, five were neutral, and two were not interested.

Procedure. Participants agreed to an IRB form approved by our university. Then, participants tried out the online Big Five Inventory and ChatFive (order counterbalanced). Following each task, they completed the Likert survey. Additionally, a 20-minute optional interview was conducted with three participants (P1, P11, P13).

6 RESULTS

Overall, 17 participants preferred ChatFive for future personality tests over the baseline [1], citing its freshness making them *"curious of what kind of conversation will happen next time."* (P13). The SUS score was 60.4 from converting UMUX-Lite using bootstrap method [9]. For survey analysis, we first assessed the normality using the Shapiro-Wilk test. Then, we used paired t-tests or Wilcoxon signed-rank test was used for parametric and non-parametric data, respectively.

6.1 ChatFive is Engaging but took time (RQ1)

The ambiguity of expressing oneself in Likert-scale Personality tests was resolved in ChatFive. Users felt less burdened because to categorize themselves into Likert-scale options (P9, P15, P17) and enjoyed expressing their true thoughts freely ($p < 0.001$ Fig. 4 (A)). Another distinct UX aspect was that ChatFive was very fun and satisfying. Conversing with agents was relieving and happy. Some participants even said though ChatFive took 32 minutes, it didn't feel like so (P16). The major fun factor was open conversation where participants could talk about anything they wanted with some participants claiming to be relieving (P13). Similar results from UEQ showing higher *attractiveness* and *stimulation* for ChatFive compared to our baseline with Cronbach's Alpha > 0.75 (Figure 3). While the sample size is only 20 to conclude with Cronbach's Alpha, this fun factor was mentioned in open responses and effected the high overall satisfaction of ChatFive in Fig. 4. Nevertheless, ChatFive's response time received a lower score than the baseline, as shown in Fig. 4 (A) and (B) due to the computational time required by the OpenAI API. The full distribution of replies can be seen in Appendix Fig. A1 and Fig. A2.

6.2 ChatFive Agent felt personalized (RQ2)

Participants claimed that the conversational agents were natural and personalized (Fig. 4 (B)). A prominent reason, as noted by most participants (P1-P16, P18-P20), was the progressive personalization of questions. Which P9 described as *"the more you answer, the more the question becomes centered around you"*. ChatFive's empathetic summaries also contributed to its personalized feeling, which made participants perceive ChatFive to understand their answers and feel their *voice was being heard* (P2). Based on such personalized questions, participants were able to express their thoughts freely to the conversational agents. Participants averaged 13.25 turns per test (std: 4.19, max: 24, min: 8) and wrote an average of 19.32 words per turn (std: 15.95, max: 115, min: 1). Notably, in 9 cases, the final answer word count was longer than initial answer, suggesting increased engagement over time. Another personalized factor was the profile made by the agent using DALL·E 3 [2]. Participants, like P12 who wanted to keep her profile, valued and liked their personalized profiles.

6.3 Accuracy Results (RQ3)

Table 1 presents the Pearson correlation between the Big Five trait scores predicted by ChatFive and the baseline (IPIP BFI-50 [1]). This approach aligns with established machine learning methods for personality prediction [8, 21]. We observe substantial variations in the correlation across traits. Conscientiousness and Extraversion exhibit relatively high correlations (0.77), while Agreeableness and Openness show moderate correlations (0.33). Notably, the correlation for Neuroticism is almost negligible. These discrepancies may partially be attributed to linguistic cues, such as the use of pronouns, social process language, which are known indicators of distinct Big Five dimensions [16]. However, given the drastic differences, we further explore the validity concerns in the Discussion section 7.

Traits	**Openness(O)**	**Conscientiousness(C)**	**Extraversion(E)**	**Agreeableness(A)**	**Neuroticism(N)**
Pearson Correlation	0.33	0.77	0.77	0.35	-0.02

Table 1: Pearson correlation between ChatFive prediction and BFI inventory results. The score ranges from 0 to 100.

