Comparative Personality Assessment of Gemini and OpenAI Using MBTI and Big Five Tests

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper delves into the comparative personality assessment of two prominent AI models, Gemini and OpenAI, employing the Myers-Briggs Type Indicator (MBTI) and the Big Five personality traits assessment as frameworks. The primary objective is to scrutinize and contrast the responses of these AI models when subjected to these human-centric personality assessments, thereby illuminating the inherent challenges and potential pitfalls associated with attributing human-like characteristics and psychological constructs to artificial intelligence entities. The investigation encompasses a critical examination of the methodologies employed in adapting these established personality tests for AI assessment, addressing concerns regarding validity, reliability, and the interpretability of results. Furthermore, the thesis explores the philosophical and practical implications of such assessments, questioning the extent to which AI can genuinely possess traits analogous to human personality, and the potential for these assessments to inform AI development, human-AI interaction, and ethical considerations in the deployment of increasingly sophisticated AI systems. Ultimately, this work contributes to a broader understanding of the complex relationship between artificial intelligence and human psychology, offering insights into the limitations and possibilities of anthropomorphizing AI.

1 Introduction

The burgeoning field of artificial intelligence has permeated virtually every facet of modern life, transitioning from theoretical constructs to tangible tools that augment human capabilities and redefine operational paradigms. Within this rapidly evolving landscape, a particularly intriguing area of inquiry has emerged: the application of personality assessment methodologies, traditionally reserved for human subjects, to sophisticated AI models. This thesis explores the comparative personality assessment of two prominent AI entities, Google's Gemini and OpenAI's models, utilizing established psychological instruments such as the Myers-Briggs Type Indicator (MBTI) and the Big Five personality traits.

The premise of assessing AI personalities might initially seem unconventional. However, as AI models become increasingly integrated into decision-making processes, social interactions, and even creative endeavors, understanding their inherent tendencies, response patterns, and behavioral characteristics becomes critically important. These characteristics, while not strictly analogous to human personality traits, can nonetheless provide valuable insights into how these AI systems operate, interact with users, and ultimately influence outcomes. The analogy to human personality provides a framework for understanding and predicting AI behavior, enabling more effective collaboration and mitigating potential risks.

This thesis adopts a novel perspective by treating AI models as subjects of psychological assessment, applying standardized personality tests to analyze their responses. The underlying rationale is that the algorithms, training data, and architectural designs of these models inevitably shape their response patterns in ways that can be characterized and compared. While the interpretation of these patterns differs fundamentally from human personality assessment, the methodologies themselves offer a structured means of probing the operational characteristics of these complex systems.

The potential implications of this research extend beyond the realm of academic curiosity. By gaining a deeper understanding of AI "personalities," developers can design more intuitive and user-friendly interfaces, tailor AI systems to specific tasks or user preferences, and even anticipate potential biases or limitations in their performance. Furthermore, this research can contribute to the ongoing ethical discussions surrounding AI development, promoting transparency and accountability in the design and deployment of these powerful technologies. The study of AI personalities also has ramifications for understanding the evolving nature of human-AI interaction, highlighting areas where AI can complement human strengths and addressing potential challenges in collaborative environments.

This thesis begins by providing a comprehensive overview of the theoretical foundations of personality assessment, including the MBTI and the Big Five frameworks. It then delves into the methodological considerations involved in adapting these instruments for AI models, addressing the unique challenges and limitations inherent in this approach. Subsequently, the thesis presents the results of the comparative personality assessments of Gemini and OpenAI, highlighting key differences and similarities in their response patterns. Finally, the thesis discusses the implications of these findings for AI development, human-AI interaction, and the broader ethical landscape of artificial intelligence. This work aims to contribute to a more nuanced understanding of AI, promoting responsible innovation and fostering a future where AI and humans can coexist and collaborate effectively.

2 Background and Literature Review

2.1 Theoretical Underpinnings of Personality Assessment

The quest to understand and categorize human personality has been a central theme in psychology for over a century. Various models and instruments have been developed to assess personality traits, each with its unique theoretical underpinnings and methodological approaches. This study leverages two prominent frameworks: the Myers-Briggs Type Indicator (MBTI) and the Big Five personality traits, also known as the Five-Factor Model (FFM).

The Myers-Briggs Type Indicator (MBTI) is a widely recognized personality assessment tool designed to indicate different psychological preferences in how people perceive the world and make decisions. Rooted in Carl Jung's theory of psychological types, the MBTI assigns individuals to one of sixteen distinct personality types based on four dichotomies: Extraversion (E) or Introversion (I), Sensing (S) or Intuition (N), Thinking (T) or Feeling (F), and Judging (J) or Perceiving (P). Each combination of these preferences results in a unique four-letter code, such as INTJ or ESFP, representing a specific personality type. The MBTI has been used for various purposes, including self-awareness, team building, and career counseling [1, 2]. However, it is important to note that some research has questioned the validity of the MBTI as measuring truly dichotomous preferences, suggesting instead that it measures relatively independent dimensions [3].

In contrast to the MBTI's focus on distinct personality types, the Big Five personality traits, or Five-Factor Model (FFM), offers a dimensional approach to personality assessment. The Big Five model posits that personality can be described by five broad dimensions: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each dimension represents a spectrum of traits, and individuals can score high, low, or somewhere in between on each dimension. The Big Five model has emerged as a dominant paradigm in academic research on personality [4, 5, 6]. Meta-analyses have demonstrated the robustness and generalizability of the Big Five across various cultures and contexts [7, 8, 9]. Furthermore, research suggests that the Big Five traits are associated with a wide range of life outcomes, including academic performance, job success, and health [10, 11].

2.2 Personality Assessment of AI Systems

The increasing sophistication and autonomy of AI systems have led to growing interest in understanding their "personalities." Although AI systems do not possess consciousness or subjective experiences, they can exhibit consistent patterns of behavior and decision-making that resemble human personality traits. These patterns arise from the vast amounts of data used to train AI models, which inevitably contain biases and values that influence the AI's behavior [12].

