June 2024

“HEY SIRI, DON’T MAKE ME MAD” – OVERCOMING USER ANNOYANCES WITH VOICE ASSISTANTS

Christina Strauss
Witten/Herdecke University, christina.strauss@uni-wh.de

Michael Dominic Harr
University of Duisburg-Essen, michael.harr@icb.uni-due.de

Reinhard Schütte
University of Duisburg-Essen, reinhard.schuette@icb.uni-due.de

Simon Wimmer
Stuttgart Media University, sw207@hdm-stuttgart.de

“HEY SIRI, DON’T MAKE ME MAD” – OVERCOMING USER ANNOYANCES WITH VOICE ASSISTANTS

Short Paper

Christina Strauss, Witten/Herdecke University, Witten, Germany, christina.strauss@uni-wh.de

Michael Harr, University of Duisburg-Essen, Essen, Germany, michael.harr@icb.uni-due.de

Reinhard Schütte, University of Duisburg-Essen, Essen, Germany, reinhard.schuette@icb.uni-due.de

Simon Wimmer, Stuttgart Media University, Stuttgart, Germany, sw250@hdm-stuttgart.de

Abstract

This study examines the effects of integrating a technically advanced and more human-like large language model into a voice assistant to assess, how technical advancements mitigate user annoyances. Therefore, a generative pre-trained transformer was integrated into Siri and made available to 23 interview participants. Preliminary results reveal a decrease in user-reported annoyances, showing that the integration not only improves technical accuracy but also enhances the perceived humanness of interactions. However, subsequent interviews indicated that the distinction between the effects of technical advancements and the infusion of humanness emerged as critical, indicating a complex interplay between these factors. It is therefore planned to differentiate between technical and human improvements in the further development of this article. The results contribute to the discourse on optimizing voice assistants by pinpointing the reduction of user annoyances as a pivotal factor in improving user experience, suggesting pathways for future enhancements in voice assistant platforms.

Keywords: Voice Assistants, Siri, User Annoyance, Large Language Model, Digitalization.

1 Introduction

The enormous expansion and improvement of artificial intelligence (AI) (e.g., Mariani et al., 2022) has enabled substantial progress in the development of Large Language Models (LLMs) (Kusal et al., 2022), revolutionizing the interaction between human and AI. Notably, Generative Pre-trained Transformers (GPTs) have emerged as a frontier in natural language processing (NLP), offering unprecedented capabilities in human-like text generation and comprehension (Deng & Lin, 2022). This evolution holds promise for voice assistants (VAs)—also called conversational agents (CAs) (Schöbel et al., 2023)—like Google Assistant or Siri, which have become an integral part of life (Poushneh, 2021). Nevertheless, despite their widespread adoption to satisfy both hedonic and utilitarian needs (e.g., Pitardi & Marriott, 2021), user annoyances (UA) were identified as hindrances to the effective use of VAs, undermining the potential for truly fluid and human-like interaction and therefore exerting a detrimental impact on the overall user experience (UX) (Demaeght et al., 2022). Existing research has extensively explored the technical capabilities and UX design of VAs (e.g., Oesterreich et al., 2023), anthropomorph communication styles (e.g., Brendel et al., 2023), and the refinement of automatic speech recognition (ASR) through LLMs (e.g., Min & Wang, 2024). Despite these advancements, there remains a gap in comprehending how the integration of advanced and more human-like LLMs specifically mitigate UA concerning the quality and comprehensibility of responses. Moreover, research has yet to thoroughly explore the distinction between advancements resulting from technical enhancements and those arising from enhanced human-like interaction qualities. We close this gap, hypothesizing that integrating an advanced and more human-like LLM into an existing VA will alleviate UA. Thus, we contribute to research on “the next generation of CAs” (Schöbel et al., 2023, p. 14) by answering the following research question: *“To what extent do technically advanced and human-like LLMs mitigate UA concerning the quality, conversational design, and comprehensibility of responses in VA interactions?”*

To address the research question, we, first, establish a baseline of UA with a given LLM; second, we integrate an advanced LLM (“Siri Pro”); third, we assess the mitigation on UA.

