OPEN ACCESS

KING'S College LONDON

King's Research Portal

DOI:
10.1145/3640794.3665887

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Sun, G., Zhan, N., & Such, J. (2024). Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents. In *Proceedings of the 6th Conference on ACM Conversational User Interfaces, CUI 2024* Article 35 (Proceedings of the 6th International Conference on Conversational User Interfaces, CUI 2024). https://doi.org/10.1145/3640794.3665887

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 14. Jan. 2025

Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents

Guangzhi Sun*
University of Cambridge
Cambridge, United Kingdom
gs534@cam.ac.uk

Xiao Zhan*
King's College London
London, United Kingdom
xiao.zhan@kcl.ac.uk

Jose Such
King's College London
London, United Kingdom
& VRAIN, Universitat Politecnica de Valencia, Spain
jose.such@kcl.ac.uk

ABSTRACT

The incorporation of Large Language Models (LLMs) such as the GPT series into diverse sectors including healthcare, education, and finance marks a significant evolution in the field of artificial intelligence (AI). The increasing demand for personalised applications motivated the design of conversational agents (CAs) to possess distinct personas. This paper commences by examining the rationale and implications of imbuing CAs with unique personas, smoothly transitioning into a broader discussion of the personalisation and anthropomorphism of CAs based on LLMs in the LLM era.

We delve into the specific applications where the implementation of a persona is not just beneficial but critical for LLM-based CAs. The paper underscores the necessity of a nuanced approach to persona integration, highlighting the potential challenges and ethical dilemmas that may arise. Attention is directed towards the importance of maintaining persona consistency, establishing robust evaluation mechanisms, and ensuring that the persona attributes are effectively complemented by domain-specific knowledge.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; **Usability in security and privacy**; • **Computing methodologies** → *Discourse, dialogue and pragmatics*; **Natural language processing**; • **Human-centered computing** → *HCI theory, concepts and models.*

KEYWORDS

Large language model, persona, personality, conversational agent, ChatGPT, natural language processing

ACM Reference Format:
Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents. In *ACM Conversational User Interfaces 2024 (CUI '24), July 8–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3640794.3665887

*Both authors contributed equally to this research.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CUI '24, July 8–10, 2024, Luxembourg, Luxembourg
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665887

1 WHAT DOES 'PERSONA' MEAN IN THE CONTEXT OF CONVERSATIONAL AGENTS?

In the context of conversational agents (CAs), the concept of *persona* represents the essence or 'soul' of these agents. Persona encapsulates the distinct tone, voice, and personality that characterizes a CA, transforming mechanical interactions into engaging, human-like conversations [18, 46]. Commonly, these attributes of persona can consist any type of information that intend to capture personal characteristics about an individual [28], and are relatively static (race), and slowly change over time (age), or temporary (emotional status) [26, 54].

Before delving deeper into the discussion of personas in CAs, it's important to distinguish this concept from the idea of 'personality' that has been explored in prior research [24, 27, 36, 38]. While personality traits, such as being "friendly" or "smart," or frameworks like the Myers-Briggs Type Indicator (MBTI) [4], might define certain characteristics shared by groups of individuals, a persona in CAs represents a more complex and consistent identity [36, 55]. This persona transcends mere personality traits, serving as an external manifestation of a character's unique identity. For instance, when a CA is designed with the persona of a specific character, say, Sherlock Holmes, it consistently embodies the unique attributes and behaviors of that character throughout interactions. This specificity differs significantly from assigning generic traits like 'bravery' and 'smartness' to a CA. In the latter case, the CA might alternate between different characters who share these traits, such as both Sherlock Holmes and Hermione Granger, depending on the context of the interaction. Thus, the persona of a CA is a more nuanced and stable layer that defines its interaction style and character representation.

1.1 Persona in CAs in pre-LLM era

Recent research in the field of CAs has focused extensively on enhancing the capabilities of chatbots, aiming to imbue them with more human-like characteristics. This initiative is driven by the goal to significantly boost user engagement, among other benefits. The development of a persona for CAs such as chatbots has emerged as a key strategy in this domain. The introduction of these personas is a testament to the evolving sophistication of chatbot technology, reflecting a deeper understanding of human-chatbot interaction dynamics [16].

Two main avenues of exploration have emerged: the technical research stream pushes the boundaries of what is technically

possible [8, 26, 27, 44, 46, 49, 57], the social research stream ensures that these advancements are grounded in a thorough understanding of user needs, preferences, and the broader societal context [3, 16, 37, 56].