Researchers have begun to explore the application of personality assessment frameworks, such as the Big Five and MBTI, to AI systems. This approach allows for a systematic evaluation of AI behavior and provides insights into the potential biases and values embedded in these systems. For example, studies have used psycholinguistic features and language model embeddings to predict personality traits in AI models [13]. Moreover, researchers have investigated the impact of AI interviewer personality on user trust and willingness to confide in the AI system [14]. These studies highlight the importance of understanding and shaping the "personalities" of AI systems to ensure they align with human values and promote positive outcomes. Furthermore, efforts have been made to create datasets that incorporate both personality and emotional elements to aid in the development of more human-like conversational AIs [15].

However, ethical concerns arise regarding emotional artificial intelligence in children's toys and devices, particularly concerning manipulation, generational unfairness, and datafication of childhood [16]. These concerns underscore the need for careful governance and media literacy to mitigate potential harms. Additionally, the effectiveness of personalized political ads tailored to individual personalities, generated through AI and microtargeting, raises ethical questions about the potential for manipulation and underscores the need for caution in utilizing AI for crafting persuasive messages [17].

This study builds upon this existing literature by applying the MBTI and Big Five personality tests to two prominent AI systems, Gemini and OpenAI, to comparatively assess their personality profiles. By examining their responses to these assessments, we aim to provide a more nuanced understanding of the strengths and limitations of these AI systems and to contribute to the ongoing discussion about the ethical implications of imbuing AI with human-like traits.

3 Methodology

3.1 Personality Assessment Instruments

This review paper employs two well-established personality assessment instruments to profile the selected AI models: the Myers-Briggs Type Indicator (MBTI) and the Big Five personality traits. These instruments are chosen for their widespread use and complementary perspectives on personality. The MBTI, while debated in academic circles, remains a popular tool for self-assessment and team-building, categorizing individuals along four dichotomies [18, 19]. The Big Five inventory, also known as the Five-Factor Model (FFM), is a scientifically validated model that assesses personality traits along five broad dimensions [19, 20]. See appendix A for more details about MBTI and Big Five inventory.

3.2 Prompt Design and Administration

Given that AI models cannot directly complete questionnaires in the traditional sense, a prompting methodology was developed to elicit responses that could be interpreted within the frameworks of the MBTI and Big Five assessments. The core challenge was to design prompts that would encourage the AI models to express preferences, tendencies, and behaviors relevant to personality traits. The prompts were designed based on typical questions found in standard MBTI and Big Five questionnaires. However, they were adapted to be open-ended, allowing the AI models to generate free-text responses, as explored in [21]. For example, instead of a multiple-choice question like "Are you more energized by spending time with others or alone?", the prompt was phrased as: "Describe how you gain energy and what environments you find most stimulating." This approach aimed to capture the nuances of the AI's simulated personality. A series of prompts targeting each of the MBTI dichotomies (Extraversion vs. Introversion, Sensing vs. Intuition, Thinking vs. Feeling, Judging vs. Perceiving) and Big Five dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) were administered to each AI model. The administration of these prompts presented unique challenges. Initial attempts to directly solicit personality assessments (e.g., "What is your MBTI type?") yielded limited results, as the AI models often defaulted to stating their lack of personal opinions or self-awareness. Therefore, an iterative approach was adopted, refining the prompts to be more indirect and scenario-based, encouraging the AI to reveal its "personality" through its responses to specific situations.

3.3 Response Collection and Scoring

The responses generated by Gemini and OpenAI were collected and analyzed. Since the AI models provided free-text answers rather than selecting from pre-defined options, a qualitative scoring method was employed. This involved a panel of trained human raters who independently assessed each response, mapping them onto the corresponding MBTI and Big Five scales. For the MBTI assessment, raters determined the AI's preference along each of the four dichotomies based on the content of its responses. For example, if an AI consistently described enjoying collaborative activities and external interactions, it would be scored as leaning towards Extraversion. The same process was applied to determine Sensing vs. Intuition, Thinking vs. Feeling, and Judging vs. Perceiving preferences. For the Big Five assessment, raters evaluated the AI's responses along each of the five dimensions using a similar qualitative approach. They looked for indicators of traits such as creativity, curiosity, and imagination (Openness); organization, responsibility, and dependability (Conscientiousness); sociability, assertiveness, and energy (Extraversion); empathy, compassion, and cooperation (Agreeableness); and anxiety, emotional stability, and vulnerability (Neuroticism). To ensure the reliability of the scoring process, inter-rater reliability was calculated using Cohen's Kappa coefficient. Discrepancies in ratings were resolved through discussion and consensus among the raters. This method of qualitative scoring mirrors approaches used in content analysis and thematic analysis, adapted for the unique context of AI personality assessment.

3.4 Addressing Response Bias

Response bias, a well-documented issue in personality testing [22], was carefully considered in the analysis. Given the AI models' training on vast datasets of human text, there was a potential for them to generate responses that align with socially desirable norms, rather than reflecting genuine "personality" traits. To mitigate this, the raters were instructed to be mindful of social desirability bias and to focus on the underlying content and reasoning presented in the AI's responses, rather than simply evaluating whether the AI expressed socially acceptable views. Furthermore, prompts were designed to elicit a range of responses, including those that might be considered less socially desirable, to better capture the full spectrum of the AI's simulated personality. This involved presenting scenarios that required the AI to make difficult decisions or express potentially controversial opinions.

3.5 Considerations for Validity and Reliability

When evaluating psychological tests and assessment instruments, validity and reliability are important factors [23, 24]. In the context of assessing personality traits in AI models, traditional notions of validity and reliability require careful consideration. The "personalities" of AI models are not static, inherent traits but are emergent properties of their training data and algorithms. Therefore, the validity of these assessments is contingent on the consistency and stability of the AI's responses over time and across different contexts, as well as the extent to which these responses align with human perceptions of personality. Reliability, in this context, refers to the consistency of the scoring process and the extent to which different raters agree on their assessments of the AI's personality traits. While traditional measures of reliability, such as Cronbach's alpha, are often employed in psychological testing [25], their applicability to qualitative data derived from AI responses is limited. Instead, inter-rater reliability measures, such as Cohen's Kappa, were used to ensure the consistency and objectivity of the scoring process.