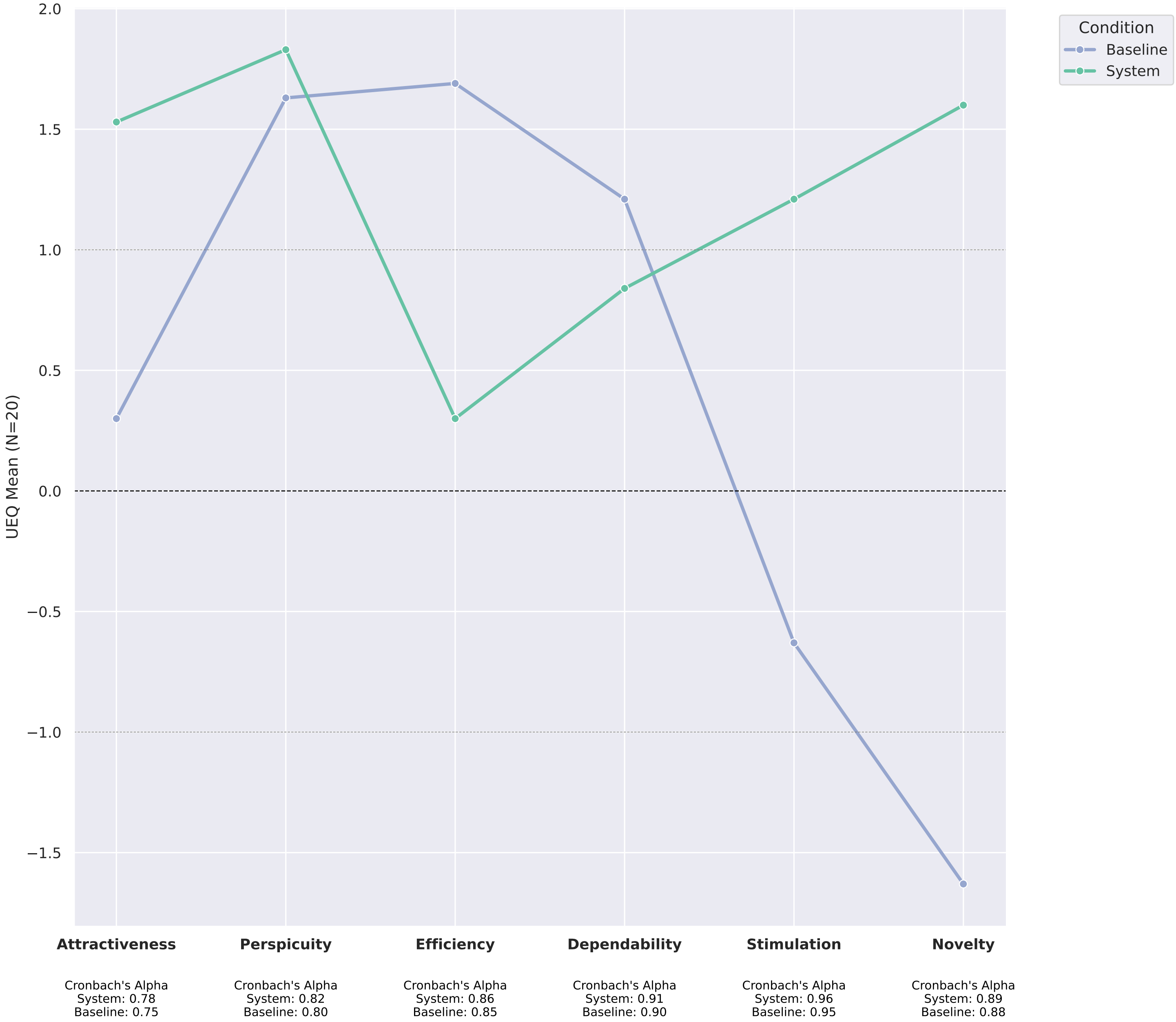


Figure 3: UEQ results of Big Five Inventory Baseline(A) [1] and ChatFive(B)

7 DISCUSSION

Validity of ChatFive. Deviating from standardized, structured questionnaires impacts the psychometric properties, such as construct validity and reliability. As an exploratory work contributing a novel perspective to personality assessment, establishing the validity of this approach was the primary initial focus, given the inherent challenges in assessing reliability across the varying individual instances of ChatFive’s questions.

Section 6.3 showed that ChatFive’s Big Five predictions diverged from the baseline. Analysis of the conversation logs revealed this stemmed from losing middle contextual elements in longer conversations - common in other LLMs [25, 33]. To improve the overall accuracy, ChatFive will analyze each question-answer pair, reducing analytic load, and separating predictions by trait. Further, to