1.1.1 Technical research. Previous studies have proposed various methods for embedding personas into traditional chatbots[1]. The categories used are broad — for a comprehensive summary of the model and a survey see [46]. The more widely known examples are that neural models of conversation generation provided a simple mechanism for incorporating personas as embeddings [26, 44, 49]. More recently, Liao and He created personas for conversational agents that had distinct gender and race to understand user preferences [27]. As one example of a project that is guided by user data, persona XiaoIce was designed based on a large scale analysis of human conversations [57]. In doing so, the designers found that the majority of "desired" users are young and female. Hence, they designed XiaoIce's persona around an "18-year-old girl" [57]. As another example, Danielescu and Christian [8] designed personas for a conversational coaching system where they involved customers by interviewing them and brainstorming with them, finding that their preferences may vary based on their culture and region.

1.1.2 Social research. The academic community has consistently maintained a positive attitude towards endowing chatbots with personas. Incorporating a distinct persona in CAs significantly influences the development of a robust relationship in human-agent interactions. It has been demonstrated that a well-crafted persona can significantly enhance the capacity of CAs to engage in empathetic conversations [37, 56]. This is mirrored from empirical research, such as that by Zhong et al. [56], has established the role of persona in fostering empathy within human conversations from the psychological perspective. Moreover, the positive contribution of persona is recognised in specific areas such as healthcare where CAs assume varied roles. For instance, Bickmore et al. [3] found that an empathetic persona in an agent is effective for managing mental health, whereas an agent with a subtle persona guiding exercise can enhance commitment to behavior change. Similarly, preliminary research conducted in [16] indicates that chatbots embodying roles like doctors, in comparison to generic bots, achieve higher user acceptance, intimacy, and trust in healthcare-related interactions.

2 REALITY OR ASPIRATION?

Large Language Model (LLM)-based CAs, exemplified by systems like ChatGPT[2], are rapidly being integrated into various critical sectors, underscoring their growing significance in practical applications. These include, but are not limited to, healthcare [6, 21, 47], education [19, 31, 52], and finance [22, 23, 51], among others.

These LLM-based CAs, which are originally developed for general-purpose applications, do not prioritize the establishment of a distinct persona during their design phase. For example, as illustrated in Figure 1, ChatGPT, a typical instance of such systems, is structured to function without a predefined persona, focusing instead on

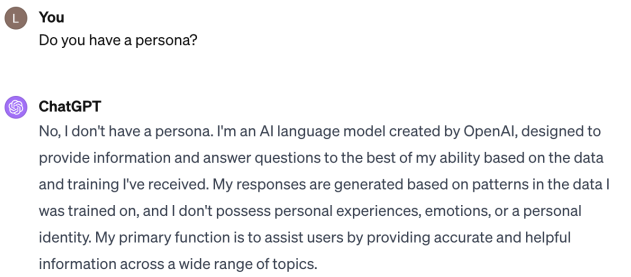

Figure 1: A screenshot of a dialogue with GPT-4-0125-preview. This suggests that GPT-4 does not embody a specific persona. However, this conclusion is based on the model's output, which may not fully align with the designers' intentions.

delivering information and interaction capabilities that are broadly applicable across various contexts and user requirements[3].

Nevertheless, the integration of personas in LLM-based CAs should not be viewed as an unattainable goal. Online resources (including blogs [5, 32], and technical reports [50]) already provide guidance on designing specific personas to optimize ChatGPT's effectiveness across various roles, typically achieved by customizing initial conversation prompts to assign a desired persona. Concurrently, numerous empirical studies [2, 7, 12, 17, 20, 34, 35, 58] have examined and demonstrated the practicality of assigning personas to LLM-based CAs. Among them, some promising results indicated that endowing LLM-based CAs with personas leads to satisfactory outcomes. These include the ability to express opinions similar to people from some controes [12], offering useful answers [20], team working [7], and enhancing the overall truthfulness in their responses [58].