Preliminary results of our study showed the following: after testing “Siri Pro,” users hardly reported any UA in subsequent interviews. However, participants’ statements suggest that this was less due to greater perceived humanness than to technical improvements. This distinction leads to the need for a follow-up investigation that distinguishes between technical improvement and humanization of the LLM to reveal which of these factors leads to a reduction in UA. We assume that the technical improvement of Siri will lead to reduced UA, whereas greater humanness will lead to higher satisfaction and a better UX.

2 Theoretical Background

2.1 Voice Assistants and Generative AI

VAs exemplify manifestations of AI designed to engage in natural language communication with users based on audible interactions (Kulkarni et al., 2019; Kusal et al., 2022). VAs are defined as software programs executed within the provider’s data centers (Hörner, 2019). These assistants, despite distinct features, adhere to consistent core functions across applications, including home device control, music playback, call initiation, news retrieval, and information searches (Krol & Boßow-Thies, 2020). Their operational paradigm predominantly unfolds in five sequential stages: Activation, Speech Input, Speech Processing, Execution, and Speech Output (Krol & Boßow-Thies, 2020). Upon activation, the user’s analog message, along with metadata, is recorded and converted into digital signals, while simultaneously filtering out background noise (Hoy, 2018). These signals are then transmitted to the cloud providers (Krol & Boßow-Thies, 2020). This process involves the use of NLP techniques, starting with ASR and Speech-to-Text (STT) conversion (Subhash et al., 2020). Once the request is machine-readable, NLU techniques are employed to extract intentions, entities, context, meaning, and relationships from the user input (Kulkarni et al., 2019). The actual execution is fundamentally based on the understanding provided by the NLU component. The task can be executed directly on the device, such as setting an alarm (Krol & Boßow-Thies, 2020), or additional data can be retrieved from external application programming interfaces (APIs). The final step transforms machine-readable information into human-understandable language using NLG. This generated text is converted into a digital signal through Text-to-Speech (TTS) and presented to the user via the device’s screen or speakers (Hoy, 2018). Within these sub-processes, several limitations emerge that hinder usability and lead to UA (Demaeght et al., 2022), such as difficulties in comprehension (e.g., X. Wang & Nakatsu, 2013), intent classification and entity recognition (e.g., Villa et al., 2023), maintaining conversational context (e.g., Chen & Tseng, 2022), or generating contextually relevant responses (e.g., Graf & Zessinger, 2022). In contrast, the advent and recent advancements of generative AI bring a new paradigm centered on generative models that leverage machine learning (ML) and deep learning (DL) to create responses based on pre-trained knowledge (Goodfellow et al., 2014). These models can be used particularly in the field of NLP, comprising natural language understanding (NLU) and natural language generation (NLG), in which case they are referred to as LLMs (Bengio et al., 2003; Jeon & Lee, 2023). GPTs, rooted in transformer architectures and coupled with generative AI, exemplify an influential approach in NLP (Iskender, 2023). These models attain a comprehensive linguistic understanding, allowing them to generate contextually coherent language (Su & Yang, 2023). In the context of VAs, the added value of LLMs and GPTs lies in their non-domain-specific alignment, providing a wide range of possibilities for users. For instance, LLMs offer more enhanced technical opportunities (e.g., programming, image generation) and therefore excel in handling “out-of-domain” questions that traditional VAs struggle with (Hassija et al., 2023). Moreover, they enhance the degree of “product smartness” (Graf & Zessinger, 2022, p. 16), through increased autonomy and the capability for natural dialoge (Bahrini et al., 2023). Hence, LLMs enable interactions that feel personal and emotional (e.g., Kusal et al., 2022), thereby enhancing VAs anthropomorphism (Tsai & Chuan, 2023).