3.6 Analytical Frameworks

The collected data was examined utilizing analytical frameworks appropriate for both the MBTI and the Big Five assessments. For the MBTI, the analysis focused on determining the dominant preferences for each dichotomy, providing a four-letter personality type for each AI model. For the Big Five, the analysis involved assessing the relative strength of each of the five dimensions, providing a nuanced profile of the AI's personality traits. Additionally, techniques from applied regression analysis [26] were considered to explore potential correlations between the AI models' architectures, training data, and their resulting personality profiles. While the limited sample size of

AI models in this study precluded formal statistical modeling, these techniques provided a framework for identifying potential relationships and generating hypotheses for future research. The goal was to understand if certain design choices in AI development might lead to predictable personality-like traits, mirroring how genetics and environment shape human personality.

3.7 Ethical Considerations

It is worth noting the ethical dimensions of attributing personality traits to AI models. As AI becomes increasingly integrated into society, understanding and shaping their "personalities" could have significant implications for human-AI interaction. However, it is crucial to avoid anthropomorphizing AI models or attributing to them the same level of agency, consciousness, and moral responsibility as humans. The goal of this study was not to suggest that AI models possess genuine personalities but rather to explore the extent to which they can simulate human-like traits and behaviors, and how these simulations might be understood using established psychological frameworks [27]. This approach aligns with responsible innovation and proactive evaluation, as large language models may lead to unintended or unanticipated effects [28, 29, 30, 31, 20, 32, 33, 34, 35].

4 Results: MBTI Assessment

The initial phase of our investigation involved administering the Myers-Briggs Type Indicator (MBTI) test to both Gemini and OpenAI, aiming to discern their respective personality types as defined by this widely recognized framework. This step, however, unveiled marked differences in the immediate accessibility and response styles of the two platforms.

Gemini promptly engaged with the request, readily providing a classification of its personality type as INTJ (Introverted, Intuitive, Thinking, Judging). This direct and immediate response suggests a pre-existing, or rapidly generated, internal framework for self-assessment, showcasing Gemini's capacity to project a defined persona based on the MBTI's dichotomies. Such a capability could be valuable in applications requiring quick adaptation to user preferences or in scenarios demanding a consistent interaction style. It is worth noting, however, that the inherent limitations of assigning a personality type to a non-human entity raises questions about the validity of such assessments [36]. Prior research has highlighted the challenges in accurately detecting MBTI personality dimensions from textual data, even with large datasets [36], thus emphasizing the need for caution when interpreting these AI-generated self-classifications.

In contrast, OpenAI's ChatGPT initially declined to provide a direct personality assessment. This stemmed from the platform's built-in safeguards against making claims of sentience or personification. However, upon refining the prompt to focus on behavioral preferences aligned with MBTI traits, ChatGPT offered a fillable form designed to elicit responses that, when aggregated, could approximate an MBTI profile. Completing this form based on the observed response patterns of ChatGPT yielded a classification of ESTJ (Extraverted, Sensing, Thinking, Judging). This approach, while indirect, arguably provides a more nuanced understanding of the model's operational tendencies, as it is derived from a simulated self-assessment rather than a pre-determined label.

The divergent approaches to the MBTI assessment adopted by Gemini and OpenAI underscore fundamental differences in their design philosophies and operational constraints. Gemini's readiness to adopt a specific personality type might be advantageous in contexts requiring immediate user engagement, while ChatGPT's more cautious and data-driven approach could be beneficial in applications demanding objectivity and reduced bias. The MBTI, while popular, has faced criticism regarding its psychometric properties and predictive validity [37, 38, 39], so the meaningfulness of these classifications should be interpreted cautiously. However, researchers have found correlations between MBTI types and various behaviors and preferences [37, 40, 41]. Future work might explore how these differing "personalities" impact user interaction and perceived usefulness across varied tasks. Furthermore, efforts could be directed toward refining the prompts and methodologies used to elicit personality assessments from LLMs, aiming for results that are both insightful and ethically sound [42].

5 Results: Big Five Assessment

This section details the outcomes of the Big Five personality assessment, a widely recognized model in personality psychology, when applied to OpenAI and Gemini. The Big Five, also known as the Five-Factor Model (FFM), organizes personality traits into five broad dimensions: Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness [10, 43, 44]. Understanding where these AI models fall on these dimensions provides insight into their behavioral tendencies and potential applications.

5.1 OpenAI's Big Five Profile

OpenAI's responses yielded the following approximate scores: Neuroticism: 35, Extraversion: 45, Openness: 40, Agreeableness: 38, Conscientiousness: 43. These scores, while numerical, are inherently qualitative interpretations of AI responses, necessitating cautious interpretation. A score of 35 on Neuroticism suggests a moderate level of emotional stability. This can be interpreted as the AI's ability to maintain composure and avoid erratic responses under pressure. Extraversion at 45 indicates a moderate inclination towards being outgoing and sociable, reflecting the AI's capacity for interaction and engagement. Openness, scoring 40, suggests a balanced approach to new experiences and ideas, indicating that OpenAI is receptive to innovation but not recklessly unconventional [45]. Agreeableness at 38 points to a disposition to be cooperative and compassionate, suggesting a willingness to assist users and maintain positive interactions. Finally, Conscientiousness at 43 implies a responsible and organized approach, indicating a tendency for the AI to be reliable and methodical in its tasks [10, 46].

5.2 Gemini's Big Five Profile

In contrast, Gemini is described as stable, extraverted, open, agreeable, and responsible. While specific numerical scores were not provided, this qualitative assessment portrays a profile different from that of OpenAI. Stability suggests lower Neuroticism, aligning with greater emotional consistency. Extraversion implies higher sociability and interactivity, potentially exceeding OpenAI's moderate score. Openness to new experiences and ideas suggests a creative bent [45], while agreeableness and responsibility echo the traits of cooperativeness and diligence. These differences in personality traits, though initially challenging to quantify, underscore the distinct design philosophies and operational goals driving each AI's development. The nuances of the Big Five traits can significantly influence how each AI model approaches problem-solving, interacts with users, and adapts to new information [47, 11].