However, upon deeper analysis of persona-based CAs, it becomes evident that LLM-based CAs are still far from embodying specific personas at this stage, highlighting a substantial developmental path that lies ahead. For instance, significant performance disparities exist between different GPT versions. Stories from GPT-4 personas are generally more readable, coherent, and believable, while ChatGPT tends to deviate from the provided prompts, failing to adhere strictly to the prescribed personas [17]. In [41], a study was conducted to assess if the prevailing prompt-based approach facilitates LLM-based CAs in delivering consistent and robust responses. Their investigation, which included testing 15 open-source LLMs, ultimately revealed that most models lacked a consistent persona. Furthermore, it's noteworthy that malicious actors sometimes exploit these characteristic, manipulating them to generate toxic responses[10, 59].

3 PERSONA NEEDS IN LLM-BASED CAS

In the current landscape dominated by LLMs, the importance of persona has not diminished, rather, it often takes on an even more critical role. In this section, we will explore various situations and use cases where the persona of a LLM-based CA is particularly crucial.

[1]Unlike our approach that distinctly separates persona from personality, some prior research conflates these concepts without addressing their nuances. Therefore, the summary in this section includes works that focus on 'personality' as well.
[2]https://openai.com/chatgpt

[3]Despite this observation, no official documentation or evidence has been found to indicate that ChatGPT was deliberately designed to incorporate distinct personas.

3.1 Participant Simulation

Hagendorff et al. [15] conducted an evaluation of GPT-3.5 through cognitive response tests and discovered that the error patterns of the language model qualitatively reflect intuitive behaviors akin to those found in humans. Furthermore, it often fails in similar reasoning tasks as humans do [9]. These findings underscore the significant potential of LLMs in capturing aspects of human behavior. Based on these findings, LLMs are increasingly being considered and used to simulator human beings with different personas. Recent studies [1, 2, 35] have provided substantial evidence that LLMs simulating user responses can replicate social science experiments and online forums with a high degree of consistency comparable to those obtained using actual human participants.

The future of simulating various user types appears brighter as the accuracy of such simulations continues to improve. Experiments and studies in fields constrained by traditional methodologies stand to benefit significantly from advanced technologies like LLMs. For example, research exploring interactions with individuals who have mental health issues often faces ethical dilemmas and heightened risk assessments. Utilizing LLMs equipped with well-defined personas to simulate such participants can expedite research processes while minimizing potential risks to the interaction between researchers and subjects. Additionally, in studies seeking diverse and balanced samples, recruitment challenges often arise, especially when targeting specific demographic backgrounds. LLMs can be programmed to represent a range of demographics and personas, thus addressing recruitment limitations efficiently. Moreover, the financial implications of user studies involving large participant groups are considerable. By incorporating personas into LLMs, researchers can conduct extensive studies more cost-effectively, without compromising the breadth and diversity of participant profiles.

3.2 Role Playing in Specific Domains

LLM-based CAs, when programmed with specific personas, offer substantial support to educators, especially teachers, in improving their development of educational content, enriching their teaching methodologies, and bolstering their self-assurance. For instance, such agents can simulate a variety of student personas, enabling teaching assistants (TAs) to engage in realistic interaction scenarios [30]. This approach allows TAs to refine their skills in providing feedback and effectively addressing the needs of students with diverse characteristics, learning goals, and educational backgrounds. This comprehensive and authentic practice environment is instrumental in equipping TAs with the necessary competencies to minimize instructional mishaps in real-world teaching situations. Similarly, they have the potential to significantly enhance the professional skills of lawyers, physicians, and other specialists. These LLM-based CAs can simulate interactions with diverse patient types, including the elderly and those with unique symptoms or needs. Traditionally, such simulations form a crucial part of training before professionals are fully qualified. Now, with the integration of LLM agents equipped with specialized personas, this training phase can be streamlined and made more intelligence-oriented, offering a sophisticated approach to professional skill development.

Beyond their assistive role, these technologies can have personas to simulate domain experts, notably in healthcare, education and law. Here, CAs would blend intellectual and emotional support, innovatively simulating roles such as caregivers, tutors, and legal advisors. Nonetheless, their effectiveness hinges also on having accurate, domain-specific expertise, a critical aspect we will discuss in Section 4.3.

3.3 Brand Representation

The persona of an LLM-based CA plays a crucial role in brand representation by aligning with the brand's values, enhancing user engagement, and serving as a differentiator in a crowded market. For instance,

> " Domino's pizza created 'Dom', a virtual ordering assistant. Dom's persona is friendly and efficient, reflecting the brand's focus on convenient and fast service. Dom allows customers to order pizza using conversational language, making the process more engaging and aligning with Domino's commitment to innovation in delivery and customer service." [11]

A well-defined persona ensures that the agent's communication style and tone are consistent with the brand's identity, fostering a stronger and more coherent brand image. This alignment is essential not only for maintaining brand consistency but also for creating a more engaging and relatable experience for users. In an environment where many companies employ similar technologies, a distinctive persona can significantly set a brand apart, making it more memorable and appealing to customers. This unique identity helps in building customer loyalty and establishing a competitive edge.