2.2 User Annoyances with Voice Assistants

UA represent significant impediments to user satisfaction and engagement (Demaeght et al., 2022) in the context of VAs. Since good technical functionality alone is often no longer enough for a system to be accepted (Castañeda et al., 2007), the pursuit of positive UX (including other factors, such as emotions and affect) takes center stage (e.g., Hassenzahl & Tractinsky, 2006). While the absence of UA is not sufficient to generate a positive UX (Hassenzahl & Tractinsky, 2006; Herzberg, 1959), the overarching goal "to create outstanding quality experiences rather than merely preventing usability problems" (Hassenzahl & Tractinsky, 2006, p. 95) implies, however, that the elimination of UA with VAs is still a fundamental precondition for superlative UX. UA arise from a variety of sources, including limitations of technology, interaction design, or unmet user expectations. Recent research shows that rigid rule-based VAs, restricting users to predefined paths, or technical deficiencies, like unintended activations, could lead to UA (e.g., Demaeght et al., 2022; Oesterreich et al., 2023; Thorat & Jadhav, 2020). Furthermore, the interaction design within VAs could lead to UA if it fails to engage users in human-like and goal-oriented conversations, for instance, due to limited NLU or intent classification capabilities (Min & Wang, 2024; Sargsyan et al., 2023). Technological advances, while providing opportunities for innovation, can paradoxically induce UA if they do not meet users' expectations or if they introduce new complexities (Jain et al., 2023; Sharma et al., 2022). When voice assistants do not perform as anticipated, users may experience frustration, leading to avoidance and disengagement from the technology (Sharma et al., 2022). One major factor that increases satisfaction is humanness. Prior research has shown that a more human-like AI is more accepted, liked, and trusted (Qiu & Benbasat, 2009; Staffa & Rossi, 2016). Furthermore, recent research by Oesterreich et al. (2023, p. 64) indicates, that "perceived humanness may be more important for users […] than system response accuracy", implying that users seek more from VAs than technical efficiency. This reflects a desire for interactions that mimic human communication, valuing characteristics such as empathy, understanding, and conversation naturalness. Therefore, it can be hypothesized that, while technical improvements will influence UA, greater humanness will lead to higher satisfaction and a better UX.

3 Method

First, to investigate the baseline of UA with Siri, we conducted an online survey with German-speaking respondents via SurveyMonkey. Participants were recruited from two German universities via email invitations for survey participation. The study utilized opportunistic sampling, requiring participants to have experience with VAs, regardless of usage frequency. Recognizing that our LLM integration primarily impacts NLP and response generation (i.e., NLU and NLG), eight UA were distilled from 26 annoyances identified by Demaeght et al. (2022), that can be influenced by the integration of an LLM (excluding, for example, instances such as false triggering). In detail, we meticulously chose eight items from the categories of response quality, comprehensibility, and conversational design, as outlined in Demaeght et al.'s (2022, p. 264) survey, who integrate findings from "various studies and literature on voice [UX]." To access their appearance in Siri, the questions were inverted to avoid biases and questioned on a 6-point Likert scale ranging from 1 ("strongly disagree") to 6 ("strongly agree"). 21 participants (seven female, 14 male; all regular Siri users) aged between 18 to 64 completed the survey.

Second, "Siri Pro," a prototypical VA directly connected to the OpenAI API, was developed to replace Apple's NLP component. By opting for Siri—one of the most used VAs (Suhr, 2020)—an existing user base was leveraged, capitalizing on users' familiarity with Siri. We integrated the "GPT-3.5-turbo" model of the "Chat Completion" endpoint from Chat GPT into Siri through an Apple Shortcut. This model was selected as it was the only available option in May 2023, in contrast to GPT-4 models, which were inaccessible for open research due to waiting lists. Considering the goal of mitigating UA through contextually relevant conversations with Siri Pro, the optimized "Chat Completion" endpoint was selected over the "Text Completion" endpoint (OpenAI, 2023a). In detail, user requests post-ASR-/STTphase are sent as API requests in JavaScript Object Notation (JSON) format to the Chat Completion endpoint of the OpenAI API (OpenAI, 2023a). The response is then processed based on individual parameters and returned in the same format to the device, where Siri transforms the written response from Siri Pro through TTS functionalities to articulate it audibly to the user (Apple, 2023; Hoy, 2018).

