Parents, You Are Not the Only Ones Struggling With Childrens' Questions: Exploring Childrens' Preference for Clarification Methods Used by Voice Assistants

TESSA VAN BELOIS, University of Twente, The Netherlands

Parents are constantly asked questions by their children. Now there are systems on the market designed to answer those questions, but how do they ask for clarification when faced with an ambiguous query? Voice Assistants like Siri, Amazon Alexa, and Google Assistant are growing in popularity across different age demographics, including children. While research has been conducted on different methods of query clarification for adults, there is little understanding of children's perspectives on this matter. This paper aims to bridge this gap by emulating the study done with adult participants by Kiesel et al (2018) [8]. That study was adapted to fit within the limitations of this research and to be appropriate for a younger research population. We compare two different methods a voice assistant can use for asking for clarification, in order to find which one is preferred by children, based on how they perceive the functionality of the Voice Assistant and their level of enjoyment when using them. Our findings include that both versions have their merits, and further research needs to be done to find what method of clarification a voice assistant designed for children should employ.

Additional Key Words and Phrases: Query Clarification, Voice Assistants, Ambiguity

1 INTRODUCTION

Voice Assistants (VAs), like Amazon Alexa and Google Assistant, are widespread and utilized across various age groups. While these systems are designed with an adult user in mind, many children have access to, and interact with them [5]. VAs generally work in a simple question-answer format, where the user asks the system questions, and the system answers, or says it does not understand the question and needs further explanation. This is usually enough for adults who know how to rephrase questions, and have a clearer understanding of what information they want to get from the system. Even so, various research has been done on more complex clarification mechanism to improve adults' user experience [9, 10]. Queries from children are often vague and thus harder to answer for VAs, as they tend to have limited proficiency in their language and less general knowledge [6]. Additionally, children have difficulty rephrasing their questions and elaborating their intent when the system fails to comprehend their initial query [11, 13]. Because of this, a VA that is specifically designed for children should have a different approach to asking the user for clarification.

Children have a harder time interacting with Artificial Intelligence systems, like chatbots or Voice Assistants, due to their lack of proficiency in writing clear and complete queries. We want to find out whether we can improve the way a VA, designed for children, asks for clarification on a query it does not understand [4]. In a recent paper [2], the researchers compared a simple question-answer AI with a conversational one. Their aim was to find out whether children prefer a more conversational interaction with an AI. While there was no clear result from the research, the children did seem to prefer an interaction that resembled a conversation more than a simple question and answer. Previous research has been conducted on how humans ask clarifying questions [3], as well as how systems might generate such questions in response to query ambiguity [7, 14].

1.1 The original study: *Towards Voice Query Clarification*

Research has been done around the topic of query clarification, but mainly on adults [12, 15]. The design of the study used in this paper is based on the study performed by Kiesel et al (2018) [8]. This paper focuses on the decrease of user satisfaction when asked for clarification from an AI, versus getting the right answer immediately. The researchers compared seven different styles of clarification, including giving three meanings or categories and asking to verify whether the specified meaning is the correct one. The results seem to indicate that adult users do not mind needing to clarify their query, as opposed to getting the right answer back immediately. The users do however dislike it when a system interprets their input incorrectly. The preferred clarification method is the 3-meanings response method, where the user is asked to choose from 3 meanings of the ambiguous word that the system did not understand. The paper does stress that these results are dependent on the length of the possible answers and the English proficiency of the user. The results did however include that in most cases users prefer to also be able to clarify themselves, instead of having to choose one of the three options.

The study described in this paper is not a true replication of the original study, it has been adapted to fit a smaller scope and a younger research population. Instead of the 7 different methods they compared, two were selected for this study. The first method, referred to as the *open-question method*, is comparable to how most Voice Assistants work now. When asked a query the system does not fully comprehend, it will reply with something along the lines of: "I do not understand your query, could you provide context or rephrase the question?". The alternative method that will be compared in this study will be further referred to as the *multiple-choice* clarification method, "Did you possibly mean this, that, or thus?". This method might be preferable for children, considering their possible trouble formulating queries and lesser grasp on what information they want to gain. Based on the preference of the participants, whether they enjoy using each version of the system and think they work effectively, the aim is to determine which of these methods of clarification is more suitable for a younger audience.

TScIT 40, February 2, 2024, Enschede, The Netherlands

This goal is summarized in the following research question:

Research Question: Do children indicate a preference between the open-question and multiple-choice clarification methods employed by a Voice Assistant, when evaluating the effectiveness and user experience?

To try to answer this question, a within-subject study was conducted, where fourteen participants were asked to perform four research tasks, with the help of a VA in the form of a Furhat robot. After performing two tasks with one version of the system, the participants gave their opinion on system usability and user experience through a short questionnaire, using a 5-point Likert scale. The same process was repeated with the other version, employing the other method of asking for clarification. The results of the questionnaires, as well as notes taken during the study, were collected and processed and were used for both quantitative and qualitative analyses.

2 METHOD OF RESEARCH

The methodology of this research is based on the study done in Kiesel et al's paper [8]. Lowering the age of the participants brings some potential complications; less time with each participant, a shorter attention span, and an increased difficulty of understanding and performing the tasks. Because of these limitations, and a general time constraint for the research with 15 minutes per participant, our study was adapted to be shorter and more in-depth. In the next section, the similarities and differences to this research will be discussed, as well as an overview of how the study was conducted.

2.1 Setup

The 15 minutes per participant were divided into an introduction, two sessions with two tasks each, and a brief closing interview. There was one researcher present to interact with the participant, command the robot and take notes. The participants sat facing the robot and the researcher, who facilitated the interaction with a laptop and kept oversight. They were given a sheet of paper with the tasks and a pen.

A Furhat robot took the place of the Alexa assistant used in the original research. The Furhat is a social robot designed to perform research with, it gives the researcher more control over the interactions during the study. The decision was made based on multiple factors: as opposed to an Alexa it is capable of interacting in Dutch and the fun and novel nature of the Furhat keeps the participants more engaged and makes participating in the study more appealing. To make the Furhat able to respond to the queries by the participants, the combined functionality with OpenAI's developer space was employed. Using the prompt-based chatbot function, two chatbots were made, each with their own clarification method.

2.2 Participants

Multiple students at the *Libanon Lyceum Rotterdam*, a Dutch high school, were asked to participate in the study during class time. The research was set up in collaboration with teachers at the school. Children whose parents consented joined the experiments. 14 participants joined the study, pooled from 3 first and one second grade class, with their ages ranging from 12 to 15.

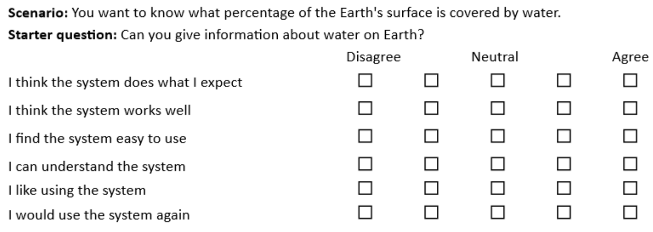
2.3 Independent Variable

The independent variable of both studies is the type of clarification method. In the original study, 7 different methods of clarification were compared, in 13 tasks assigned to each participant. In our study, we compared two methods, the open-question method and the method that was most popular in Kiesel et al's study. The open-question method is similar to the way current AI chatbots or VAs prompt users to provide additional information or rephrase their questions. The alternative method that is compared, the multiple-choice method, gives the user 3 possible options to choose from.

2.4 Measurements

As defined in the research question, the methods are assessed on two major components, the effectiveness of communication and the user experience. While the participants interact with the system, multiple elements are measured to define the effectiveness of the communication. For each task, it is noted how well the participant was able to complete it, based on whether they were able to finish the task within the set time limit and get the information they were asked to retrieve, as compared to the predefined answers. The speed of the interactions is recorded in the amount of conversation turns.

In addition to being effective, the system should be enjoyable for the children to use. Therefore, the evaluation also considers both the systems' appeal to children. After performing two tasks with the first version of the system, the children fill in a questionnaire, using the Likert scale 1. These questions were adapted from the original study, to fit the research and the participants better. Since the participants only had to fill it in twice, instead of 13 times, the initial questionnaire of 4 questions was extended into 6 questions, with three questions aimed at the user experience, and three at the perceived effectiveness of the system. A brief interview was conducted for a more qualitative comparison of the two versions of the system.



Scenario: You want to know what percentage of the Earth's surface is covered by water.
Starter question: Can you give information about water on Earth?

	Disagree		Neutral		Agree
I think the system does what I expect	□	□	□	□	□
I think the system works well	□	□	□	□	□
I find the system easy to use	□	□	□	□	□
I can understand the system	□	□	□	□	□
I like using the system	□	□	□	□	□
I would use the system again	□	□	□	□	□

Fig. 1. One example task with the questionnaire. Note that the study was performed in Dutch, and thus this has been translated to English

2.5 Procedure

In the study, the two methods are compared in a within-subject type experiment, where each participant interacts with both versions of the system in two consecutive sessions. Before performing the tasks, the participants are briefed on what is expected of them, and given an explanation of the tasks and the system. Just like in the original study, each task consists of a short description of the scenario, with a corresponding start question, Figure 1. The initial query is vague by design, to force ambiguity. The participant has to start the interaction with the starter question, and the system answers using the assigned clarification method. With this setup, we ensure that the participant interacts at least once with the method of clarification assigned to the session, during each task. After the initial response from the system, the interaction phase between the participant and the voice assistant (referred to as the system) starts. Here the participant is free to interact with the system, the goal being to solve the scenario. When the scenario has been solved, or the time has run out, the participant moves on to the next task. Participants were presented with a sheet of paper with two tasks, with a scenario and query (named Starter question) each, and a questionnaire, see Figure 1. After the interaction phase was completed, the participants were debriefed and asked several questions to compare the two methods of clarification.

2.5.1 The Tasks. Since there are two tasks for each version of the system, a distinction was made between the type of ambiguity in the starter question. The different types of ambiguity implemented are syntactic and polysemic [1]. In practice it means that per session, for one of the tasks the initial query was vague because the question could be interpreted different ways, and in the other the query contained a word that can have different meanings, like the word '*bank*' in Dutch, that can mean both bank and couch.

2.5.2 The System. A Furhat robot took the place of the Alexa assistant used in the original research. The Furhat is a social robot, designed to perform research with. The decision was made based on multiple factors: An Alexa is not capable of interacting in Dutch, the Furhat gives more control during the study, and the fun and novel nature of the Furhat keeps the participants more engaged and makes joining in the study more appealing. To make the Furhat able to respond to the queries of the participants, the combined functionality with OpenAI's developer space was employed. Using the prompt-based chatbot function, two chatbots were made, each with their clarification method.

3 RESULTS

3.1 Quantitative Data from The Questionnaires

The six questions of the questionnaire were divided into two themes. The first three questions were aimed at the perceived functionality of the system, or whether the participants thought the system worked well. The second set of questions aimed to find out whether the participants enjoyed using the system. After the data of the first participants was collected, the reliability of the questionnaire was tested using the Cronbach Alpha test. Both sets of questions were found to be reliable, which means that for each set, the three questions can be combined into a scale. The answers to the questions in the scale can be combined, as they are proven to be internally correlated. The questionnaire can also be used in further research, as it has been proven to be reliable.

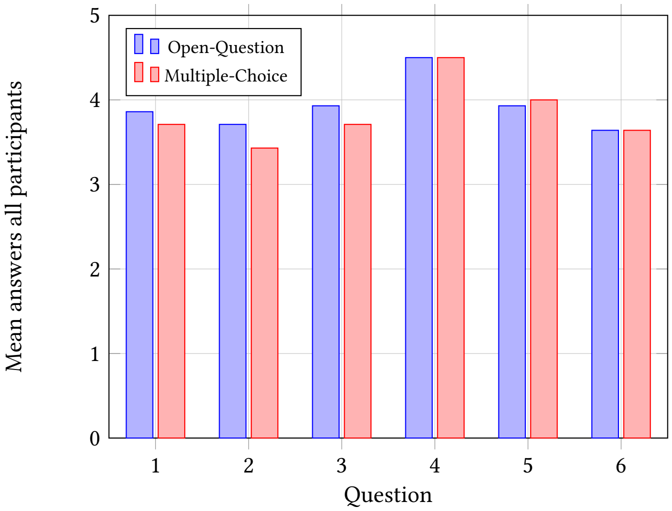


Fig. 2. Means of the answers of all participants per question.

Statistical analysis was performed on the data from the questionnaires, both on the scales and the questions individually. After running parametric and non-parametric tests, for example, the repeated measure ANOVA, we concluded that there is no significant difference between the two methods of clarification, which means that we cannot conclude that either of the methods is better than the other. In Figure 2, the means of the answers per questions of all participants are shown. Even though most participants indicated a preference for one of the two versions, this did not seem to influence whether they would prefer to use the system again, or not, looking at the means for question six. Question four was about if the system is understandable, however, the original Dutch word is not aimed at comprehension - if they understood what the system meant - but more whether the system was easy to hear. Since neither the voice nor any of the other settings of the robot were different between the sessions, the equality of the means is explicable. While there is no statistically significant difference, it is still worthwhile to look at the individual answers of the participants. Looking at the means of all the questions together per participant, we can see that two of the participants did not show a preference for either of the versions of the system in the questionnaire. Out of the other twelve participants, exactly six showed a preference for the open-question method and exactly six for the multiple-choice method. This explains why the means of the two methods for each question are so similar.

3.2 Qualitative Data from The Interviews

After the interaction phase, the researcher explained the exact difference between the two versions of the system. The participants were then asked for their preference and reasoning, in a short interview. The results of which system they said they preferred can be seen in figure 3. Four participants preferred the multiple-choice method, with their reasoning being that it was easier to use and helped them ask the right questions. One participant said (statement translated from Dutch) "When you don't know exactly what you want to

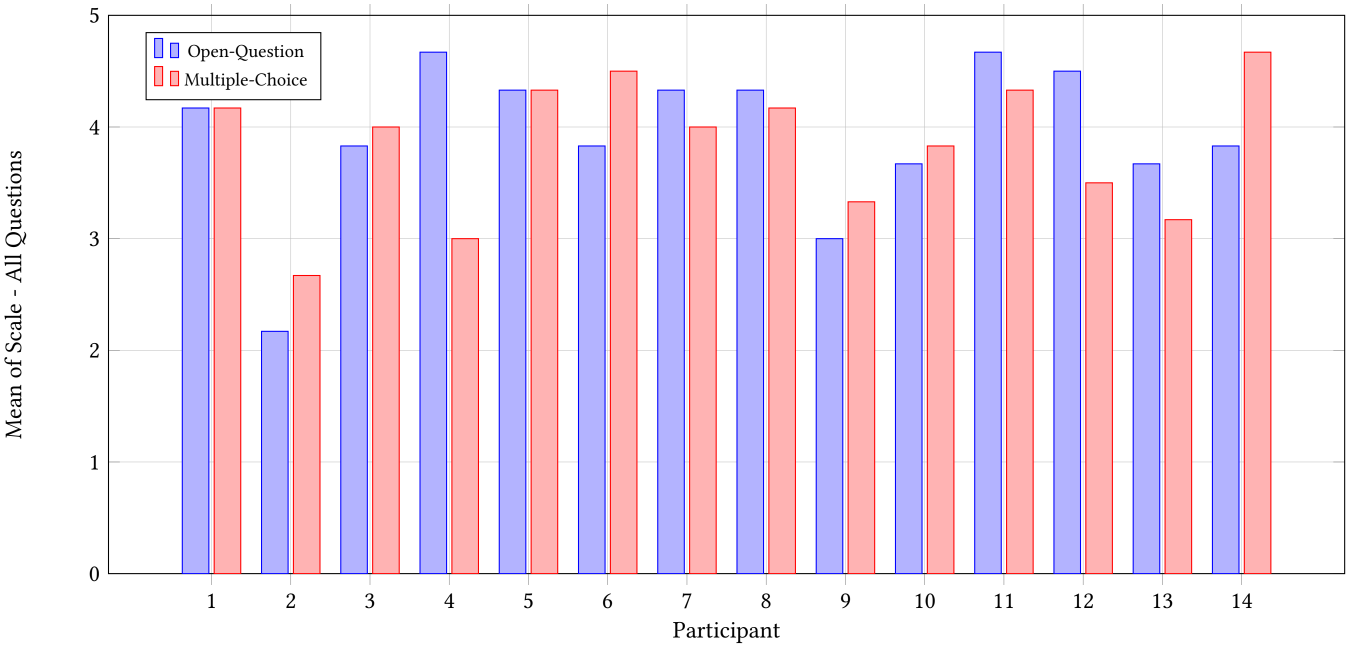


Fig. 3. Means of Answers to Questionnaires per Participant.

know, he can help you ask a question." Six participants indicated a preference for the open-question method, with two specifically mentioning that it was because none of the options the multiple-choice method provided were the ones they wanted. Three indicated that they had no preference between the versions. Notably, one participant initially favored having more options; however, when the right answer was not one of the options, the participant did not know how to proceed.

