Generating Proactive Suggestions based on the Context: User Evaluation of Large Language Model Outputs for In-Vehicle Voice Assistants

Lesley-Ann Mathis
Fraunhofer Institute for Industrial Engineering IAO and University of Stuttgart
lesley-ann.mathis@iao.fraunhofer.de

Can Günes
University of Stuttgart, Institute of Human Factors and Technology Management IAT
can_guenes@outlook.de

David Lerch
Fraunhofer Institute for Optronics, System Technologies and Image Exploitation IOSB
david.lerch@iosb.fraunhofer.de

Kathleen Entz
Fraunhofer Institute for Optronics, System Technologies and Image Exploitation IOSB
kathleen.entz@iosb.fraunhofer.de

Frederik Diederichs
Fraunhofer Institute for Optronics, System Technologies and Image Exploitation, IOSB
frederik.diederichs@iosb.fraunhofer.de

Harald Widlroither
Fraunhofer Institute for Industrial Engineering IAO
harald.widlroither@iao.fraunhofer.de

ABSTRACT

Large Language Models (LLMs) have recently been explored for a variety of tasks, most prominently for dialogue-based interactions with users. The future in-car voice assistant (VA) is envisioned as a proactive companion making suggestions to the user during the ride. We investigate the use of selected LLMs to generate proactive suggestions for a VA given different context situations by using a basic prompt design. An online study with users was conducted to evaluate the generated suggestions. We demonstrate the feasibility of generating context-based proactive suggestions with different off-the-shelf LLMs. Results of the user survey show that suggestions generated by the LLMs GPT4.0 and Bison received an overall positive evaluation regarding the user experience for response quality and response behavior over different context situations. This work can serve as a starting point to implement proactive interaction for VA with LLMs based on the recognized context situation in the car.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI); Interaction paradigms; Natural language interfaces.

KEYWORDS

proactive voice assistants, LLMs, context-based suggestions, user evaluation

ACM Reference Format:
Lesley-Ann Mathis, Can Günes, David Lerch, Kathleen Entz, Frederik Diederichs, and Harald Widlroither. 2024. Generating Proactive Suggestions based on the Context: User Evaluation of Large Language Model Outputs for In-Vehicle Voice Assistants. In *ACM Conversational User Interfaces 2024 (CUI '24), July 08–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3640794.3665568

CUI '24, July 08–10, 2024, Luxembourg, Luxembourg

ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665568

1 INTRODUCTION

Large Language Models (LLMs) play an increasingly important role for the development of AI-based assistants given their out-of-the box ability to enable a multi-turn conversational interaction [8]. In the automotive industry, several manufacturers started to integrate LLMs into their voice assistant (VA) in the car. For example, Mercedes-Benz launched a beta-program in the US in 2023 which enhances their native VA with ChatGPT to cover a wider range of conversational topics and to enable a more interactive and natural conversational behavior[1]. Similarly, manufacturers like Stellantis have announced to work on LLM integrations for their cars for the European market[2].

Current in-car assistants solely react upon the user's input, either by pressing a push-to-talk button on the steering wheel or by calling a trigger phrase [21]. On the path towards more intelligent digital companions, the development of proactive VAs can be seen as the next step in the evolution of natural language interfaces in the car [1, 17]. For example, future VAs could suggest points of interests along the route [24] or make suggestions for a mindfulness exercise for stressed drivers [12]. To model and implement a proactive VA, key questions to be answered are *when* to approach the user, *what* is the content of the approach and *how* to approach him [19, 28]. Especially the user's context situation is essential for initiating a proactive interaction [1, 35], presuming that the system perceives and understands the user's environment and current activity [10]. LLMs have recently proven to be suitable for different language generation tasks based on a given instruction [33]. In particular, the advantage of LLMs is the ability to mimic human conversation without having to rely on predefined dictionaries and grammars, as traditional approaches in natural language processing [8]. We thus see the potential to use LLMs for creating proactive suggestions

[1] https://group.mercedes-benz.com/innovation/digitalisation/connectivity/car-voice-control-with-chatgpt.html, last accessed January 02, 2024
[2] https://www.media.stellantis.com/uk-en/ds/press/ds-automobiles-is-first-european-manufacturer-to-integrate-artificial-intelligence-system-chatgpt-into-its-cars, last accessed January 02, 2024