Finally, to assess whether Siri Pro mitigates UA, we conducted semi-structured interviews following guidelines from the literature (e.g., Adeoye-Olatunde & Olenik, 2021; Dearnley, 2005). Consistent with our preliminary study, participant recruitment was executed through email at the same two universities, employing opportunistic sampling with identical prerequisites for VA experience. First, Siri Pro was tested by participants. All interactions between participants and Siri Pro were recorded in separate notes. Then, the participants were asked the same questions as in the baseline survey. They were also interviewed on their impressions of the prototype and individual concerns. The average duration of an interview was 30:02 mins (17:20 mins to 54:41 mins). In total, 23 participants (five female, 18 male) aged between 19 and 31 (median age: 26) participated in the second part of the study. 19 participants used more than one VA, demonstrating varied usage frequencies: daily (16), weekly (6), or monthly (1). Common uses included playing music (14), navigation (11), setting alarms (11), and controlling smart home devices (8), highlighting the diverse applications of VAs in everyday tasks and entertainment.

4 Mitigating User Annoyances With Siri Pro

4.1 Preliminary Study and Post-Experiment Survey

The first step was to find out to what extent users encounter UA with Siri (see Table 1). It turns out that users already experienced relatively few annoyances with Siri: answers were comprehensible (mean: 5), repetitions occurred rarely (mean: 4.29) and users were understood (mean: 4.14). However, correcting oneself was not possible (mean: 2.62) and answers were sometimes not accurate enough (mean: 3.86).

	Preliminary Study (n=21)		**Post-Experiment Survey (n=23)**	
Question	**Mean**	**SD**	**Mean**	**SD**
When interacting with Siri/Siri Pro, did you have the feeling that...				
... Siri/Siri Pro has understood you in terms of content?	4.14	1.06	5.57	0.59
... Siri/Siri Pro remembers information that has already been given and you didn’t have to repeat it.	3.95	1.43	5.61	0.66
... Siri’s/Siri Pro’s statement or question is understandable.	5.00	0.71	5.22	0.67
… Siri/Siri Pro’s rarely repeats itself.	4.29	1.10	5.13	0.63
… Siri’s/Siri Pro’s answer is satisfactory.	3.95	0.97	4.87	0.81
... the results of your question are well presented.	3.95	1.20	4.09	1.16
... Siri’s/Siri Pro’s responses are accurate enough.	3.86	1.01	5.39	0.72
... there was a possibility of correction without having to start all over again.	2.62	1.32	5.30	1.10

Table 1. User Perceptions of Siri and Siri Pro.

The results of the post-experiment survey with Siri Pro showed that participants reported hardly any UA and attributed better performance to the new version in all respects (see Table 1). Especially the recall of information (mean: 5.61), the understanding of the content (mean: 5.57), and the accuracy of answers (mean: 5.39) were rated positively. The lowest ratings were given to the presentation of the answers (mean: 4.09) and satisfaction with answers (4.87), but here, too, Siri Pro outperformed its predecessor.

4.2 In-Depth Interviews

The results of subsequent qualitative interviews revealed that most users based their assessments primarily on technical aspects. Greater humanness was only rarely mentioned explicitly. All participants stated that they were convinced of the factual and substantive accuracy of Siri Pro's answers (despite an analysis of the conversations revealing that around 10% of the prototype's answers during the experiment actually contained incorrect information). Specifically, one participant said *"Yes, I do believe that the answers are correct, that was particularly evident with the bird of the year"* even though it was precisely this question that Siri Pro did not answer correctly. However, it also became apparent that many participants were aware of the risk of incorrect answers and would double-check important answers. One participant said, for example: *"In general, I find the answers trustworthy, but I would not blindly rely on them for very critical decisions such as health issues or exam dates at university; instead, I would check the statements again."* Participants also expressed concerns. As in previous studies (e.g., Bitkom, 2022), data protection was mentioned most frequently. Possible manipulation of answers by the VA operators as well as prejudices, injustices, and distortions in the answers were noted as well.