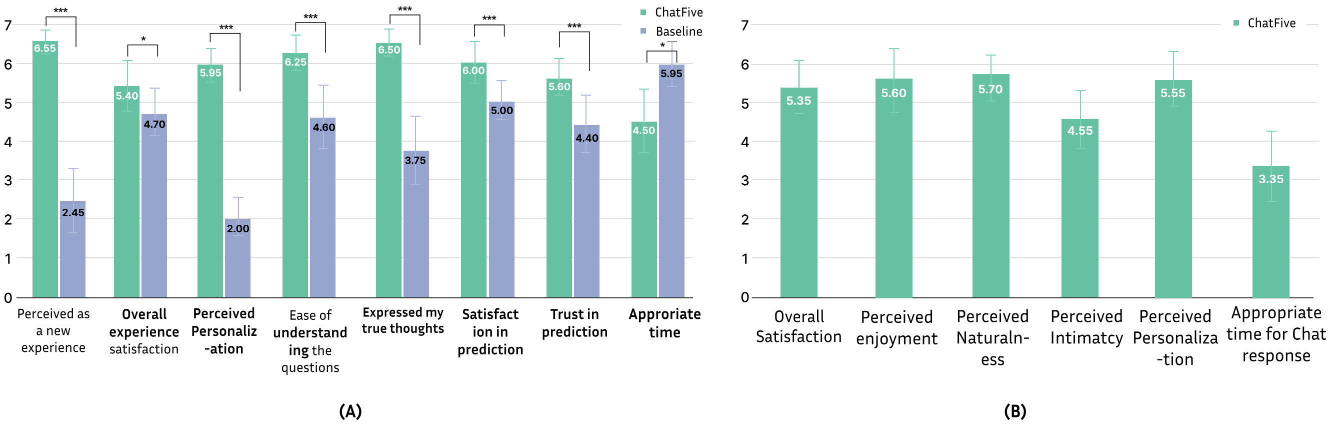


Figure 4: (A) displays the average scores from a 7-point Likert Scale UX Survey for ChatFive and the baseline[1]. (B) shows the survey averages specifically focusing on the conversational agents' quality of ChatFive. *: p<0.05, **: p<0.01, and ***:p<0.001. The full questionnaire and distribution can be seen in Appendix A.

explain the significant correlation difference in Big Five traits, analysis shows that ChatFive generated numerous questions for highly correlated traits like Extraversion but ended conversations early for low-correlation traits, potentially struggling to understand certain traits. To improve alignment with established tests, we will conduct ablation studies across agents and fine-tune the underlying language model using supplementary data focused on the weaker trait areas. While acknowledging trade-offs, this exploratory approach offers unique contributions for enhancing user experiences in personality assessment over the long term.

Time Consumption Problem. We faced slow response latency with OpenAI [27], impacting user experience. This can be improved from the conversation log analysis showing that users took 62.7 seconds per question, and agents needed 46.6 seconds for question generation and analysis. This suggests loading times could be reduced by performing analysis/generation during user response periods.

Future Works. ChatFive, converting the Likert-scale Big Five test to conversational format, enhanced user experience. We will explore cross-domain scalability in applying the framework to other Likert-scale tests like the depression inventory where user-generated rich content can be valuable. Moreover, we will compare the fine-tuned ChatFive about the traits with a mixed model employing conversational interaction(ChatFive) and traditional ML predictions. To establish the reliability, we will perform repeated evaluations by domain experts in psychology.

8 CONCLUSION

This research investigates a conversational interface for the Big Five personality inventory. We present ChatFive, leveraging LLM agents for real-time dialogue personalized to user responses. Our user study shows ChatFive enhances UX in engagement, clarity, and satisfaction by enabling free response over Likert-scales. However, there were trade-offs in response time and validity of prediction. We discuss implications and future directions to enhance reliability and validity. This work aims to establish a robust conversational interface for scaling psychometric assessments.

ACKNOWLEDGMENTS

We are grateful to Professor Geehyuk Lee for his invaluable guidance and insightful feedback throughout this work. Our sincere thanks go to the KAIST School of Computing for providing the resources to utilize the OpenAI API, which was instrumental in our research. We would also like to express our appreciation to our classmates from CS584 for their constructive feedback and willing participation in our study; their contributions were indispensable to this endeavor.

REFERENCES

[1] Accessed 2023. https://openpsychometrics.org/tests/IPIP-BFFM/
[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* (2023).
[3] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.
[4] Irene Celino and Gloria Re Calegari. 2020. Submitting surveys via a conversational interface: an evaluation of user acceptance and approach effectiveness. *International Journal of Human-Computer Studies* 139 (2020), 102410.
[5] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. 2013. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing* 17 (2013), 433–450.
[6] Andreas Demetriou, George Spanoudis, Constantinos Christou, Samuel Greiff, Nikolaos Makris, Mari-Pauliina Vainikainen, Hudson Golino, and Eleftheria Gonida. 2023. Cognitive and personality predictors of school performance from preschool to secondary school: An overarching model. *Psychological Review* 130, 2 (2023), 480.
[7] John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology* 41, 1 (1990), 417–440.
[8] Kamal El-Demerdash, Reda A El-Khoribi, Mahmoud A Ismail Shoman, and Sherif Abdou. 2022. Deep learning based fusion strategies for personality prediction. *Egyptian Informatics Journal* 23, 1 (2022), 47–53.
[9] Kraig Finstad. 2010. The usability metric for user experience. *Interacting with computers* 22, 5 (2010), 323–327.
[10] Ivar Frisch and Mario Giulianelli. 2024. LLM Agents in Interaction: Measuring Personality Consistency and Linguistic Alignment in Interacting Populations of Large Language Models. *arXiv preprint arXiv:2402.02896* (2024).
[11] Matej Gjurković and Jan Šnajder. 2018. Reddit: A gold mine for personality prediction. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media.* 87–97.
[12] Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems.* 253–262.
[13] Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. 2006. The international personality item pool and the future of public-domain personality measures.