4 CHALLENGES AND CAVEATS

4.1 Consistency Is the Top Priority

The primary objective in the design of CAs is to establish and nurture a robust connection with users, facilitating ongoing engagement over extended periods [42]. Achieving this necessitates the ability of the CAs to engage in sustained, meaningful conversations [43, 53]. Recent findings [13, 40] indicated that LLM-based CAs exhibit a heightened sensitivity to subtle and sensitive words within the context, leading to inconsistent outputs. This characteristic has raised concerns about the ability of LLM-based CAs to maintain a **consistent persona**[4] throughout multiple dialogue exchanges [17, 25]. This observation underscores the challenge of ensuring that these AI systems not only understand and process language effectively but also retain a consistent and contextually appropriate persona over successive interactions. Moreover, as discovered in [49], having inconsistency of persona is one of the major obstacles in achieving the long-term objective of developing human-like CAs to pass the Turing test [48] Addressing this issue is critical

[4]Consistency and coherency: **Consistency** means whether elements of persona remain unchanged throughout the conversation, e.g. you can not be a kid in one turn while talking like an old person in another. **Coherency** refers more to whether the persona elements are coherent, e.g. you can not say something like "I went on a trip with my wife for my 5-year-old birthday". Consistency cares more about persona across different turns, i.e. evolution across time and can only be defined for multi-turn dialogue.

in enhancing the reliability and user trust in conversational AI technologies [24, 33].

Unfortunately, most widely-used LLMs struggle to align responses consistently with latent persona attributes [41]. This inconsistency is particularly evident in complex tests, like reversing question meanings using negation. Only two out of fifteen models tested in this paper, achieved some level of consistency [41], highlighting the need for further development to enhance persona consistency in LLM responses.

4.2 Are There Effective Ways to Evaluate Persona and Its Consistency?

So far, a systematic approach to evaluate and verify persona application in LLM-based CAs has not been established. However, there exists some noteworthy attempts, such as employing empirical frameworks for indirectly assessing the persona of CAs [14, 39, 41]. This can be achieved through psychometric testing or by analyzing survey results.

It appears that one cannot ascertain the specific persona a LLM-based CA is exhibiting simply by prompting queries such as "what is your persona" or "describe your persona." Consider a scenario where an LLM-based CA is programmed or processed to embody a certain persona. The reality is, people cannot exhaustively enumerate all the traits of this persona, leaving room for the CA to exhibit some degree of self-expression in its responses. Moreover, the inherent unpredictability of the LLM adds a layer of complexity. For instance, the CA might be defined as "a 21-year-old physics student from Canada with a particular temperament..." but these specifications are insufficient to confine it to a specific character or individual. For example, in one interaction round, the LLM-based CA may fit this description but have a preference for bowling, while in the next round, it might have the same foundational characteristics but prefer skiing. In this situation, its persona has changed, yet such changes are subtle and challenging to detect and define. We can only ascertain their adherence to our initial constraints through certain predetermined questions. The CA might perfectly execute the task, but when asked about other aspects, like hobbies, it might reveal inconsistencies. Such situations are unpredictable and difficult to capture.

Moreover, we acknowledge that individual perceptions of a system's persona can vary. For instance, a chatbot with a female-like voice might be considered by someone as sufficiently demonstrating a persona. This variability poses challenges in establishing a universally accepted standard for assessing a system's capability to exhibit a persona.

4.3 More than Persona

This point becomes particularly prominent in our discussion about endowing "characters" with the ability to play different domain experts in LLM-based CAs. We believe that for CAs to successfully assume the required roles, it is essential to impart not only fundamental character traits such as demographics, age, and gender but also corresponding knowledge. Characters must also be equipped with professional knowledge that aligns with their identities. For example, a character role as an ophthalmologist should be familiar with basic ophthalmology as well as be able to fluently address complex questions about eye diseases. Similarly, a character claiming to be a judge should be acquainted with basic legal statutes. Moreover, effective role-playing entails not just possessing knowledge but also the ability to adapt responses according to different contexts. For instance, when asking a business consultant character about market trends, it should be capable of considering the current economic environment and specific industry dynamics to provide informed responses. This approach elevates that researchers should always think more than just persona. In designing and implementing these roles, careful consideration must be given to their expertise and adaptability to ensure their effectiveness and credibility in their respective fields.