5 Discussion

In this study, we demonstrated that the implementation of a refined LLM in the Siri VA can mitigate UA. Our findings contribute to a better understanding of user interaction with VAs. There are several key observations and conclusions that can be drawn from our data. The improvement in understanding the content of the prototype could primarily be attributed to the use of LLMs instead of Apple's proprietary components. LLMs pre-trained on large amounts of data show a strong ability to extract the user's intent and build a deep understanding of the language and its relationships. The improvement in context retention and the low number of repetitions could be attributed to the ability of the LLM to make references to messages from a long time ago. This context preservation is limited in the various GPT versions by the maximum number of tokens for all messages in a conversation. However, this is continuously increased with each further development of the models (OpenAI, 2023b). This presumably also contributes to the reduction of repetitions, as the prototype can remember previously given answers. The (slightly) improved intelligibility of the prototype's utterances could also be due to the architecture of the LLMs, which enables enhanced language comprehension through training with large amounts of data. This understanding also allows them to generate responses in different styles, from rather simple to more cognitively demanding expressions. However, this adaptation seems to have a rather small impact, as Siri was already very well understood. The generally higher satisfaction could be the result of better results in all individual dimensions due to the generally expanded options. In our studies, users were on average more satisfied with all sub-aspects than before, which could have led to a generally higher level of satisfaction. The – relatively – poor results in terms of visualization could be because some test subjects (partly consciously, partly unconsciously) moved outside the limits of the system. For example, some test subjects asked for images to be displayed or apps to be started, but this was not possible due to the type of technical integration, leaving the algorithm with only voice output. The improvement in the accuracy of the answers could be partly due to the length of the prototype's answers, which exceeds that of traditional VAs. Additionally, the large amount of training data might have improved accuracy. The conversational nature of the prototype led to a significant improvement in the correction options and thus the greatest difference between Siri and Siri Pro. Our findings, in concert with insights by Oesterreich et al. (2023), delineate a nuanced interplay between technical enhancements and the humanization of VAs, underscoring that while technical advancements are crucial for reducing UA, the integration of perceived humanness amplifies user satisfaction. This complex relationship suggests that optimizing voice assistants for enhanced user experience necessitates a balanced focus on both augmenting technical functionalities and fostering human-like interactions, thereby catering to users' deeper desires for empathetic, understanding, and engaging communication. The analysis of the conversation histories showed that actual conversations took place between the participants and Siri Pro, with sometimes more than 15 messages. This indicates that VAs with integrated and enhanced LLMs can potentially become communication partners where instructions are no longer limited to specific commands. This has profound implications for VA providers in practice, who should leverage the evolving capabilities of LLMs to not only mitigate UA, but to evolve VAs into more versatile communication partners, moving beyond simple “command-response” interactions.

6 Further Steps in the Paper Development Process

Our paper is still at an early stage, which is why several further development steps are planned: First, the number of participants in all surveys is to be increased to enable statistically valid comparisons between groups. Additionally, we aim to make a clearer distinction (as far as possible) between a purely technical improvement of Siri and humanization. Previous studies have shown that greater humanness of AI can lead to higher user satisfaction (e.g., Qiu & Benbasat, 2009; Staffa & Rossi, 2016). This does not have to be explicitly reflected in fewer UA but mixing the two concepts can distort the results. Our current hypothesis is that a purely technical improvement leads to fewer UA (which humanization cannot reduce any further), whereas the implementation of a more human-like AI leads to higher satisfaction. Additionally, we assume that, in some cases, a more human-like AI might be able to overcompensate technical flaws. As these ideas were echoed in our qualitative interviews, we would also like to explicitly ask about the concept of user satisfaction in more detail in the questionnaire during the follow-up survey to be able to show possible interaction effects between these concepts. In addition, we plan to further expand the theory section of this paper. So far, we have broadly covered the concept of user experience, but the theoretical background of avoiding dissatisfaction (UA) and producing satisfaction (in our case, by humanization) has only been touched upon. Furthermore, we plan to revise the discussion section based on the results of our follow-up interviews as well as our experimental distinction between technical and human-like improvements of Siri Pro. Finally, we would like to add a limitation section. As our studies are not yet sufficiently mature, such a section would not yet have been expedient. Nevertheless, it is important to categorize our results in the final thesis and to show to what extent they are transferable and where the limits of our study are.