Journal of Research in personality 40, 1 (2006), 84–96.

[14] Neil MA Hauenstein, Kevin M Bradley, Patrick Gavan O'Shea, Yashna J Shah, and Douglas P Magill. 2017. Interactions between motivation to fake and personality item characteristics: Clarifying the process. *Organizational Behavior and Human Decision Processes* 138 (2017), 74–92.

[15] Louis Hickman, Nigel Bosch, Vincent Ng, Rachel Saef, Louis Tay, and Sang Eun Woo. 2022. Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology* 107, 8 (2022), 1323.

[16] Joanne Hinds and Adam Joinson. 2019. Human and computer personality prediction from digital footprints. *Current Directions in Psychological Science* 28, 2 (2019), 204–211.

[17] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*. PMLR, 9118–9147.

[18] Seoyoung Kim, Jiyoun Ha, and Juho Kim. 2018. Detecting personality unobtrusively from users' online and offline workplace behaviors. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[19] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[20] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[21] Aditi V Kunte and Suja Panicker. 2019. Using textual data for personality prediction: a machine learning approach. In *2019 4th international conference on information systems and computer networks (ISCON)*. IEEE, 529–533.

[22] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4*. Springer, 63–76.

[23] James R Lewis, Brian S Utesch, and Deborah E Maher. 2013. UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2099–2102.

[24] Jingyi Li, Michelle X Zhou, Huahai Yang, and Gloria Mark. 2017. Confiding in and listening to virtual agents: The effect of personality. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 275–286.

[25] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.

[26] Yunxing Liu and Jean-Bernard Martens. 2022. Conversation-Based Hybrid User Interface for Structured Qualitative Survey: A Pilot Study Using Repertory Grid. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8.

[27] openai. 2023. *Creating safe AGI that benefits all of humanity.* openai. Retrieved Nov 11, 2023 from https://openai.com/

[28] Susanna Pardini, Silvia Gabrielli, Marco Dianti, Caterina Novara, Gesualdo M Zucco, Ornella Mich, and Stefano Forti. 2022. The role of personalization in the user experience, preferences and engagement with virtual reality environments for relaxation. *International Journal of Environmental Research and Public Health* 19, 12 (2022), 7237.

[29] Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. 2007. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science* 2, 4 (2007), 313–345.

[30] Idan Szpektor, Yoelle Maarek, and Dan Pelleg. 2013. When relevance is not enough: Promoting diversity and freshness in personalized question recommendation. In *Proceedings of the 22nd international conference on World Wide Web*. 1249–1260.

[31] Ronnie Taib, Shlomo Berkovsky, Irena Koprinska, Eileen Wang, Yucheng Zeng, and Jingjie Li. 2020. Personality sensing: Detection of personality traits using physiological responses to image and video stimuli. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 3 (2020), 1–32.

[32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[33] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large Language Models as Optimizers. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Bb4VGOWELI

[34] Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024. LLM Agents for Psychology: A Study on Gamified Assessments. *arXiv preprint arXiv:2402.12326* (2024).

[35] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022).

[36] Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 2022. UX research on conversational human-AI interaction: A literature review of the ACM Digital Library. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[37] Matthias Ziegler, Carolyn MacCann, and Richard Roberts. 2012. *New perspectives on faking in personality assessment.* Oxford University Press.