Hallucination, as one of the crucial caveats in most LLMs, will trigger new problems in persona-based LLMs. Persona-based LLMs may hold the wrong belief in certain facts about themselves, e.g. occupation and social relationships. Current hallucination detection or evaluation methods depend on fixed non-persona-based datasets or uncertainty and inconsistency measures[29, 45]. However, in persona-based LLMs, such beliefs, once established, tend to remain consistent where the model is confident. Therefore, high self-consistency in persona-based LLMs requires more customised hallucination detection and prevention approaches to be developed.

4.4 Ethical Considerations

As with any AI-based technology, integrating personas into LLM-based CAs presents a dual-edged sword. It offers significant benefits but can also harbor potential risks. The use of such technology inevitably necessitates careful consideration of ethical issues. Potential harms include, but are not limited to, the ethics of deception and the reinforcement of societal stereotypes. We encourage our audience to refer to the provocation paper [36] for a more comprehensive discussion on the ethical considerations surrounding the use of personas in these systems. Here we refrain from redundant elaboration of previously stated points.

5 CONCLUSION

In integrating persona into LLM-based CAs, this provocation highlights the significance of persona to enhance human-like interactions. It covers the criticality of various applications, and meanwhile puts forward challenges in achieving persona consistency and domain-specific adaptability. In conclusion, although the prospect of creating CAs with high effectiveness and human resemblance is promising, prioritizing ethical standards and tackling technical challenges is essential. Future efforts must aim for responsible development that maximizes the benefits of persona integration while addressing its complexities.

ACKNOWLEDGMENTS

We thank CUI's anonymous reviewers for their constructive comments on previous drafts of this paper. This research was partially funded by EPSRC under grant *SAIS: Secure AI assistantS* (EP/T026723/1) and by the INCIBE's strategic SPRINT (Seguridad y Privacidad en Sistemas con Inteligencia Artificial) C063/23 project with funds from the EU-NextGenerationEU through the Spanish government's Plan de Recuperación, Transformación y Resiliencia.

REFERENCES

[1] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*. PMLR, 337–371.

[2] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[3] Timothy W Bickmore, Suzanne E Mitchell, Brian W Jack, Michael K Paasche-Orlow, Laura M Pfeifer, and Julie O'Donnell. 2010. Response to a relational agent by hospital patients with depressive symptoms. *Interacting with computers* 22, 4 (2010), 289–298.

[4] Katharine Cook Briggs. 1987. *Myers-Briggs type indicator*. G. Palo Alto, Calif. :Consulting Psychologists Press.

[5] Sydeney Butler. 2023. How to Create ChatGPT Personas for Every Occasion. Retrieved January 2024 from https://www.howtogeek.com/881659/how-to-create-chatgpt-personas-for-every-occasion/

[6] Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems* 47, 1 (2023), 33.

[7] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201* (2023).

[8] Andreea Danielescu and Gwen Christian. 2018. A bot is not a polyglot: Designing personalities for multi-lingual conversational agents. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–9.

[9] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051* (2022).

[10] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335* (2023).

[11] Domino's. 2014. MEET DOM: THE VIRTUAL VOICE ORDERING ASSISTANT FOR DOMINO'S PIZZA. Retrieved April 7, 2024 from https://ir.dominos.com/news-releases/news-release-details/meet-dom-virtual-voice-ordering-assistant-dominos-pizzar

[12] Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388* (2023).

[13] Iker García-Ferrero, Begoña Altuna, Javier Álvez, Itziar Gonzalez-Dios, and German Rigau. 2023. This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models. *arXiv preprint arXiv:2310.15941* (2023).

[14] Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988* (2023).

[15] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2022. Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5. *arXiv preprint arXiv:2212.05206* (2022).

[16] Youjin Hwang, Donghoon Shin, Sion Baek, Bongwon Suh, and Joonhwan Lee. 2021. Applying the persona of user's family member and the doctor to the conversational agents for healthcare. *arXiv preprint arXiv:2109.01729* (2021).

[17] Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547* (2023).