References

Adeoye-Olatunde, O. A., and Olenik, N. L. (2021). Research and scholarly methods: Semi-structured interviews. *Journal of the American College of Clinical Pharmacy*, *4*(10), 1358–1367.

Apple. (2023). *Voice trigger system for Siri*. URL: https://machinelearning.apple.com/research/voice-trigger (visited on November 10, 2023).

Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., and Pasehvar, M. (2023). ChatGPT: Applications, Opportunities, and Threats. *2023 Systems and Information Engineering Design Symposium (SIEDS)*, 274–279. IEEE.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*(6), 1137–1155.

Bitkom. (2022). Die Zukunft der Consumer Technology - 2022. URL: https://www.bitkom.org/sites/main/files/2022-08/220823_CT_Studie_2022.pdf (visited on October 30, 2023).

Brendel, A. B., Hildebrandt, F., Dennis, A. R., and Riquel, J. (2023). The Paradoxical Role of Humanness in Aggression Toward Conversational Agents. *Journal of Management Information Systems*, *40*(3), 883–913.

Castañeda, J. A., Muñoz-Leiva, F., and Luque, T. (2007). Web acceptance model (WAM): Moderating effects of user experience. *Information & Management*, *44*(4), 384–396.

Chen, E., and Tseng, Y.-H. (2022). A Decision Model for Designing NLP Applications. *Companion Proceedings of the Web Conference 2022*, 1206–1210. New York, NY, USA: ACM.

Dearnley, C. (2005). A reflection on the use of semi-structured interviews. *Nurse Researcher*, *13*(1), 19–28.

Demaeght, A., Nerb, J., and Müller, A. (2022). A survey-based study to identify user annoyances of German voice assistant users. In: F. Fui-Hoon Nah & K. Siau (Eds.), *HCI in Business, Government and Organizations. Lecture Notes in Computer Science* (pp. 261–271). Cham: Springer Gabler.

Deng, J., and Lin, Y. (2022). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, *2*(2), 81–83.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 3, pp. 2672–2680).

Graf, E., and Zessinger, D. (2022). Alexa, know your limits: developing a framework for the accepted and desired degree of product smartness for digital voice assistants. *SN Business & Economics*, *2*(6), 1–33.

Hassenzahl, M., and Tractinsky, N. (2006). User experience - a research agenda. *Behaviour & Information Technology*, *25*(2), 91–97.

Hassija, V., Chakrabarti, A., Singh, A., Chamola, V., and Sikdar, B. (2023). Unleashing the Potential of Conversational AI: Amplifying Chat-GPT's Capabilities and Tackling Technical Hurdles. *IEEE Access*, *11*(November), 143657–143682.

Herzberg, F. (1959). *Motivation to work*. New York, USA: Wiley & Sons.

Hörner, T. (2019). Marketing mit Sprachassistenten. In *Marketing mit Sprachassistenten*. Wiesbaden: Springer Fachmedien Wiesbaden.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, *37*(1), 81–88.

Iskender, A. (2023). Holy or Unholy? Interview with Open AI's ChatGPT. *European Journal of Tourism Research*, *34*, 3414.

Jain, S., Basu, S., Ray, A., and Das, R. (2023). Impact of irritation and negative emotions on the performance of voice assistants: Netting dissatisfied customers' perspectives. *International Journal of Information Management*, *72*(April), 102662.

Jeon, J., and Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, 1–20.

Krol, B., and Boßow-Thies, S. (2020). Akzeptanz von Sprachassistenten zur Steuerung von Smart Home Services. In: R. Buchkremer, T. Heupel, & O. Koch (Eds.), *Künstliche Intelligenz in Wirtschaft & Gesellschaft* (pp. 517–541). Wiesbaden: Springer Gabler.

Kulkarni, P., Mahabaleshwarkar, A., Kulkarni, M., Sirsikar, N., and Gadgil, K. (2019). Conversational AI: An overview of methodologies, applications and future scope. *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, 1–7. Pune, India: IEEE.

Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Mishra, S., and Abraham, A. (2022). AI-based conversational agents: A scoping review from technologies to future directions. *IEEE Access*, *10*, 92337–92356.