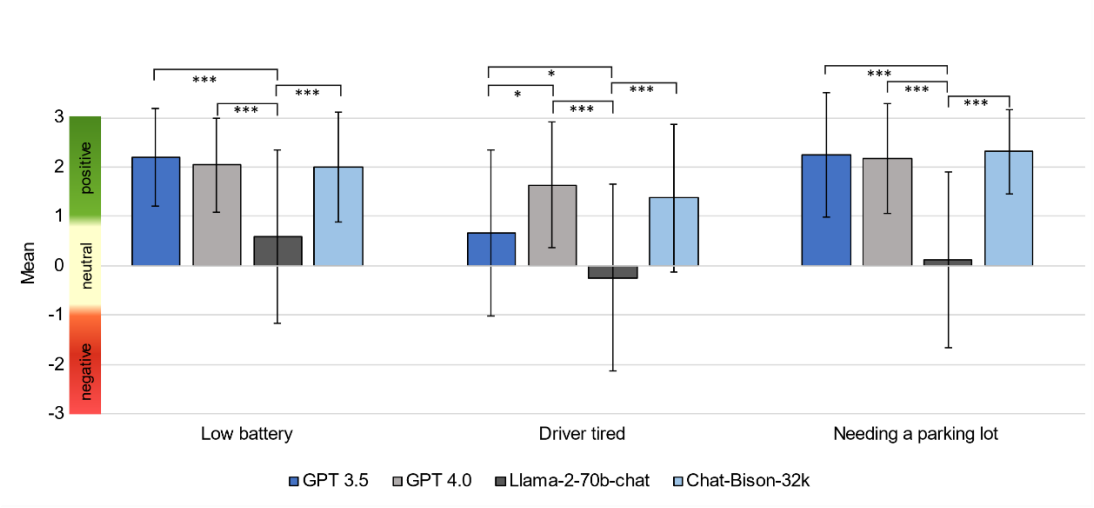

Figure 1: Mean values for *response quality*, error bars depict *SD* (* $p < .05$, ** $p < .01$, *** $p < .001$)

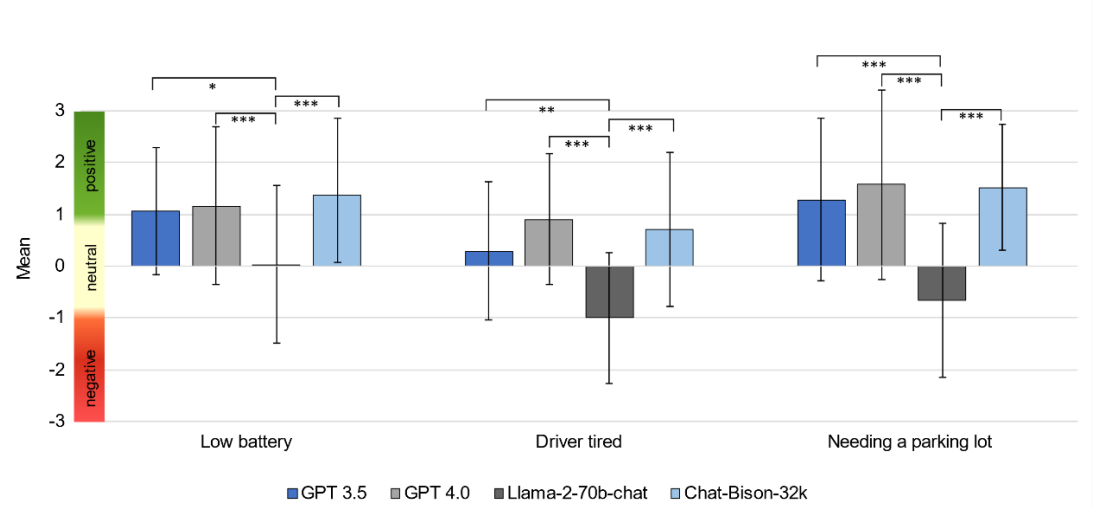

Figure 2: Mean values for *response behavior*, error bars depict *SD* (* $p < .05$, ** $p < .01$, *** $p < .001$)

to GPT3.5 (Z = 0.814, p = .021), GPT4.0 (Z = 1.640, p < .001) and Bison (Z = -1.314, p < .001). The comparison between GPT3.5 and GPT4.0 is also significant (Z = -.826, p = .018). For the situation *needing a parking lot*, response quality is rated significantly lower for Llama compared to GPT3.5 (Z = 1.686, p < .001), GPT4.0 (Z = 1.453, p < .001), and Bison (Z = -1.605, p < .001). None of the other post-hoc comparisons for response quality yielded a significant difference.