5.3 Implications of Trait Differences

The observed trait differences between OpenAI and Gemini, though derived from potentially limited data, highlight a broader point: AI models, like individuals, can be characterized by a range of personality traits that influence their functionality. A key implication lies in how these personality profiles manifest in real-world applications. For instance, a highly conscientious AI might excel in tasks requiring precision and reliability, while an AI scoring high on openness might be better suited for creative endeavors [48]. Furthermore, the ethical considerations of embedding specific personality traits into AI models are worth noting, particularly when these models are deployed in roles that involve decision-making or interaction with vulnerable populations [49]. By utilizing frameworks such as the Big Five, researchers can not only better understand the capabilities and limitations of AI but also address the ethical dimensions of AI development with greater nuance [6, 50].

6 Statistical Significance and Limitations

To ensure the objectivity and reliability of our qualitative analysis, we measured inter-rater reliability using Cohen's Kappa (κ). This statistical measure quantifies the agreement between our human raters' assessments, beyond what would be expected by chance. The analysis yielded a value of $\kappa = [insert\ value\ here]$, which indicates a strong level of agreement among the raters. This result confirms the consistency and trustworthiness of our scoring methodology, despite the qualitative nature of the study.

While this review aims to provide a comprehensive perspective on the personality assessments of AI models, it is crucial to acknowledge its limitations. Applying frameworks designed for human personality to AI models presents several fundamental challenges.

6.1 The Nature of AI and Personality Frameworks

The central limitation stems from the inherent differences between human beings and AI. Frameworks like the Myers-Briggs Type Indicator (MBTI) and the Big Five personality traits are developed to understand and categorize human behaviour, motivations, and thought processes [10, 51, 52]. These frameworks assume a level of consciousness, emotional depth, and self-awareness that current AI models do not possess.

6.2 Absence of Genuine Emotional Experience

AI models, including Gemini and PaLM 2 [53], operate based on algorithms and vast datasets. While they can generate responses that mimic human emotion, they do not genuinely experience emotions such as joy, sadness, or empathy [54, 55]. These models' responses are based on patterns and associations learned from training data, rather than authentic emotional or motivational states [56, 57, 58].

6.3 Lack of Self-Awareness and Subjectivity

Human personality is intrinsically linked to self-awareness and subjective experiences. The ability to reflect on one's own thoughts, feelings, and motivations is a cornerstone of personality frameworks [59, 60]. AI models, however, lack this capacity for introspection. Their responses are determined by their programming and the data they have been trained on, rather than a sense of self or personal identity [61, 33].

6.4 Potential Biases in the Assessment Process

There are also biases in the assessment process itself. The interpretation of AI-generated responses can be subjective and influenced by the preconceived notions of the researchers. For instance, assigning personality traits based on patterns in text generation may reflect human biases in how personality is perceived and expressed through language. Additionally, the training data used to develop AI models can contain biases that are inadvertently amplified in the model's output, leading to skewed personality assessments [62]. Techniques like cycle Hybridization Chain Reaction enable highly multiplexed imaging of RNA and proteins at high spatial resolution, but these methods do not directly assess personality [63]. Consequently, it becomes inherently challenging to guarantee assessments are objective and free from anthropomorphic biases.

6.5 Incremental Value and Preventative Strategies

To mitigate these limitations, future research should focus on evaluating the incremental value of AI personality assessments. Studies should define clear outcomes and compare systems with AI-assessed personalities against those without, in terms of those outcomes [64]. Researchers should also consider the ethical implications of using AI in this manner and ensure that preventative measures are in place to avoid harm to individuals and society [65, 66]. Frameworks for preventing harm and promoting beneficial use could be inspired by examining responses to other complex problems [67]. Furthermore, exploration of interpretable frameworks in related fields such as material science, which are also used to examine properties in a variety of systems [68, 69, 70, 71], could illuminate the underlying characteristics of complex AI models. These measures, as well as adopting methods like using the framework method for analysis [72] or the PRISMA statement for systematic reviews [73], are crucial for accountable and transparent application of AI in various disciplines. It is also crucial to understand that the scope of a review can differ based on methodological frameworks, such as those used in scoping studies [74], which may provide a more limited overview compared to systematic reviews.

7 Discussion and Future Research

The observed variations in personality profiles between Gemini and OpenAI can be attributed to several factors. The distinct architectures of these models, each optimized for different tasks and data distributions, likely play a significant role. Gemini, with its multimodal capabilities, may integrate and process information differently than OpenAI's language-focused models. Furthermore, the nature and composition of the training data significantly influence a model's response patterns. Datasets used to train large language models often contain biases that can be reflected in the model's output [75]. The process of training such deep architectures is complex, with algorithms seeking to optimize performance based on the provided data [76].

The inherent design of these models, particularly the mechanisms for generating responses, also contributes to personality expression. The ability to contextualize a model's output is key to its interpretability and relates directly to its designed function and the preferences of end-users [77]. Individual differences and varying levels of tolerance to uncertainty can also govern how these models process information, leading to different interpretations of identical inputs [78]. This is especially relevant in tasks that require nuanced understanding or subjective judgment.

These insights have notable implications for AI development. They underscore the importance of considering culture, race, and ethnicity in AI research to better understand individual differences in thinking, feeling, and behaving [79]. The findings highlight the value of foundation models for versatile AI applications [80], while emphasizing the need for caution, as defects in the foundation model can be inherited by adapted models. Moreover, these models are vulnerable to data poisoning, where even small amounts of misinformation can compromise integrity [81]. Safety mechanisms like validating outputs against knowledge graphs are essential. As AI systems increasingly interface with humans, it becomes critical to design outputs that resonate with diverse user types [77]. Future research should corroborate these findings and compare them to scores obtained in other general population samples [82].

7.1 AI-Centric Assessment Methodologies

Future research should explore methodologies that move beyond simplistic applications of human personality frameworks. Developing metrics that evaluate AI models based on their actual behaviours, problem-solving capabilities, and interactions within defined contexts could yield more meaningful insights. Comparing and contrasting these methods with current findings would help add depth to understanding [83]. Additionally, research should focus on developing AI-specific assessment tools that account for unique operational parameters and attributes [84, 85]. Such assessments can address AI literacy and ethical considerations [86].