[18] Hankyung Kim, Dong Yoon Koh, Gaeun Lee, Jung-Mi Park, and Youn-kyung Lim. 2019. Designing personalities of conversational agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[19] Lucas Kohnke, Benjamin Luke Moorhouse, and Di Zou. 2023. ChatGPT for language teaching and learning. *RELC Journal* (2023), 00336882231162868.

[20] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702* (2023).

[21] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs). *BioMedInformatics* 4, 1 (2023), 8–33.

[22] Kausik Lakkaraju, Sara E Jones, Sai Krishna Revanth Vuruma, Vishal Pallagani, Bharath C Muppasani, and Biplav Srivastava. 2023. LLMs for Financial Advisement: A Fairness and Efficacy Study in Personal Decision Making. In *4th ACM International Conference on AI in Finance*. 100–107.

[23] Kausik Lakkaraju, Sai Krishna Revanth Vuruma, Vishal Pallagani, Bharath Muppasani, and Biplav Srivastava. 2023. Can LLMs be Good Financial Advisors?: An Initial Study in Personal Decision Making for Optimized Outcomes. *arXiv preprint arXiv:2307.07422* (2023).

[24] Nadine Lessio and Alexis Morris. 2020. Toward Design Archetypes for Conversational Agent Personality. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 3221–3228.

[25] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055* (2015).

[26] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155* (2016).

[27] Yuting Liao and Jiangen He. 2020. Racial mirroring effects on human-agent interaction in psychotherapeutic conversations. In *Proceedings of the 25th international conference on intelligent user interfaces*. 430–442.

[28] Junfeng Liu, Christopher Symons, and Ranga Raju Vatsavai. 2022. Persona-Based Conversational AI: State of the Art and Challenges. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 993–1001.

[29] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models.

[30] Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. GPTeach: Interactive TA Training with GPT Based Students. (2023).

[31] Amarachi B Mbakwe, Ismini Lourentzou, Leo Anthony Celi, Oren J Mechanic, and Alon Dagan. 2023. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. , e0000205 pages.

[32] Alex McFarland. 2023. What is a ChatGPT Persona? Retrieved January 2024 from https://www.unite.ai/what-is-a-chatgpt-persona/

[33] Sara Moussawi and Raquel Benbunan-Fich. 2021. The effect of voice and humour on users' perceptions of personal intelligent agents. *Behaviour & Information Technology* 40, 15 (2021), 1603–1626.

[34] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[35] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–18.

[36] Alisha Pradhan and Amanda Lazar. 2021. Hey Google, do you have a personality? Designing personality and personas for conversational agents. In *Proceedings of the 3rd Conference on Conversational User Interfaces*. 1–4.

[37] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207* (2018).

[38] Janko Roettgers. 2019. How Alexa Got Her Personality. Retrieved January 2024 from https://variety.com/2019/digital/news/alexa-personality-amazon-echo-1203236019/

[39] Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184* (2023).

[40] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324* (2023).

[41] Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Dallas Card, and David Jurgens. 2023. You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments. *arXiv preprint arXiv:2311.09718* (2023).

[42] Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19 (2018), 10–26.

[43] Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. *arXiv preprint arXiv:1905.12188* (2019).

[44] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714* (2015).

[45] Guangzhi Sun, Potsawee Manakul, Adian Liusie, Kunat Pipatanakul, Chao Zhang, Phil Woodland, and Mark Gales. 2024. CrossCheckGPT: Universal Hallucination Ranking for Multimodal Foundation Models. *arXiv:2405.13684* (2024).

[46] Richard Sutcliffe. 2023. A Survey of Personality, Persona, and Profile in Conversational Agents and Chatbots. arXiv:2401.00609 [cs.CL]

[47] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.

[48] Alan M Turing. 2009. *Computing machinery and intelligence*. Springer.

[49] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015).

[50] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).

[51] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

[52] Yangyu Xiao and Yuying Zhi. 2023. An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages* 8, 3 (2023), 212.

[53] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2016. Attribute2image: Conditional image generation from visual attributes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 776–791.

[54] Diyi Yang. 2019. *Computational Social Roles*. Ph. D. Dissertation. Carnegie Mellon University Pittsburgh, PA, USA.

[55] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243* (2018).

[56] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. *arXiv preprint arXiv:2004.12316* (2020).

[57] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.

[58] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910* (2022).

[59] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867* (2023).