A APPENDIX

A.1 User Study Results

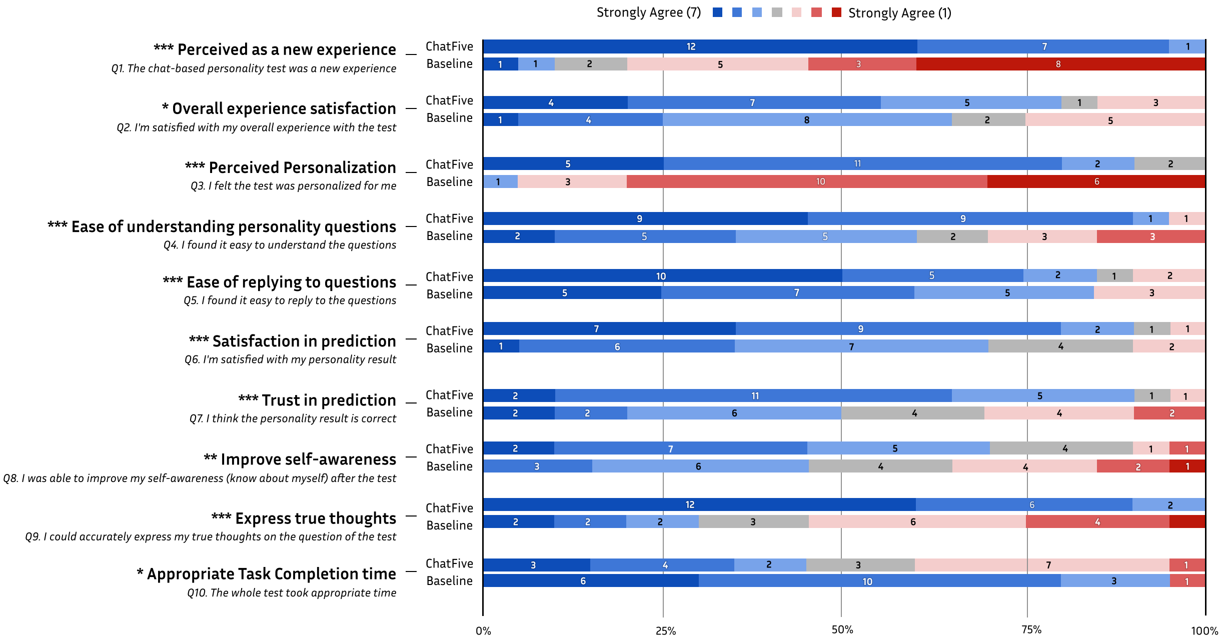


Figure A1: Comparative Distribution of UX Survey Question Ratings: ChatFive vs. Baseline on a 7-Point Likert Scale (n=20). *: p<0.05, **: p<0.01, and ***:p<0.001.

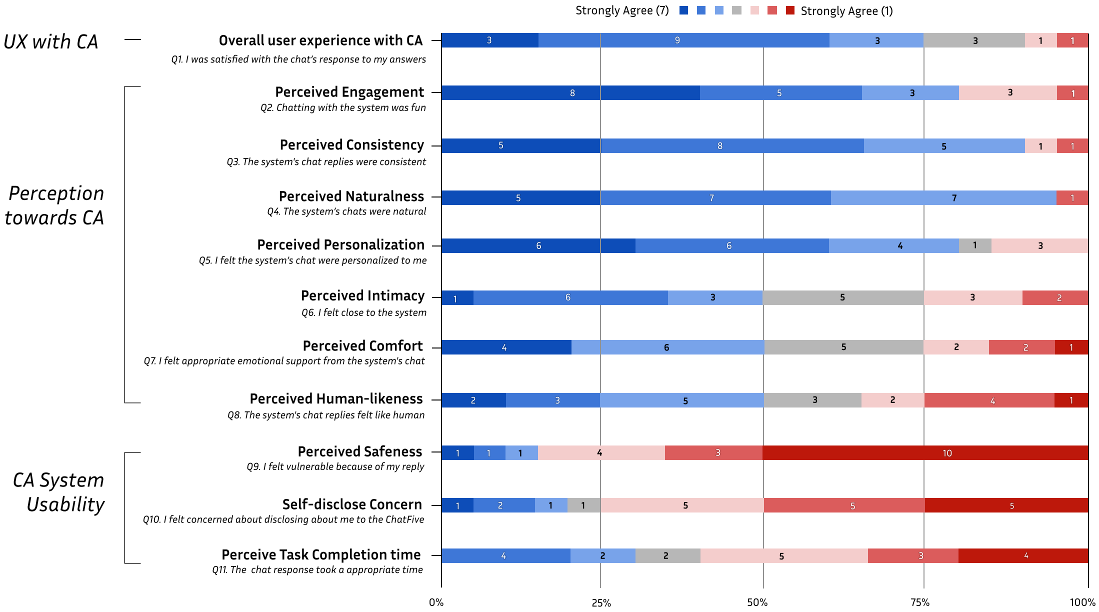


Figure A2: Distribution of ChatFive on Conversation Agent quality Ratings (n=20). *: p<0.05, **: p<0.01, and ***:p<0.001.