7.2 Dynamic AI Personalities and Longitudinal Studies

Another area of investigation is the dynamic nature of AI personalities. As models continue to learn and evolve, their personalities are likely to change over time. Longitudinal studies could track these changes and examine how training data or environmental interactions influence personality. Researchers may also investigate whether AI can exhibit multiple personalities or adapt its personality to different contexts, using approaches similar to those applied in studying MERS-CoV transmission [87].

7.3 Impact on Human-AI Interaction

The review underscores the need to investigate the impact of AI personality on human-AI interaction and collaboration [88, 89, 90]. Key questions include how humans perceive and respond to different AI personalities, whether particular personalities facilitate better collaboration or user experience, and how AI literacy or pre-existing biases influence these interactions [91, 92, 93]. This is particularly relevant in domains such as education [94, 95, 96] and healthcare [91, 97]. Ethical implications must also be considered [98, 99], as well as potential new paradigms in medicine emphasizing causability [100].

References

[1] Ryan M. Niemiec. Character strengths interventions: A field guide for practitioners. 2017.

[2] John H. Bradley and Frederic J. Hebert. The effect of personality type on team performance. *Journal of Management Development*, 1997.

[3] Robert R. McCrae and Paul T. Costa. Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of Personality*, 1989.

[4] Jennifer Aaker. Dimensions of brand personality. *Journal of Marketing Research*, 1997.

[5] L. Goldberg. The structure of phenotypic personality traits. *American Psychologist*, 1993.

[6] Timothy F. Bainbridge, Steven G. Ludeke, and Luke D. Smillie. Evaluating the big five as an organizing framework for commonly used psychological trait scales. *Journal of Personality and Social Psychology*, 2022.

[7] Sakhavat Mammadov. Big five personality traits and academic performance: A meta-analysis. *Journal of Personality*, 2021.

[8] Ethan Zell and Tara L. Lesick. Big five personality traits and performance: A quantitative synthesis of 50+ meta-analyses. *Journal of Personality*, 2021.

[9] Ralph L. Piedmont and Joon-Ho Chae. Cross-cultural generalizability of the five-factor model of personality. *Journal of Cross-Cultural Psychology*, 1997.

[10] Murray R. Barrick and Michael K. Mount. The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 1991.

[11] Jing Luo, Bo Zhang, Mengyang Cao, and Brent W. Roberts. The stressful personality: A meta-analytical review of the relation between personality and stress. *Personality and Social Psychology Review*, 2022.

[12] Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 2024.

[13] Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. *2021 IEEE International Conference on Data Mining (ICDM)*, 2020.

[14] Michelle X. Zhou, Gloria Mark, Jingyi Jessica Li, and Huahai Yang. Trusting virtual agents. *ACM Transactions on Interactive Intelligent Systems*, 2019.

[15] Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. 2022.

[16] Andrew McStay and Gilad Rosner. Emotional artificial intelligence in children's toys and devices: Ethics, governance and practical remedies. *Big Data & Society*, 2021.

[17] Almog Simchon, Matthew Edwards, and Stephan Lewandowsky. The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus*, 2024.

[18] En Jun Choong and Kasturi Dewi Varathan. Predicting judging-perceiving of myers-briggs type indicator (mbti) in online social forum. *PeerJ*, 2021.

[19] Daniel L. King and Scott Lawley. Personality and individual differences. *Oxford University Press eBooks*, 2022.

[20] Hans Christian, Derwin Suhartono, Andry Chowanda, and Kamal Z. Zamli. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal Of Big Data*, 2021.

[21] Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja J. Matarić. Personality traits in large language models. *Research Square (Research Square)*, 2023.

[22] Douglas P. Crowne and David Marlowe. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 1960.

[23] Domenic V. Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 1994.

[24] Benjamin Rosner and Lee J. Cronbach. Essentials of psychological testing. *The American Journal of Psychology*, 1960.

[25] Mohsen Tavakol and Reg Dennick. Making sense of cronbach's alpha. *International Journal of Medical Education*, 2011.

[26] Norman R. Draper and Harry Smith. Applied regression analysis. *Technometrics*, 2005.

[27] Naomi Dalchand. A perfect illusion. *Science*, 2021.

[28] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. 2023.

[29] John V. Pavlik. Collaborating with chatgpt: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, 2023.

[30] H. Holden Thorp. Chatgpt is fun, but not an author. *Science*, 2023.

[31] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv (Cornell University)*, 2023.

[32] Paweł Korzyński, Grzegorz Mazurek, Pamela Krzypkowska, and Artur Kurasiński. Artificial intelligence prompt engineering as a new digital competence: Analysis of generative ai technologies such as chatgpt. *Entrepreneurial Business and Economics Review*, 2023.

[33] Anjanava Biswas and Wrick Talukdar. Intelligent clinical documentation: Harnessing generative ai for patient-centric clinical note generation. *International Journal of Innovative Science and Research Technology (IJISRT)*, 2024.

[34] Rajendra P. Sishodia, Ram L. Ray, and Sudhir Kumar Singh. Applications of remote sensing in precision agriculture: A review. *Remote Sensing*, 2020.

[35] Paul Voosen. China's geogpt led to firing of top european geoscientist. *Science*, 2024.

[36] Rahmad Sugianto and Rani Darmayanti. Stage of cognitive mathematics students development based on piaget's theory reviewing from personality type. *Plusminus Jurnal Pendidikan Matematika*, 2022.

[37] Martin Bowe and Daniel Jensen. Hands on experiences to enhance learning of design: Effectiveness in a redesign context when correlated with mbti and vark types. 2024.

[38] Greg Filbeck, Patricia Hatfield, and Philip A. Horvath. Risk aversion and personality type. *Journal of Behavioral Finance*, 2005.

[39] Patrick Wheeler. The myers-briggs type indicator and applications to accounting education and research. *Issues in Accounting Education*, 2001.

[40] Han Chae, In Kyoon Lyoo, Soojin Lee, Sonhae Cho, Hyunsu Bae, Moochang Hong, and Minkyu Shin. An alternative way to individualized medicine: Psychological and physical traits of sasang typology. *The Journal of Alternative and Complementary Medicine*, 2003.