For response behavior, a significant difference was found for *low battery* ($\chi^2(3) = 29.263$, p < .001), *driver tired* ($\chi^2(3) = 54.212$, p < .001), and *needing a parking lot*, ($\chi^2(3) = 53.015$, p < .001). The conducted post-hoc tests with Dunn-Bonferroni corrections applied show for *charging station* that response behavior for Llama is rated significantly lower compared to GPT3.5 (Z = .826, p = .018), GPT4.0 (Z = 1.140, p < .001) and Bison (Z = -1.337, p < .001). Similarly for *driver tired*, there is a significant difference between Llama compared to GPT3.5 (Z = 1.093, p = .001), GPT4.0 (Z = 1.779, p < .001) and Bison (Z = -1.686, p < .001). For *needing a parking lot*, response behavior for Llama is rated significantly lower than for GPT3.5 (Z = 1.430, p < .001), GPT4.0 (Z = 1.709, p < .001) and Bison (Z = -1.558, p < .001). None of the other differences yielded a significant result.

4.2 Most Preferred Suggestion per Context Situation

Findings for the most preferred suggestion (see Fig. 3) show that for the context situation *low battery*, Bison's output is favored most often. Optional comments by participants who chose Bison as favorite (N = 21) emphasize that they value the brevity and conciseness of the suggestion (11 out of 14 comments), as indicated in this

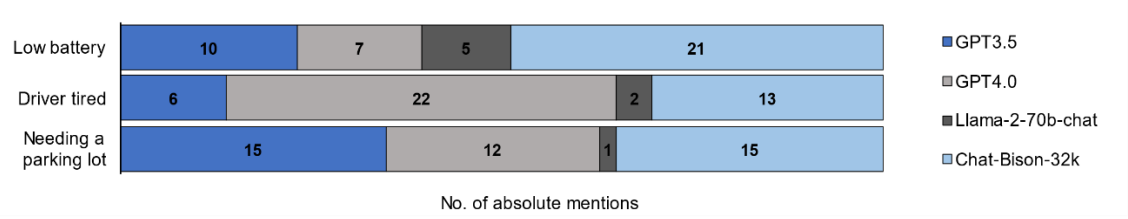

Figure 3: Most preferred proactive suggestion per context situation ($N = 43$)

comment (P38): "The important information was communicated discreetly, simply and in the shortest possible way." In contrast, for the situation *driver tired*, GPT4.0 is most often chosen as favorite ($N = 22$). The comments again reveal that brevity of this statement is important (9 out of 10 comments): It is preferred as it "sounds friendly but is also short" (P37) and "sounds more natural" (P23) than the other three suggestions. For needing a parking lot, outputs by GPT3.5 and Bison are favored equally often as a first choice (each $N = 15$). This is also reflected in the reasoning of one participant (P43), who chose the output of Bison as favorite: "Like the first option [GPT3.5], it is short and limited to the essentials. I like both as they only differ by one word. In addition, there is also an exact specification of 5 minutes [until arrival], which I personally prefer". In line with the results for response quality and behavior, Llama's output is favored by only few people in every context situation.

Some participants also point out negative aspects about the suggestions. One participant criticizes the statements for *driver tired* by GPT3.5, GPT4.0 and Llama because of "a too personal reference to an alleged sensation (you seem to be tired)" (P26). The suggestion by Bison is less personal for this context and thus preferred by the participant, with the formulation "I see that you've been riding for a while". Controversary arises in the comments regarding the use of a personal style as in "Would you like me to find you a parking space nearby?". While this statement is valued for "[. . .] creating a more personal basis and trust." (P38), another participant prefers more anonymous utterances without being personally addressed "[. . .] to feel freer in the decision" (P07).

5 DISCUSSION

The following section discusses the implications of our findings, the limitations of our work as well as starting points for future research.