Mariani, M. M., Perez-Vega, R., and Wirtz, J. (2022). AI in marketing, consumer research and psychology: A systematic literature review and research agenda. *Psychology & Marketing*, *39*(4), 755–776.

Min, Z., and Wang, J. (2024). Exploring the Integration of Large Language Models into Automatic Speech Recognition Systems: An Empirical Study. In: B. Luo, L. Cheng, Z. G. Wu, H. Li, & C. Li (Eds.), *Communications in Computer and Information Science* (Vol. 1968, pp. 69–84). Singapore: Springer.

Oesterreich, T. D., Anton, E., Schuir, J., Brehm, A., and Teuteberg, F. (2023). How can I help you? Design principles for task-oriented speech dialog systems in customer service. *Information*

Systems and E-Business Management, *21*(1), 37–79.

OpenAI. (2023a). *API reference*. URL: https://platform.openai.com/docs/api-reference (visited on November 10, 2023).

OpenAI. (2023b). Models. URL: https://platform.openai.com/docs/models (visited on November 15, 2023).

Pitardi, V., and Marriott, H. R. (2021). Alexa, she's not human but… Unveiling the drivers of consumers' trust in voice-based artificial intelligence. *Psychology and Marketing*, *38*(4), 626–642.

Poushneh, A. (2021). Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. *Journal of Retailing and Consumer Services*, *58*, 1–10.

Qiu, L., and Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems*, *25*(4), 145–182.

Sargsyan, L., Choi, S., and Kim, S.-H. (2023). Investigating User Preferences for Conversation Design of Voice Assistant Systems using Linguistic Features. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *67*(1), 91–96.

Schöbel, S., Schmitt, A., Benner, D., Saqr, M., Janson, A., and Leimeister, J. M. (2023). Charting the Evolution and Future of Conversational Agents: A Research Agenda Along Five Waves and New Frontiers. *Information Systems Frontiers*.

Sharma, A., Dwivedi, R., Mariani, M. M., and Islam, T. (2022). Investigating the effect of advertising irritation on digital advertising effectiveness: A moderated mediation model. *Technological Forecasting and Social Change*, *180*(May), 121731.

Staffa, M., and Rossi, S. (2016). Recommender interfaces: The more human-like, the more humans like. In: A. Agah, J. Cabibihan, A. Howard, M. Salichs, & H. He (Eds.), *Social Robotics - Lecture Notes in Computer Science* (Vol. 9979, pp. 200–210). Springer, Cham.

Su, J., and Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education*, *6*(3), 355–366.

Subhash, S., Srivatsa, P. N., Siddesh, S., Ullas, A., and Santhosh, B. (2020). Artificial intelligence-based voice assistant. *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability*, 593–596. London, UK: IEEE.

Suhr, F. (2020). *Postbank Digitalstudie 2020*. URL: https://www.postbank.de/unternehmen/medien/meldungen/2020/juni/immer-mehr-deutsche-sprechen-mit-digitalen-sprachassistenten.html (visited on October 30, 2023).

Thorat, S. A., and Jadhav, V. (2020). A Review on Implementation Issues of Rule-based Chatbot Systems. *International Conference on Innovative Computing & Communications*.

Tsai, W.-H. S., and Chuan, C.-H. (2023). Humanizing Chatbots for Interactive Marketing. In: C. L. Wang (Ed.), *The Palgrave Handbook of Interactive Marketing* (pp. 255–273). Cham: Springer International Publishing.

Villa, L., Carneros-Prado, D., Sánchez-Miguel, A., Dobrescu, C. C., and Hervás, R. (2023). Conversational Agent Development Through Large Language Models: Approach with GPT. In: J. Bravo & G. Urzáiz (Eds.), *Lecture Notes in Networks and Systems* (Vol. 835, pp. 286–297). Cham: Springer.

Wang, X., and Nakatsu, R. (2013). How Do People Talk with a Virtual Philosopher: Log Analysis of a Real-World Application. In: J. C. Anacleto, E. W. G. Clua, F. S. C. da Silva, & S. Fels (Eds.), *Lecture Notes in Computer Science* (Vol. 8215, pp. 132–137). Berlin: Springer.