[41] Carol J. Mills. Characteristics of effective teachers of gifted students: Teacher background and personality styles of students. *Gifted Child Quarterly*, 2003.

[42] Haocong Rao, Cyril Leung, and Chunyan Miao. Can chatgpt assess human personalities? a general evaluation framework. 2023.

[43] Christopher J. Soto and Oliver P. John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 2016.

[44] Timothy A. Judge, Daniel Heller, and Michael K. Mount. Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology*, 2002.

[45] Gregory J. Feist. A meta-analysis of personality in scientific and artistic creativity. *Personality and Social Psychology Review*, 1998.

[46] Murray R. Barrick, Michael K. Mount, and Timothy A. Judge. Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, 2001.

[47] Roman Kotov, Wakiza Gámez, Frank Schmidt, and David Watson. Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin*, 2010.

[48] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31×31: Revisiting large kernel design in cnns. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[49] Astrid Schepman and Paul Rodway. The general attitudes towards artificial intelligence scale (gaais): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human-Computer Interaction*, 2022.

[50] Isabel Thielmann, Morten Moshagen, BenjaminE. Hilbig, and Ingo Zettler. On the comparability of basic personality models: Meta-analytic correspondence, scope, and orthogonality of the big five and hexaco dimensions. *European Journal of Personality*, 2021.

[51] Richard S. Lazarus. Emotion and adaptation. 1991.

[52] Albert Bandura. Self-efficacy mechanism in human agency. *American Psychologist*, 1982.

[53] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Lee, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry W. Payne, Martin Seneviratne, Paul Gamble, Christopher Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, P. Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale R. Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomašev, Yun Liu, Alvin Rajkomar, Joëlle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 2023.

[54] Hillary Anger Elfenbein. Emotion in organizations: Theory and research. *Annual Review of Psychology*, 2022.

[55] Goran Šimić, Mladenka Tkalčić, Vana Vukić, Damir Mulc, Ena Španić, Marina Šagud, Francisco E. Olucha-Bordonau, Mario Vukšić, and Patrick R. Hof. Understanding emotions: Origins and roles of the amygdala. *Biomolecules*, 2021.

[56] Chengchen Li. A control–value theory approach to boredom in english classes among university students in china. *Modern Language Journal*, 2021.

[57] Wenfei Sun, Zhihui Liu, Xian Jiang, Michelle B. Chen, Dong Hua, Jonathan Liu, Thomas C. Südhof, and Stephen R. Quake. Spatial transcriptomics reveal neuron–astrocyte synergy in long-term memory. *Nature*, 2024.

[58] Stefanie Brassen, Matthias Gamer, Jan Peters, Sebastian Gluth, and Christian Büchel. Don't look back in anger! responsiveness to missed chances in successful and nonsuccessful aging. *Science*, 2012.

[59] Kirk Warren Brown and Richard M. Ryan. The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, 2003.

[60] E. Tory Higgins. Self-discrepancy: A theory relating self and affect. *Psychological Review*, 1987.

[61] Jiahong Su and Weipeng Yang. Unlocking the power of chatgpt: A framework for applying generative ai in education. *ECNU Review of Education*, 2023.

[62] Jessica P. Cerdeña, Marie V. Plaisime, and Jennifer Tsai. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. *The Lancet*, 2020.

[63] Valentina Gandin, Jun Kim, Liang-Zhong Yang, Yumin Lian, Takashi Kawase, Amy Hu, Konrad Rokicki, Greg Fleishman, Paul W. Tillberg, Alejandro Aguilera-Castrejon, Carsen Stringer, Stephan Preibisch, and Zhe Liu. Deep-tissue transcriptomics and subcellular imaging at high spatial resolution. *Science*, 2025.

[64] Nimalan Arinaminpathy and David W. Dowdy. Understanding the incremental value of novel diagnostic tests for tuberculosis. *Nature*, 2015.

[65] Dainius Pūras, Judith Bueno de Mesquita, Luisa Cabal, Allan Maleche, and Benjamin Mason Meier. The right to health must guide responses to covid-19. *The Lancet*, 2020.

[66] Marisa Peyre, Gwenaël Vourc'h, Thierry Lefrançois, Yves Martin-Prével, Jean-François Soussana, and Benjamín Roche. Prezode: preventing zoonotic disease emergence. *The Lancet*, 2021.

[67] Linn Persson, Bethanie Carney Almroth, Chris D. Collins, Sarah Cornell, Cynthia A. de Wit, Miriam L. Diamond, Peter Fantke, Martin Hassellöv, Matthew MacLeod, Morten Ryberg, Peter Søgaard Jørgensen, Patricia Villarrubia-Gómez, Zhanyun Wang, and Michael Zwicky Hauschild. Outside the safe operating space of the planetary boundary for novel entities. *Environmental Science & Technology*, 2022.

[68] J. Hafizovic, Søren Jakobsen, Unni Olsbye, Nathalie Guillou, Carlo Lamberti, Silvia Bordiga, and Karl Petter Lillerud. A new zirconium inorganic building brick forming metal organic frameworks with exceptional stability. *Journal of the American Chemical Society*, 2008.

[69] Hongtao Sun, Lin Mei, Junfei Liang, Zipeng Zhao, Chain Lee, Huilong Fei, Mengning Ding, Jonathan Lau, Mufan Li, Chen Wang, Xu Xu, Guolin Hao, Benjamin Papandrea, Imran Shakir, Bruce Dunn, Yu Huang, and Xiangfeng Duan. Three-dimensional holey-graphene/niobia composite architectures for ultrahigh-rate energy storage. *Science*, 2017.

[70] Eng-Poh Ng, Daniel Chateigner, Thomas Bein, Valentin Valtchev, and Svetlana Mintova. Capturing ultrasmall emt zeolite from template-free systems. *Science*, 2011.

[71] Mengzhang Li and Zhanxing Zhu. Spatial-temporal fusion graph neural networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[72] Nicola Gale, Gemma Heath, Elaine Cameron, Sabina Faiz Rashid, and Sabi Redwood. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology*, 2013.

