Methods for Analyzing Human-Conversational GenAI Interaction: A Pilot Study for Requirements Development and Validation

X[1], X[1], and X[1]

[1]X

Keywords: Human-Machine Interaction ; Conversational Generative Artificial Intelligence ; System Engineering, Requirements Engineering ; Pilot study.

1 Introduction

Defined as AI systems capable of producing "new and diverse content, in various formats, and for different tasks, by leveraging generative models" (García-Peñalvo and Vázquez-Ingelmo, 2023), the conversational GenAI (CGenAI) has already inspired researchers to study their integration in system engineering (Perreau *et al.*, 2024) after being valued by requirements engineering (**e.g.**; Arora *et al.*, 2023). Looking at previous research, the investigation of human-CGenAI interactions (H-CGenAI.I) seems to suffer from a lack of methodological tools adapted to their specific generative nature, increasing the difficulty of their research and testing in the field. Furthermore, the concentration of mainly theoretical work or self-demonstration has shown a minority of empirical demonstration with end users, denoting an absence of validation of the proposed guidelines or templates (Rapp *et al.*, 2023). To reduce this knowledge gap, this pilot study focused on developing and trialling a mixed approach to support a process-based assessment. The research method is evaluated on a case study that aims to analyse the impact of prompting methodical recommendations on the definition of requirements, as part of the research question: How do prompting guidelines and templates impact the quality of the requirements definition? This research aims to propose new methods of H-CGenAI.I analysis with the integration of human-science perspectives and practices. As far as the authors know, no previous research has proposed this approach for a system engineering context. The results and observations of the method application in a use case are hoped to support the specification, design, and evaluation of the enabling system composed of the system engineer and the CGenAI.

2 Methods

To allow a cross-comparison of subjective and objective views for H-CGenAI.I, a mixed approach was preferred using four data collection methods: questionnaires and video, dialogue, and semi-structured interview recording. Three main personas were defined based on their experience level of knowledge in Systems Engineering (SE), Model-Based SE (MBSE), and the use of conversational AI. After completing a free consent form, the participants were given a notice containing a glossary of technical keywords used in the instructions of three exercises that must be read according to the order of the experimental tasks. The participants had to complete a questionnaire to assess their general knowledge in requirements engineering and conversational GenAI. Then, they were given 10 minutes to read the first exercise. The instructions were defined around an electric toothbrush system, which didn't require specific technical knowledge to be conceptualized. The volunteers had to derive functional requirements from the system - seen from a black box (external) perspective - before validating them. Both tasks were repeated in three exercises where the participants were free to require the assistance of ChatGPT 4.0: – 1) without prompting guidelines for the system function "To transform electric power into mechanical brushing power", 2) with prompting guidelines for the system function "To send last brushing duration", and 3) with prompting templates for the system function "To inform on status ON/OFF/IDLE". Finally, researchers conducted a semi-directed interview with each participant within a limit of one hour.

3 Results

Three participants were selected for the experiment, each participant matching one persona profile: a PhD and MBSE researcher with a high level of knowledge of CGenAI, a MBSE and SE practitioner with less than four previous CGenAI experience of use and a SE apprentice with daily use of CGenAI. The research collected nine dialogues associated with three interviews transcribed. A quantitative study was planned to calculate quality scores for each modality based on criteria defined in SAE ARP4754A (2015) (SAE, 2015), ISO/IEC/IEEE 29148 (2018) (ISO/IEC/IEEE, 2018), and the INCOSE Guide to Writing Requirements (INCOSE, 2022). However, the poor quality of the generated requirements made the approach unpracticable and unusable.

A qualitative analysis was also conducted on the obtained materials. The coding was elaborated from the scientific literature and adapted with a grounded theory approach. Whereas the quantitative approach was found inconclusive, the qualitative analysis allowed a deeper investigation of the chronology of the task, the forms of guidances used respectively by the CGenAI and the user, the frequency of use of the templates and guidelines and their associated needs explicitly written in the different prompts. From the behavioural side, the method succeeded in retrieving the decision-making following the CGenAI proposition and the intentions detailed during the interviews. This information was compared to the management of the requests, the attention allocation, and the social coordination with role distribution. As a last indicator, the study of the dialogues offered the analysis of trouble sources and repair attempts through their length and position from their target.

4 Discussion

As looked at in the research question about the influence of prompting guidelines and templates on the quality of requirement definition, the proposed mixed methods succeed in investigating the reason for the performance and highlight the necessity of defining new indicators for H-CGenAI.I. However, without the factual indications of previous personal experiences, the pre-assessment questionnaire failed to differentiate the participant profiles precisely, however observed with the first prompt exchange with the GenAI. Although the study enabled us to understand the dynamic of the interactions, the analysis still lacks a clear view of the participant's internal requirements. As a participant representation might differ from the prescriptive criteria, the behaviours and production couldn't be entirely related to the capacity of adaptation of the technology, questioning the user-centric or task-centric evaluation perspective. Additionally, the limited number of participants restrains the scope of the analysis. Future studies should consider that point by integrating auto-evaluative measures of status and completeness towards the task. The authors also recommend specifying the coding grid with the observed activity characteristics, including the activity needs, requirements, and constraints.

References

Arora, Chetan, Grundy, John, and Abdelrazek, Mohamed (2023). *Advancing Requirements Engineering through Generative AI: Assessing the Role of LLMs*, arXiv: 2310.13976[cs]. **available at**: http://arxiv.org/abs/2310.13976 (accessed 17 Sept. 2024).

García-Peñalvo, Francisco and Vázquez-Ingelmo, Andrea (2023). "What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI", *IJIMAI*, Vol. 8 No. 4, p. 7. ISSN: 1989-1660. DOI: 10.9781/ijimai.2023.07.006. **available at**: https://www.ijimai.org/journal/sites/default/files/2023-11/ijimai8_4_1.pdf (accessed 28 Jan. 2024).

INCOSE, Requirements Working Group (2022). *Guide to Writing Requirements*,

ISO/IEC/IEEE, 29148 (2018). *Systems and software engineering Life cycle processes — Requirements engineering*,

Perreau, Emilie, Pinquié, Romain, and Masclet, Cédric (2024). "ModACT -Proposition d'une méthodologie de recherche pour l'étude des interactions humains-agent conversationnel textuel en milieu écologique : application pour des tâches d'Ingénierie Système." *ModACT 2024*. Lab for User Cognition & Innovative Design (LUCID) and Luxembourg Institute of Science and Technology (LIST). Paris, France. **available at**: https://hal.science/hal-04681012.

Rapp, Amon *et al.*, (2023). "Collaborating with a Text-Based Chatbot: An Exploration of Real-World Collaboration Strategies Enacted during Human-Chatbot Interactions", *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23: CHI Conference on Human Factors in Computing Systems. ACM: Hamburg Germany, pp. 1–17. ISBN: 978-1-4503-9421-5. DOI: 10.1145/3544548.3580995. **available at**: https://dl.acm.org/doi/10.1145/3544548.3580995 (accessed 2 Oct. 2023).

SAE, ARP4754A (2015). *ARP4754A Guidelines For Development Of Civil Aircraft and Systems*,