5.1 Using LLMs to Generate Context-Based Proactive Suggestions: User Perception and Practical Implementation

We derived a basic prompt structure which could be successfully applied to generate proactive suggestions with four different LLMs based on a given context situation. LLMs show promising capabilities for the implementation of proactive behavior, as no task-specific datasets or rule-based patterns must be developed beforehand. The derived prompt structure shall serve a basis to investigate the generation of context-based suggestions, where further work can build on.

Our findings reveal significant differences between the investigated LLMs, showing that suggestions generated by Llama are perceived less human-like in response behavior as well as less clear and distinct in all context situations. In contrast, suggestions by GPT4.0 meet users' expectations best regarding a human-like interaction style and receiving clear, useful answers for all investigated contexts, very closely followed by the positive evaluations for Bison. The context situation *driver tired* stands out compared to the other two, given generally lower scores for both response quality and behavior. Especially the suggestions generated by GPT3.5 and by Llama are considered less likable and less suitable. Reasons might be the extensive length of Llama's output and the combination of two alternatives entailed in the suggestion by GPT3.5 for this context situation. Thus, it might be the case that the generated outputs for functional use cases (such as when the battery is empty or a parking lot is needed) are more aligned with user expectations and leave less margin for error. In contrast, situations pertaining the user state could pose more challenges for generating suitable suggestions. As pointed out by participants in the comments, they can be perceived as imposing or even an insinuation towards the user.

From a practical point of view, the information entailed in the prompt describing the context situation can be extracted from the sensory information available in the car, such as driver monitoring systems detecting drowsiness and distraction [7] and occupant monitoring systems detecting passenger activities. In manual driving, the driving task is the driver's main activity, while in automated driving other user activities become dominant and can be derived from camera images [13, 14]. Hence, it can be helpful for a suitable timing and content of the proactive suggestion to consider the current activity. For example, it is likely to be less disturbing to start an interaction with the user when she or he is eating than reading [15]. On the other hand, motion sickness commonly develops when reading on curvy roads, so activity recognition can be used as context to trigger motion sickness warnings [6]. Speech interaction can also be used to keep drivers awake [2] or to wake up sleeping drivers, which will become a relevant state in automated driving [9, 25].

5.2 Potentials and Challenges of LLMs to Implement the Virtual Companion in the Car

The generated suggestions by the LLMs GPT4.0 and Bison show a positive user experience evaluation on nearly all context situations and are favored comparably often. Users highlight positively

the naturalness, brevity and conciseness of the generated suggestions. Thus, the created utterances come already close to those of a respectful and trustworthy dialogue partner who makes a suitable and beneficial suggestion in a given situation. LLMs might thus become a key enabler for the implementation of the virtual companion in the car [18]. Further potential lies in enhancing the derived prompt structure with user information to provide more personalized suggestions (see e.g., [34]). The qualitative feedback has shown that users perceive the style and formulations of utterances differently, e.g. regarding the use of personal pronouns by the assistant. Suitable proactive suggestions beyond functional use cases, such as parking lot suggestions, must consider user-specific factors as relevancy and preferences for proactive interactions are highly individual [37].

A common caveat when using LLMs is the steerability and reliability of their outputs [32], which have to be considered for the implementation of proactive behavior. The car is a safety critical domain, where users should not be distracted in selected situations, and proactive behavior might be perceived as intrusive or imposing by users [37]. Thus, proactive behavior using LLMs for in-vehicle VA should be integrated after sufficient testing and with a possibility to iteratively adapt to the user's needs, e.g., regarding the content and style of the proactive suggestion.

5.3 Limitations

This work has several limitations. Our study sample was rather young and limited in size so that further research is required to validate our evaluation results. A rotation of the scenarios and suggestions in the questionnaire was not applied. While we checked for careless responding in the data, fatigue might have influenced the presented results. We did not fine-tune the available LLMs which might further improve the received outputs. While the focus of our work was on showing the feasibility of generating proactive suggestions based on exemplary context situations and comparing the outputs of different off-the-shelf LLMs, our approach cannot raise a claim to completeness. The prompt structure might not be generalizable to other LLMs and it still remains to be investigated how LLMs handle more complex context situations. In addition, we focused on generating suggestions in German and findings may be language dependent. Especially the neutral to negative evaluation for suggestions generated by Llama might be different when conducting a similar study in other target languages. In addition, the chosen scales for evaluating the generated suggestions, response quality and response behavior, do not allow for detailed conclusions on specific aspects of the formulations, such as perceived politeness, naturalness, or trustworthiness. The current study's focus was on generating and evaluating the initial suggestion, while future studies could investigate complete proactive dialogues with assistants.