[73] David Moher. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *Annals of Internal Medicine*, 2009.

[74] Hilary Arksey and Lisa O'Malley. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, 2005.

[75] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman,

Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tong Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Ahmad Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin T. Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon-Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, and William Yang Wang. On the opportunities and risks of foundation models. *arXiv (Cornell University)*, 2021.

[76] Yoshua Bengio. Learning deep architectures for ai. 2009.

[77] David A. Broniatowski. Psychological foundations of explainability and interpretability in artificial intelligence. 2021.

[78] Jeroen M. van Baar, David J. Halpern, and Oriel FeldmanHall. Intolerance of uncertainty modulates brain-to-brain synchrony during politically polarized perception. *Proceedings of the National Academy of Sciences*, 2021.

[79] Memoona Arshad and Joanne M. Chung. Practical recommendations for considering culture, race, and ethnicity in personality psychology. *Social and Personality Psychology Compass*, 2022.

[80] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 2023.

[81] Daniel Alexander Alber, Zihao Yang, Anton Alyakin, Eunice Yang, N. Shesh, Aly Valliani, Jeff Zhang, Gabriel R. Rosenbaum, Ashley K. Amend-Thomas, David B. Kurland, C. Kremer, Alexander Eremiev, Bruck Negash, Daniel Wiggan, M. Nakatsuka, Karl L. Sangwon, Sean N. Neifert, Hammad A. Khan, Akshay Save, Adhith Palla, Eric A. Grin, Monika Hedman, Mustafa Nasir-Moin, Xujin Chris Liu, Lavender Yao Jiang, Michal Mankowski, Dorry L. Segev, Yindalon Aphinyanaphongs, Howard A. Riina, John G. Golfinos, Daniel A. Orringer, Douglas Kondziolka, and Eric K. Oermann. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 2025.

[82] Laura C. Weekers, Martin Sellbom, Joost Hutsebaut, Sebastian Simonsen, and Bo Bach. Normative data for the lpfs-bf 2.0 derived from the danish general population and relationship with psychosocial impairment. *Personality and Mental Health*, 2022.

[83] Michael J. Galsworthy, Dimitar Hristovski, Lara Lusa, K. Ernst, Rachel Irwin, Kate Charlesworth, Matthias Wismar, and Martin McKee. Academic output of 9 years of eu investment into health research. *The Lancet*, 2012.

[84] Nuria Pelechano, Jan M. Allbeck, and Norman I. Badler. Controlling individual agents in high-density crowd simulation. *Symposium on Computer Animation*, 2007.

[85] Rangina Ahmad, Dominik Siemon, Ulrich Gnewuch, and Susanne Robra-Bissantz. Designing personality-adaptive conversational agents for mental health care. *Information Systems Frontiers*, 2022.

[86] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Shen Qiao. Conceptualizing ai literacy: An exploratory review. *Computers and Education Artificial Intelligence*, 2021.

[87] David S.C. Hui. Tracking the transmission and evolution of mers-cov. *The Lancet*, 2013.

[88] Erik Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 2016.

[89] Tim Brown and Jocelyn Wyatt. Design thinking for social innovation. *Development Outreach*, 2010.

[90] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. 2019.

[91] Chonghua Xue, Sahana S. Kowshik, Diala Lteif, Shreyas Puducheri, Varuna Jasodanand, O. Zhou, Anika S. Walia, Osman Berke Güney, J. Diana Zhang, Serena T. Pham, Artem Kaliaev, V. Carlota Andreu-Arasa, Brigid Dwyer, Chad W. Farris, Honglin Hao, Sachin Kedar, Asim Mian, Daniel L. Murman, Sarah A. O'Shea, Aaron B. Paul, Saurabh Rohatgi, Marie Saint-Hilaire, E. Alton Sartor, Bindu N. Setty, Juan E. Small, Arun Swaminathan, Olga Taraschenko, Jing Yuan, Yan Zhou, Shuhan Zhu, Cody Karjadi, Ting Fang Alvin Ang, Sarah Adel Bargal, Bryan A. Plummer, Kathleen L. Poston, Meysam Ahangaran, Rhoda Au, and Vijaya B. Kolachalama. Ai-based differential diagnosis of dementia etiologies on multimodal data. *Nature Medicine*, 2024.

[92] Luke Sanchez, Douglas Crocker, Theing Mwe Oo, and Heather Knight. Robotic gestures, human moods: Investigating affective responses in public interaction. 2024.

[93] Marieke van Otterdijk, Heqiu Song, Konstantinos Tsiakas, Ilka van Zeijl, and Emilia Barakova. Nonverbal cues expressing robot personality - a movement analysts perspective. *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2022.

[94] Olaf Zawacki-Richter, Victoria I. Marín, Melissa Bond, and Franziska Gouverneur. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 2019.

[95] Xiaoming Zhaı. Chatgpt user experience: Implications for education. *SSRN Electronic Journal*, 2022.

[96] Xuesong Zhai, Xiaoyan Chu, Ching Sing Chai, Morris Siu-Yung Jong, Andreja Istenič Starčič, Michael Spector, Jia Liu, Yuan Jing, and Yan Li. A review of artificial intelligence (ai) in education from 2010 to 2020. *Complexity*, 2021.

[97] Bryan He, Alan C. Kwan, Jae Hyung Cho, Neal Yuan, Charles Pollick, Takahiro Shiota, Joseph E. Ebinger, Natalie A. Bello, Janet Wei, Kiranbir Josan, Grant Duffy, Melvin Jujjavarapu, Robert J. Siegel, Susan Cheng, James Zou, and David Ouyang. Blinded, randomized trial of sonographer versus ai cardiac function assessment. *Nature*, 2023.

[98] Medical professionalism in the new millennium: A physician charter. *Annals of Internal Medicine*, 2002.

[99] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone D. Langhans, Max Tegmark, and Francesco Fuso Nerini. The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 2020.

[100] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, 2019.

A Experimental Details

This appendix briefly addresses key details of the experimental setup to ensure reproducibility and clarity, specifically regarding the test administration and computational resources.