4 DISCUSSION

4.1 Limitations

With the participant pool drawing from a single school, conclusions drawn from the sample are less likely to apply to the general population. Additionally, the research team was limited to one member who had to manage the robot, interact with the participants, and document the proceedings. This resulted in less comprehensive notes being taken, which means there is less data present on the actual effectiveness of the system, or how well the tasks were completed. Having more detailed notes on each task could lead to more in-depth insights into the impact of the actual effectiveness, as opposed to the perceived effectiveness that was measured in the questionnaires.

During the study, there were many things, aside from the independent variable, that could impact the opinion of the participants. For example, the robot was not always able to understand what the participant said, due to limitations of speech-to-text function, or because the participant did not speak clearly. Other factors can include the connection to the chatbot failing, the chatbot not giving the exact answer the participant wants, or even the topic of the task itself being less interesting to the participant. All these factors can impact the way the participant fills in the questionnaire for each session.

Another big factor that was not controlled during the study was whether the correct answer was included in the three options the multiple-choice method supplied. While the use of the chatbots made for a more realistic interaction - these systems could be employed as voice assistants today - it meant that there was less control during the study. From the results of the interview, it seems that the inclusion of the answer the participant was looking for within the multiple-choice options had a big influence on whether they liked the system or not. This variable was however not controlled, so this is not proven statistically.

The use of Likert scales brings the risk of certain biases, for example, the acquiescence bias, where participants tend to agree with statements and the social desirability bias, the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others. These biases are likely present in the results, especially when performing research with children, with a researcher present.

Using the Furhat robot made the study more interesting for both the researcher and the participants. It is however likely that the use of the robot, as opposed to an Alexa or other voice assistant, led to more positive results in the questionnaires. After being told that the robot might not answer the questions correctly, one of the participants said that it was okay and that it was already impressive that the robot was able to speak at all. This could lead to the conclusion that the form of the VA contributes to its being accepted more or less, including its perceived quality. Standards could be higher for an Alexa than for a Furhat because it looks less human. The responses to the questionnaire were generally more positive, which can be explained by this in combination with the Likert scale biases. In some cases, even if the task was not completed successfully, the participant answered on the positive side of the Likert scale for the questions that aimed at functionality. For the purpose of this study, it is however not detrimental, as there are no statistically significant results to dispute, and the robot as a variable did not differ over the two sessions.

4.2 Future work

The biggest influence on the preference of the participants seems to be whether they think the correct answer is included in the options the multiple-choice method gave. Because of the way the study described in this paper was conducted, this variable was not controlled. Further research should be conducted, in a similar setup, where the researcher fully controls what the robot will reply. Or at least the initial response to the system should be set, which is doable because the start question is predefined either way. This way the researcher can control the options the multiple-choice gives, and compare the effects of having the correct one included or not. It is of course still possible that the correct answer is included, but that it is not the answer the participant is looking for, depending on what they think they want to know.

Further research could be done into a system that uses a combination of the two methods, or one of the other methods that is described in the study by Kiesel et al. Whilst the adults preferred the multiple-choice method, children might prefer one of the other options they researched. We can conclude that there is not a one-size-fits-all system and that it depends on the person using it. Maybe a combination of the two methods, where the VA supplies multiple choices only when they seem useful. This would be a much more complicated system, how would the VA know whether the options are useful or not? It is worth looking into. The research should be combined with the way children ask questions and reformulate their queries, it would be useful to look at it from a linguistic or pedagogical perspective as well, to try to gain a deeper understanding of the age group.

5 CONCLUSION

The goal of this study was to add to the existing research done on query clarification methods with adults, and the research on Voice Assistants for children, by looking specifically at two clarification methods that a VA can employ. While statistically, we cannot conclude that one of the methods is better than the other, we can see from the data that both versions have their merits. The average scores of each method per question are very similar, but for each participant individually they are quite different. For some participants, the briefness of the open-question was preferred, and the options the multiple-choice gave were more a hindrance than help. For others, the options the multiple-choice method gave helped them reformulate their queries. The results seem to indicate that neither of the clarification methods is the best in practice. The results of the interviews seem to indicate that the biggest factor that determined whether the participants enjoyed using the system employing the multiple-choice method, was the correct answer being included in the three options or not. We can also conclude from the study that the set-up could be used in further research, and especially the scales of the questionnaires are proven to be reliable and reusable. Further research needs to be done on types of clarification and the decrease in user experience when an unhelpful clarification question is asked.

ACKNOWLEDGMENTS

I would like to extend my gratitude to the Libanon Lyceum Rotterdam; to Mr. Verbeek who helped immensely with set up the study, and to the students for their participation. I would like to thank my supervisor, Prof. Dr. V. Evers, for all her support, ir. T.H.J. Beelen for his help with practical aspects, and both for their feedback. Last but absolutely not least, special thanks to my father for making it possible.

REFERENCES

[1] Giulia Baker and Michelle Aldridge. 2021. Disambiguating ambiguity: Providing a framework for classifying types of ambiguity. *Linguistics and the Human Sciences* 14, 3 (Mar. 2021), 237–260. https://doi.org/10.1558/lhs.19339

[2] Thomas Herman Johan Beelen, Khiet Phuong Truong, Roeland J.F. Ordelman, Ella Velner, Vanessa Evers, and Theo W.C. Huibers. 2022. A Child-Friendly Approach to Spoken Conversational Search. In *CIKM-WS 2022 (CEUR Workshop Proceedings)*, Georgios Drakopoulos and Eleanna Kafeza (Eds.). CEUR. 2nd Workshop on Mixed-Initiative Conversational Systems, MICROS 2022, MICROS 2022 ; Conference date: 21-10-2022 Through 21-10-2022.

[3] Pavel Braslavski, Denis Savenkov, Eugene Agichtein, and Alina Dubatovka. 2017. What Do You Mean Exactly? Analyzing Clarification Questions in CQA. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 345–348. https://doi.org/10.1145/3020165.3022149

[4] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. "Hey Google is it OK if I eat you?": Initial Explorations in Child-Agent Interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children* (Stanford, California, USA) *(IDC '17)*. Association for Computing Machinery, New York, NY, USA, 595–600. https://doi.org/10.1145/3078072.3084330

[5] Allison Druin, Elizabeth Foss, Hilary Hutchinson, Evan Golub, and Leshell Hatley. 2010. Children's roles using keyword search interfaces at home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 413–422. https://doi.org/10.1145/1753326.1753388

[6] Tatiana Gossen, Thomas Low, and Andreas Nürnberger. 2011. What are the real differences of children's and adults' web search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 1115–1116. https://doi.org/10.1145/2009916.2010076

[7] Kimiya Keyvan and Jimmy Xiangji Huang. 2022. How to Approach Ambiguous Queries in Conversational Search: A Survey of Techniques, Approaches, Tools, and Challenges. *ACM Comput. Surv.* 55, 6, Article 129 (dec 2022), 40 pages. https://doi.org/10.1145/3534965

[8] Johannes Kiesel, Arefeh Bahrami, Benno Stein, Avishek Anand, and Matthias Hagen. 2018. Toward Voice Query Clarification. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, Ann Arbor, MI, USA, 4. https://doi.org/10.1145/3209978.3210160

[9] Hadrien Lautraite, Nada Naji, Louis Marceau, Marc Queudot, and Eric Charton. 2021. Multi-stage Clarification in Conversational AI: The case of Question-Answering Dialogue Systems. *CoRR* abs/2110.15235 (2021). arXiv:2110.15235 https://arxiv.org/abs/2110.15235

[10] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[11] Nicholas Vanderschantz and Annika Hinze. 2021. Children's query formulation and search result exploration. *International Journal on Digital Libraries* 22, 4 (2021), 385–410. https://doi.org/10.1007/s00799-021-00316-9

[12] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring Conversational Search With Humans, Assistants, and Wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (<conf-loc>, <city>Denver</city>, <state>Colorado</state>, <country>USA</country>, </conf-loc>) *(CHI EA '17)*. Association for Computing

Machinery, New York, NY, USA, 2187–2193. https://doi.org/10.1145/3027063.3053175

[13] Svetlana Yarosh, Stryker Thompson, Kathleen Watson, Alice Chase, Ashwin Senthilkumar, Ye Yuan, and A. J. Bernheim Brush. 2018. Children asking questions: speech interface reformulations and personification preferences. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (Trondheim, Norway) *(IDC '18)*. Association for Computing Machinery, New York, NY, USA, 300–312. https://doi.org/10.1145/3202185.3202207

[14] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating Clarifying Questions for Information Retrieval. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 418–428. https://doi.org/10.1145/3366423.3380126

[15] Jie Zou, Mohammad Aliannejadi, Evangelos Kanoulas, Maria Soledad Pera, and Yiqun Liu. 2023. Users Meet Clarifying Questions: Toward a Better Understanding of User Interactions for Search Clarification. *ACM Trans. Inf. Syst.* 41, 1, Article 16 (jan 2023), 25 pages. https://doi.org/10.1145/3524110