6 CONCLUSION

We explored the use of four off-the-shelf LLMs to generate proactive suggestions for an in-car VA. Our basic prompt structure including a role description, the current context, an instruction and specification of the output style was successful in generating proactive suggestions with each selected LLM. Proactive suggestions by the models GPT4.0 and Bison received overall highest ratings on the user experience scales response quality and response behavior. Feedback by participants highlights the importance of brief and concise proactive suggestions as well as a natural and simple language style. In contrast, the use of a personal address in the suggestions as well as suggestions pertaining the user's state can be perceived as imposing. This work serves as starting point for using LLMs to implement proactive behavior for in-car assistants based on the recognized context.

ACKNOWLEDGMENTS

The work has been funded under the funding codes 19A21031F and 19A21031L by the Federal Ministry for Economic Affairs and Climate Action of Germany (BMWK) on the basis of a decision by the German Bundestag and by the European Union.

REFERENCES

[1] Caterina Bérubé, Marcia Nißen, Rasita Vinay, Alexa Geiger, Tobias Budig, Aashish Bhandari, Catherine R. Pe Benito, Nathan Ibarcena, Olivia Pistolese, Pan Li, Abdullah B. Sawad, Elgar Fleisch, Christoph Stettler, Bronwyn Hemsley, Shlomo Berkovsky, Tobias Kowatsch, and A. B. Kocaballi. 2024. Proactive behavior in voice assistants: A systematic review and conceptual model. Computers in Human Behavior Reports, 100411. DOI: https://doi.org/10.1016/j.chbr.2024.100411.

[2] Christer Ahlström, Johanna Wörle, Mikael Ljung Aust, and Frederik Diederichs. 2023. Road Vehicle Automation and Its Effects on Fatigue, Sleep, Rest, and Recuperation. In The Handbook of Fatigue Management in Transportation. CRC Press, 513–524. DOI: https://doi.org/10.1201/9781003213154-43.

[3] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering ChatGPT's Capabilities in Recommender Systems. In Proceedings of the 17th ACM Conference on Recommender Systems. ACM, New York, NY, USA, 1126–1132. DOI: https://doi.org/10.1145/3604915.3610646.

[4] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. In CHI Workshops, May 10. New Orleans, LA.

[5] Anand Das. 2023. Epic Battle of AI Models – Google Bard, ChatGPT-3.5, GPT-4, Bison PaLM 2, and Anthropic Claude: Unveiling the Best (2023). Retrieved March 31, 2024 from https://bito.ai/blog/epic-battle-of-ai-models-google-bard-chatgpt-3-5-gpt-4-bison-palm-2-and-anthropic-claude-unveiling-the-best/.

[6] Frederik Diederichs, Amina Herrmanns, David Lerch, Zeyun Zhong, Daniela Piechnik, Lesley-Ann Mathis, Boyu Xian, Nicklas Vaupel, Ajona Vijayakumar, Canmert Cabaroglu, and Jessica Rausch. 2024, in press. Activities that correlate with motion sickness in driving cars – an international online survey. In HCI International 2024.

[7] Frederik Diederichs, Christoph Wannemacher, Fabian Faller, Martin Mikolajewski, Manuel Martin, Michael Voit, Harald Widlroither, Eike Schmidt, Doreen Engelhardt, Lena Rittger, Vahid Hashemi, Manya Sahakyan, Massimo Romanelli, Bernd Kiefer, Victor Fäßler, Tobias Rößler, Marc Großerüschkamp, Andreas Kurbos, Miriam Bottesch, Pia Immoor, Arnd Engeln, Marlis Fleischmann, Miriam Schweiker, Anne Pagenkopf, Lesley-Ann Mathis, and Daniela Piechnik. 2022. Artificial Intelligence for Adaptive, Responsive, and Level-Compliant Interaction in the Vehicle of the Future (KARLI). In HCI International 2022 Posters, Constantine Stephanidis, Margherita Antona and Stavroula Ntoa, Eds. Communications in Computer and Information Science, 1583. Springer International Publishing, Cham, 164–171. DOI: https://doi.org/10.1007/978-3-031-06394-7_23.

[8] Vishal Goar, Nagendra S. Yadav, and Pallavi S. Yadav. 2023. Conversational AI for Natural Language Processing: An Review of ChatGPT. International Journal on Recent and Innovation Trends in Computing and Communication 11, 3s, 109–117. DOI: https://doi.org/10.17762/ijritcc.v11i3s.6161.

[9] Maria Hirsch, Frederik Diederichs, Harald Widlroither, Ralf Graf, and Sven Bischoff. 2020. Sleep and take-over in automated driving. International Journal of Transportation Science and Technology 9, 1, 42–51. DOI: https://doi.org/10.1016/j.ijtst.2019.09.003.

[10] Li Jinglu and Xiaohua Sun. 2023. Exploring Proactivity in Human-Vehicle Interaction: Insights for proactive interaction Design. In Human Factors in Transportation. AHFE International. AHFE International. DOI: https://doi.org/10.54941/ahfe1003804.

[11] Andreas M. Klein, Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. 2020. Construction of UEQ+ scales for voice quality. In Proceedings of Mensch

und Computer 2020. ACM, New York, NY, USA, 1–5. DOI: https://doi.org/10.1145/3404983.3410003.

[12] Kevin Koch, Varun Mishra, Shu Liu, Thomas Berger, Elgar Fleisch, David Kotz, and Felix Wortmann. 2021. When Do Drivers Interact with In-Vehicle Well-being Interventions? An Exploratory Analysis of a Longitudinal Study on Public Roads. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 5, 1, 1–30. DOI: https://doi.org/10.1145/3448116.

[13] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad S. Khan. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models (June 2023). Retrieved April 8, 2024 from http://arxiv.org/pdf/2306.05424v1.

[14] Manuel Martin, David Lerch, and Michael Voit. 2023. Viewpoint Invariant 3D Driver Body Pose-Based Activity Recognition. In 2023 IEEE Intelligent Vehicles Symposium (IV). IEEE. DOI: https://doi.org/10.1109/iv55152.2023.10186682.

[15] Lesley-Ann Mathis, Daniela Piechnik, Carla B. Bubeck, Selina Layer, and Harald Widlroither. 2024. "Ist jetzt ein guter Zeitpunkt?": Proaktive Ansprachen durch einen Sprachassistenten bei fahrfremden Tätigkeiten im Realverkehr. In Dokumentation des 70. Arbeitswissenschaftlichen Kongresses: Arbeitswissenschaft in-the-loop - Mensch-Technologie-Integration und ihre Auswirkung auf Mensch, Arbeit und Arbeitsgestaltung. GfA-Press, Sankt Augustin.

[16] Lesley-Ann Mathis, Kathrin Werner, and Harald Widlroither. 2023. Exploring use cases and user perception of a proactive voice assistant in automated vehicles. In Human Factors in Transportation. AHFE International. AHFE International. DOI: https://doi.org/10.54941/ahfe1003803.

[17] Anna-Maria Meck, Christoph Draxler, and Thurid Vogt. 2023. How May I Interrupt? Linguistic-Driven Design Guidelines for Proactive in-Car Voice Assistants. International Journal of Human–Computer Interaction, 1–15. DOI: https://doi.org/10.1080/10447318.2023.2266251.

[18] Anna-Maria Meck, Marion Sardone, and Jacqueline Cullmann. 2023. Will the Assistant Become the Driver, and the Driver Become the Assistant? In Proceedings of the 5th International Conference on Conversational User Interfaces. ACM, New York, NY, USA, 1–5. DOI: https://doi.org/10.1145/3571884.3603753.

[19] Florian Nothdurft, Stefan Ultes, and Wolfgang Minker. 2015. Finding Appropriate Interaction Strategies for Proactive Dialogue Systems—An Open Quest. In Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, K. Jokinen and M. Vels, Eds. Linköping University Electronic Press, 73–80.

[20] Elnaz Nouri, Robert Sim, Adam Fourney, and Ryen W. White. 2020. Proactive Suggestion Generation. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York, NY, USA, 1585–1588. DOI: https://doi.org/10.1145/3397271.3401272.

[21] Cathy Pearl. 2017. Designing voice user interfaces: Principles of conversational experiences. O'Reilly, Beijing, Boston, Farnham, Sebastopol, Tokyo.

[22] Maria Schmidt and Patricia Braunger. 2018. Towards a speaking style-adaptive assistant for task-oriented applications. In Studientexte zur Sprachkommunikation, 143–150.

[23] Maria Schmidt, Daniela Stier, Steffen Werner, and Wolfgang Minker. 2019. Exploration and assessment of proactive use cases for an in-car voice assistant. Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung, 148-155.

[24] Martin Schrepp. 2023. User Experience Questionnaire Handbook: All you need to know to apply the UEQ successfully in your projects (2023). Retrieved from https://www.ueq-online.org/Material/Handbook.pdf.

[25] Doreen Schwarze, Frederik Diederichs, Lukas Weiser, Harald Widlroither, Rolf Verhoeven, and Matthias Rötting. 2023. Are Drivers Allowed to Sleep? Sleep Inertia Effects Drivers' Performance after Different Sleep Durations in Automated Driving. Multimodal Technologies and Interaction 7, 6, 62. DOI: https://doi.org/10.3390/mti7060062.

[26] Taylor Shin, Yasaman Razeghi, Robert L. Logan, IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts (October 2020). Retrieved April 8, 2024 from http://arxiv.org/pdf/2010.15980v2.

[27] Aleksandar J. Spasić and Dragan S. Janković. 2023. Using ChatGPT Standard Prompt Engineering Techniques in Lesson Preparation: Role, Instructions and Seed-Word Prompts. In 2023 58th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST). IEEE, 47–50. DOI: https://doi.org/10.1109/ICEST58410.2023.10187269.

[28] Xiaohua Sun, Jinglu Li, and Weiwei Guo. 2023. A design approach of proactive HMI based on smart interaction. In Proceedings of the 6th International Conference on Intelligent Human Systems Integration (IHSI 2023) Integrating People and Intelligent Systems, February 22–24, 2023, Venice, Italy. AHFE International. AHFE International. DOI: https://doi.org/10.54941/ahfe1002823.

[29] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models (February 2021). Retrieved April 8, 2024 from http://arxiv.org/pdf/2102.02503.

[30] Timm Teubner, Christoph M. Flath, Christof Weinhardt, Wil van der Aalst, and Oliver Hinz. 2023. Welcome to the Era of ChatGPT *et al.* Bus Inf Syst Eng 65, 2, 95–101. DOI: https://doi.org/10.1007/s12599-023-00795-x.

[31] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian C. Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit S. Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric M. Smith, Ranjan Subramanian, Xiaoqing E. Tan, Binh Tang, Ross Taylor, Adina Williams, Jian X. Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models (2023). Retrieved April 8, 2024 from https://doi.org/10.48550/arXiv.2307.09288.

[32] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling Conversational Interaction with Mobile UI using Large Language Models. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, 1–17. DOI: https://doi.org/10.1145/3544548.3580895.

[33] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams W. Yu, Brian Lester, Du Nan, Andrew M. Dai, and Quoc Le V. 2021. Finetuned Language Models Are Zero-Shot Learners. Paper presented at ICLR 2022 (September 2021). Retrieved April 8, 2024 from http://arxiv.org/pdf/2109.01652.

[34] Yunjia Xi, Weiwen Liu, Jianghao Lin, Xiaoling Cai, Hong Zhu, Jieming Zhu, Bo Chen, Ruiming Tang, Weinan Zhang, Rui Zhang, and Yong Yu. 2023. Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models (June 2023). Retrieved April 8, 2024 from http://arxiv.org/pdf/2306.10933v4.

[35] Neil Yorke-Smith, Shahin Saadati, Karen Myers, and David N. Morley. 2012. The Design of a Proactive Personal Agent for Task Management. Int. J. Artif. Intell. Tools 21, 01, 1–30. DOI: https://doi.org/10.1142/S0218213012500042.

[36] J. D. Zamfirescu-Pereira, Richmond Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, 1–21. DOI: https://doi.org/10.1145/3544548.3581388.

[37] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah T. Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. 2022. Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. In 4th Conference on Conversational User Interfaces. ACM, New York, NY, USA, 1–14. DOI: https://doi.org/10.1145/3543829.3543834.