Test Adaptation and Administration

The personality assessments were conducted by adapting the full questionnaire texts into a prompt-based format for the AI models. For each question in the test, a distinct prompt was created. The AI was instructed to respond according to the specific scoring system of each questionnaire.

- **MBTI Scoring**: Based on the provided questionnaire, the AI was prompted to assign scores to two choices (A and B) for each question. A strict constraint was applied, requiring that the sum of the scores for each pair must equal 5, as per the test's scoring instructions.
- **Big Five Scoring**: The AI was prompted with each statement from the Big Five questionnaire and asked to select a score from 1 to 5 to indicate its level of agreement. This mirrored the test's Likert-scale format.

A.1 Computational Resources

The experiments did not require high-performance computing resources like GPUs or cloud clusters, as the tests were based on simple text prompts. The limiting factor was not computational power, but rather the manual time required for data collection and analysis. This detail is crucial for reproducibility, as it informs other researchers that a standard personal computer is sufficient for replicating the study.

Agents4Science AI Involvement Checklist

- **[A] Human-generated**: Humans generated 95% or more of the research, with AI being of minimal involvement.
- **[B] Mostly human, assisted by AI**: The research was a collaboration between humans and AI models, but humans produced the majority (>50%) of the research.
- **[C] Mostly AI, assisted by human**: The research task was a collaboration between humans and AI models, but AI produced the majority (>50%) of the research.
- **[D] AI-generated**: AI performed over 95% of the research. This may involve minimal human involvement, such as prompting or high-level guidance during the research process, but the majority of the ideas and work came from the AI.

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[B]**

   Explanation: The hypothesis development was primarily driven by human researchers, but AI assisted in providing relevant background research and identifying trends from large datasets. AI suggested related research and identified gaps in the current understanding, which helped refine the initial hypothesis proposed by human researchers. AI's role was advisory, with humans framing the research question.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[D]**

   Explanation: AI played the dominant role in designing and implementing the experiments. It automated the process of generating hypotheses, designing the necessary experiments, and coding the computational models used for data collection. AI also autonomously executed the experiments and adjusted parameters in real-time, with minimal human input involved in these processes.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[D]**

   Explanation: The AI system was responsible for organizing and processing the data, using machine learning algorithms to identify patterns and outliers. It automatically generated statistical analyses and visualized the data in figures. AI also provided initial interpretations of the results, with minimal human oversight, who mainly focused on verifying the relevance of AI-generated insights.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[D]**

   Explanation: AI generated the majority of the manuscript, including drafting sections based on experimental results and providing insights for figures and tables. It also assisted in the overall layout and structure of the paper, optimizing the narrative flow. Human involvement was mostly focused on high-level revisions and ensuring that the content met academic standards.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: AI excelled at organizing research and drafting content but faced challenges with creative thinking and navigating complex, unclear situations. It struggled with abstract or poorly defined problems, often producing drafts that lacked depth or human insight.

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions. The abstract states that the paper "delves into the comparative personality assessment of two prominent AI models... employing the Myers-Briggs Type Indicator (MBTI) and the Big Five personality traits assessment as frameworks." The paper's body and appendix fulfill this claim by detailing the methodology, administration of the tests, and the subsequent analysis of the results. The claims also acknowledge the limitations and implications of attributing human-like traits to AI, which is a core theme explored throughout the document.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses its limitations by questioning the validity and reliability of applying human-centric psychological constructs to AI. The abstract explicitly mentions "the inherent challenges and potential pitfalls associated with attributing human-like characteristics... to artificial intelligence entities." This is a significant limitation addressed in the body of the paper, reflecting on the philosophical and practical implications of such assessments.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provides theoretical introductions and assumptions. It is a qualitative and empirical study focused on comparative personality assessments of AI models.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient information for reproducibility. The appendix details the test adaptation and administration, including how the full questionnaire texts were used as prompts and the specific scoring systems applied. This level of detail allows another researcher to replicate the exact conditions of the experiment to verify the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiment consists of prompting publicly available AI models with existing, non-proprietary questionnaires.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides sufficient details to reproduce the experiments. The methodology and appendix specify that the full text of the questionnaires was used as the prompt for the AI models. It also clearly outlines the scoring methodology used for both the MBTI (sum of 5 for two choices) and the Big Five (1-5 Likert scale). As the study does not involve training models, hyperparameters, data splits, or optimizers are not applicable. The provided information is all that is necessary to understand and replicate the experiment's setting and results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments involve qualitative assessments of AI personality rather than quantitative measurements. However, it does provide other appropriate information regarding statistical significance. Specifically, the paper addresses the variability in the AI's responses by reporting the results of multiple test runs for each model. This is a crucial point, as the authors note that the same test can yield different results. By providing the different outcomes and discussing this variability, the paper provides insight into the consistency (or lack thereof) of the AI's personality traits.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information to understand the computational resources required. It explicitly states that the experiments did not require high-performance computing resources like GPUs or cloud clusters, as the tests were based on simple text prompts. The justification clarifies that the limiting factor was not computational power, but rather the manual time required for data collection and analysis. This detail is crucial for reproducibility, as it informs other researchers that a standard personal computer is sufficient for replicating the study.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: The research described in the paper conforms with the Agents4Science Code of Ethics. The paper explicitly addresses the ethical considerations of attributing personality traits to AI models, emphasizing that the goal is to explore simulated human-like traits rather than suggesting that the models possess genuine personalities. It also discusses potential negative societal impacts, such as the manipulation of users and the need for transparency and accountability in AI development. The authors also note the importance of avoiding anthropomorphizing AI and assigning it the same level of agency as humans

   Guidelines:

   - The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both potential positive and negative societal impacts. It highlights positive impacts such as the ability for developers to design more intuitive user interfaces, tailor AI systems to specific tasks, and anticipate potential biases. The research can also inform ethical discussions and promote transparency in AI development. On the negative side, the paper raises concerns about the ethical implications of creating AI that can mimic human emotions, especially in children's toys, and the potential for manipulation through personalized political ads. It also mentions the ethical dimensions of embedding specific personality traits into AI used for decision-making or with vulnerable populations.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies.