Received November 29, 2021, accepted January 11, 2022, date of publication January 20, 2022, date of current version February 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3145323

Experimentation for Chatbot Usability Evaluation: A Secondary Study

RANCI REN[1], MIREYA ZAPATA[2], JOHN W. CASTRO[3], OSCAR DIESTE[4], AND SILVIA T. ACUÑA[1]

[1]Departamento de Ingeniería Informática, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Madrid, Spain
[2]Research Center of Mechatronics and Interactive Systems (MIST), Universidad Tecnológica Indoamérica, Quito 170103, Ecuador
[3]Departamento de Ingeniería Informática y Ciencias de la Computación, Universidad de Atacama, Copiapó 1532297, Chile
[4]Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Spain

Corresponding author: John W. Castro (john.castro@uda.cl)

This work was supported in part by the Spanish Ministry of Science, Innovation and Universities under Research Grant PGC2018-097265-B-I00, in part by the MASSIVE Project under Grant RTI2018-095255-B-I00, and in part by the Madrid Region Research and Development Program (Project FORTE, P2018/TCS-4314).

ABSTRACT Interest in chatbot development is on the rise. As a usability evaluation is an essential step in chatbot development, the number of experimental studies on chatbot usability has grown as well. As a result, we think a systematic mapping study is opportune. We analyzed more than 700 sources and retrieved 28 primary studies. By aggregating the research questions and examining the characteristics and metrics used to evaluate the usability of chatbots in experiments, it is possible to identify the state of the art in chatbot usability experimentation. We conducted a systematic mapping study to identify the research questions, characteristics, and metrics used to evaluate the usability of chatbots in experiments. Most experiments adopted a within-subjects design. On the other hand, few experiments provided raw data, and only one of the identified papers was part of a family of experiments. Effectiveness, efficiency, and satisfaction are usability characteristics used to identify how well users can learn and use chatbots to achieve their goals and how satisfied users are during the interaction. Generally, the experimental results revealed that chatbots have several advantages (e.g., they provide a real-time response and they improve ease of use) and some shortcomings (e.g., natural language processing, which is rated as the weakness most in need of improvement). This research offers an overview of chatbot usability experimentation. The increasing interest in this area is very recent, as works did not start to be published until 2018. Chatbot usability experiments should be more replicable to improve the reliability and transparency of the experimental results.

INDEX TERMS Usability, chatbots, experiments, family of experiments, systematic mapping study.

I. INTRODUCTION

A chatbot, also known as chatterbot, is domain-specific text-based software that supports human users with specific services [1], [2]. Joseph Weizenbaum developed the first dialog system (ELIZA) in the 1960s. ELIZA is considered to be the first chatbot [3]. Remarkable advances in deep learning, natural language (NL) processing, and machine learning are causing a seismic shift. Thus, chatbots are now better at interpreting a natural language phrase by the user and sending back the response in a similar way to users [4]. In turn, this has created unlimited possibilities and productive and useful experiences based on chatbots that can access and interact with digital services in many different applications [3], [5], [6].

In the current on-demand, real-time world, users expect the information they want to be only a click away. Chatbots have always played the role of information or service provider, especially in the e-commerce business world [4]. In recent years, chatbots have been pervasive, as e-commerce demand (e.g., online consulting, online payment) has grown and barriers to chatbot creation (like advanced technical expertise) have receded. People can create their own chatbot on social media platforms like Facebook Messenger, Twitter, and WeChat without sophisticated programming knowledge and other highly specialized technical skills. Some sites, like ChatBot (chatbot.com) or appypie (appypie.com/chatbot/builder), help novices develop simple chatbots using drag-and-drop interfaces.

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna DUlizia.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

Currently, chatbots are being applied in a range of different contexts, where they (i) help TV viewers interact with their TVs [7], (ii) recommend music [8], and (iii) perform collaborative modeling as part of the software development process [9]. In other words, chatbots have more or less infinite applications in many fields aimed at facilitating interaction. We are interested in exploring the state of the art of experimentation to evaluate chatbot usability. Therefore, our research does not place any limits on the context of use, since chatbots can potentially exist in any area.

Not all users are ready to place their trust in chatbots in preference to other communication channels, like email, due to their perceived poor understanding and quality of response [10]. In this context, the chatbot is still far from reading users' minds. Therefore, it is necessary for better integration between usability evaluation and the chatbot [11].

Usability evaluation, a growing field that is still being defined, refers to how well users can learn and use software to meet their requirements and addresses how satisfied users are during the process [12]. In software engineering (SE), usability is commonly considered to be one of many non-functional requirements and quality characteristics [13], where it has come to be recognized as a crucial tool for success in the competitive commercial world [14]. The right choice of evaluation methodology must be applied for the current research question or issue [12]. Apparently, chatbot usability evaluation is not yet a mature field [11].

In general, a chatbot usability evaluation learns and borrows experience from experimentation in software engineering (ESE). In order to explore chatbot usability experimentation, we conducted a preliminary survey, which failed to find any previous studies or literature reviews providing a consolidated view. As a result, we conducted a systematic mapping study (SMS) with the aim of: (i) exploring the state-of-the-art on chatbot usability experimentation, (ii) identifying the research questions that were investigated in experiments about chatbot usability, and (iii) defining the metrics used in experiments to measure chatbot usability in SE. Finally, our findings address the research questions and topics raised in this research in order to pinpoint the topics requiring future work. This research provides an informative review of the status quo of chatbot usability experimentation. Our contribution is designed to provide a map of everything that has been published, since we included all reported references in the literature of our SMS on chatbot usability experimentation. This map includes the usability characteristics used to measure the results and the categorization of the metrics used to evaluate the experimental results, the sample size of the experiment, the types of subjects participating in the experiment, the experimental design and procedure, the implemented tasks of the experiment, measurement instruments and statistical techniques, as well as any replications carried out. With this information, researchers interested in conducting experiments and/or replications related to chatbot usability will have access to a baseline accounting for all the aspects that they should consider (such as experimental design). Our research is a practical step towards a better understanding of chatbot usability experimentation, and its primary audience is researchers in the areas of human-computer interaction (HCI), SE, and chatbot development.

The paper is organized as follows. Section 2 outlines the main concepts of usability, and related work about chatbot usability evaluations and families of experiments. In Section 3, we explain the research method, the research questions of our study and the search strategies that were used in this article. In Section 4, we present the answers to each of the research questions. Section 5 provides a discussion of the results. We discuss the threats to validity of this study in Section 6. Finally, we outline the conclusions of our study and future work in Section 7.

II. BACKGROUND

To conduct the SMS, we referred back to a baseline study [11]. Ren *et al.* [11] found that chatbots and their respective usability evaluation were popular topics by 2015. This was when the number of publications started to grow, and many articles have been published every year since then. However, findings with respect to the ideal usability experiment were inconclusive in [11].

A. CHATBOT USABILITY EVALUATIONS

Usability is a common concern in SE. The International Organization for Standardization ISO 9241-11:1998 put forward a generalized definition of usability as *"the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use"* [15]. The ISO/IEC 25010:2011 software quality model categorized usability as a sub-characteristic of system/software product quality properties. It was defined as a subset of quality in use consisting of effectiveness, efficiency, and satisfaction, for consistency with its established meaning [16].

Below we review and explore the differences between the five existing research papers on the usability aspects of chatbots [17]–[22] and our research.

Abd-Alrazaq *et al.* [17] discussed the technical metrics used to evaluate healthcare chatbots. They started by describing the 65 studies that they included addressing detailed features (e.g., study design). They then categorized the technical metrics used to measure healthcare chatbots into four groups: global metrics, metrics related to response generation, metrics related to response understanding, and metrics related to esthetics. Their scoping review findings show that usability is the most commonly assessed aspect of healthcare chatbots.

Hobert [18] conducted a literature review to investigate which methods are suited for evaluating pedagogical chatbots in interdisciplinary research domains. While they declared 25 papers as the case base, they did not detail the number of papers that were left at the end of each screening process. Their findings revealed that many different evaluation approaches were adopted in research on educational conversational agents. Besides, they pointed out that researchers tended to analyze specific aspects in terms of their discipline. They concluded that comprehensive evaluations that analyzed pedagogical conversational agents from different perspectives were usually missing. They suggested that future research needs to test whether evaluating multiple goals in one research study is practicable and provides adequate contributions.

Rapp *et al.* [19] conducted a systematic literature review on how users interact with text-based chatbots. After applying the grounded theory literature review method on three electronic databases, they included 83 studies published from 2010 to 2020 in the review. They firstly discussed the features of the identified research in terms of publication venues, research methodologies, and chatbot characteristics. They concluded that experiments were the most commonly used methods to evaluate text-based chatbots. Secondly, they described the themes (e.g., chatbot user experience (UX)) and sub-themes. Note that although usability and UX have a lot in common, UX may be seen as an extension of usability. They singled out five main research themes in 83 studies. They then analyzed and identified the main methodological methods used in these papers.

Following the phenomenological method and Cooper's taxonomy of literature reviews, Tariverdiyeva and Borsci [20] conducted a literature review to explore aspects that influence a user's perception of the chatbot. They proposed a list of 27 key factors that affect users' perceptions of usability, including, but not limited to, response time, perceived ease of use, user privacy and ethical decision making.

To comprehensively review the quality attributes of chatbots and identify appropriate quality assurance approaches, Radziwill and Benton [21] conducted a literature review covering 32 conference papers and 10 journal articles. They outlined the quality attributes organized in terms of efficiency, effectiveness and satisfaction according to the concept of usability. Based on the analytic hierarchy process, they then synthesized the approaches across primary studies to recommend a compound technique.

In the knowledge that voice-based conversational agents had advanced over recent years and voice-related publications had increased correspondingly over the last 5 years, Seaborn and Urakami [22] conducted a rapid review of quantitatively measured voice-based system (including chatbot) UX through experiments. After reviewing the published full user studies based on ACM Digital Library and IEEE Xplore databases, they analysed independent variables (IVs), dependent variables (DVs) and the relationship between the IVs and DVs. They found that there is little consensus, and most user studies are lab-based studies. They also found that many studies adopted and focused on usability measurements, and usability is well-represented in both IV categories and DV categories. In view of their findings, they concluded that there is a solid foundation of usability research, and voice-based systems appear to satisfy basic usability criteria.

Taken together, a growing body of literature has investigated chatbot usability, including investigations on a specific types of chatbots [17]–[19], a qualitative study of critical factors that affect users' perception of chatbots [20], a discussion on chatbot quality attributes [21], and a quantitative user experience research on voice-based chatbots [22]. To the best of our knowledge, however, there is no work specifically investigating chatbot usability experimentation. Therefore, our research should fill this gap.

B. FAMILY OF EXPERIMENTS

As mentioned above, ESE plays a role in chatbot usability evaluation, and the experimental process could be used as a checklist and guideline. Once the experiment has been conceived, the general steps of the experimental process are: scoping, planning, operation, analysis and interpretation, presentation, and packaging, after which the chatbot usability experimental report can be drafted [23]. These steps were adopted in [9], for example, to evaluate a chatbot named SOCIO. SOCIO is a collaborative modeling tool to construct models or meta-models through social networks. This study employed a two-sequence, two-period within-subjects crossover design. The usability of chatbot SOCIO was determined by the attributes of efficiency, effectiveness, satisfaction, and the quality of the results. By comparison with Creately, a tool serving a similar purpose, the statistical results showed that chatbot SOCIO performance was superior in terms of efficiency and satisfaction and some aspects of diagram quality.

Nevertheless, the scientific community unanimously agrees that, with few exceptions, single experiments are of limited value. The accuracy of the baseline experiment results can only be established by replicating and contrasting results [24]. A family of experiments is a set of experimental replications where the experimental design and protocol is known. A family of experiments provides access to the data (raw or aggregated) for each experiment and contains at least three experiments with at least two different technologies testing the same response variable [25]. Families of experiments provide greater statistical power due to the higher number of experimental subjects [26].

More and more families of experiments are being run in SE [25]. As Basili *et al.* [27] stated in 1999, *"families of experiments refer to a group of experiments that pursue the same goal and build a body of knowledge by combining and generalizing the result"*.

Families of experiments are necessary to investigate the effects of alternative values for important attributes of the experimental models, vary the strategy with which detailed hypotheses are investigated, and make up for certain threats to validity that often arise in realistically designed experiments [26]. However, they are not infallible. SE families of experiments share common limitations: they tend to be comprised of fewer studies than those usually gathered in systematic literature reviews (SLRs) and usually study fewer response variables than SLRs, etc. [25].

In particular, families of experiments provide software engineering researchers with some advantages for evaluating the effectiveness of SE tools [28]–[32]: (i) families of experiments provide access to raw data so that researchers can apply consistent measurements and analysis techniques to analyze the experiments, and, hence, increase the statistical power of the findings; (ii) researchers conducting families of experiments may opt to reduce the number of changes made throughout the experiments, which can increase the internal validity of joint conclusions, and (iii) families increase the reliability of the findings, since joint conclusions are not affected by already published results. Due to the strengths of families of experiments, we pay special attention to the adoption of families of experiments in chatbot usability evaluation.

III. RESEARCH METHOD

The secondary study reported in this paper has been developed following the guidelines established by Kitchenham *et al.* [33] and Petersen *et al.* [34] to perform a literature review using a SMS in the fields of SE and HCI. To conduct the research, the first SMS phase is dedicated to identifying the need and corresponding databases for the review, including goals and research questions, and also the search strategy as detailed below.

A. OBJECTIVES AND RESEARCH QUESTIONS

A SMS in SE is a type of secondary study designed to give an overview of a research area by classifying and categorizing published research reports and results and providing a visual summary or map [34]. Since the field of our study is relatively unexplored, a SMS is a good option for this study [23].

The main objective of this study was to map the chatbot usability experiments with respect to aspects of publication status, investigated research questions and metrics measured in the experiments. This gave rise to the following research questions (RQ):

RQ1: What is the state of the art of chatbot usability experimentation?

RQ2: What research questions did chatbot usability experiments investigate?

However, experimental research in SE has not yet been standardized [35]. In view of this, we propose a third research question, namely:

RQ3: How do experiments evaluate chatbot usability?

B. SEARCH STRING SELECTION

We identified the search string keywords as part of a previous study [11]. We ran a pilot study testing different combinations of keywords and analyzing the results for the different databases used. This study was defined in [11]. Finally, we selected the search string (see Table 1) that optimized both the quantity of hits and the share of each database in the process.

TABLE 1. Selected keywords.

Keywords		
"usability" **OR** "usability techniques" **OR** "usability practice" **OR** "user interaction" **OR** "user experience"	**AND**	"chatbots" **OR** "chatbots development" **OR** "conversational agents" **OR** "chatterbot" **OR** "artificial conversational entity" **OR** "mobile chatbots"

C. DATABASES AND SEARCH PROTOCOL

The IEEE Xplore, ACM Digital Library, SpringerLink, Scopus, and ScienceDirect academic databases (DBs) were used in the SMS process. Following the advice of Kitchenham *et al.* [33], we used more than one database to prevent any possible database-derived bias [36]. The search fields used were determined by the options provided by each DB. Table 2 summarizes the search fields used for each DB.

TABLE 2. Search fields by databases.

DBs	Search Fields
IEEE Xplore	"Abstract"
ACM Digital Library	"Abstract"
SpringerLink	"Title OR Abstract OR Keywords"
Scopus	"Title OR Abstract OR Keywords"
ScienceDirect	"Title OR Abstract OR Keywords"

The selection criteria used to retrieve the primary studies are summarized below.

Inclusion criteria:

- The abstract or title mentions an issue regarding chatbots and usability **OR**
- The abstract mentions an issue related to usability engineering or HCI techniques **OR**
- The abstract mentions an issue related to user experience **AND**
- The paper describes a chatbot usability experiment.

Exclusion criteria:

- The paper does not report an evaluation or an experiment related to chatbot usability **OR**
- The paper does not report any issue related to chatbots and usability **OR**
- The paper does not report any issue related to chatbots and user interaction **OR**
- The paper does not report any issue related to chatbots and user experience **OR**
- The paper is written in a language other than English.

D. SEARCH PROCESS

We reviewed papers about experiments describing chatbot usability published from January 2014 to June 2021. The search was conducted in three phases. The first search phase was run in October 2018, including papers published from January 2014 to October 2018. The second search phase was run in June 2020 and contained papers published from November 2018 to June 2020. The third search phase was run in June 2021 and contained papers published from

July 2020 to June 2021. Since most databases were not searchable based on post month, we searched based on post year (e.g., 2018 to 2020) during the second and third phases and then eliminated duplicate results with previous SMS. Additionally, we searched for publications in the tables of contents of the proceedings of HCI conferences and HCI journals from 2014 to 2021. We have uploaded the lists of HCI conferences and HCI journals that we searched to supplementary material (shorturl.at/dxMR5).

Once we had identified the search strings and defined the search fields (Table 2), we started our search process. A total of 718 papers (referred to as retrieved papers) were found in the different DBs, HCI conferences, HCI journals or were recommended by external HCI experts. In particular, external HCI experts recommended 5 journal articles, 10 conference papers and 8 papers, which account for the 23 papers from other sources.

Then, the duplicate papers were removed from the retrieved papers, 560 papers were filtered to the group of non-duplicate retrieved papers. A peer review was carried out on these 560 papers applying the inclusion and exclusion criteria to the title and abstract. Discrepancies were resolved through discussion. As a result, we identified 113 candidate papers.

To determine if the candidate papers were relevant to chatbot usability and the execution of chatbot usability experiments, we reviewed each candidate paper again using the inclusion and exclusion criteria. However, this time we read the papers in full (i.e., a full-text review). The results were cross-checked by two experts from the HCI and ESE fields.

Finally, a total of 28 were selected as the experiment papers used in this study. Of the experiment papers, seven were sourced from outside the database: two were retrieved from HCI conference, and five were recommended by external HCI experts (our search did not identify these papers due to the defined search strings). These papers were included as other sources in Figure 1 and Table 3. During the search process, we were not able to review one of the candidate papers. As it was not downloadable, it was discarded from this analysis. The results of the selection were assessed by two of the authors who are experts in HCI and ESE, and any disagreement was discussed and resolved. The steps for conducting the review are shown in Fig. 1. Table 3 reports the number of papers taken from each group: most experiment papers were taken from the Scopus database. The 28 experiment papers used in the analysis and extraction of the results are shown in Appendix A.

With the aim of solving disagreements between researchers in the primary study selection process, we evaluated inter-rater reliability by applying two assessments [37]: (i) percentage agreement [38], and (ii) Cohen's Kappa coefficient (k) [39]. For the first assessment, the observed percentage agreement was 87%, indicated by the total number of papers on which both researchers reached an agreement (488 papers), divided by the total number of reviewed papers (560 papers) (see Table 4). This is considered acceptable.

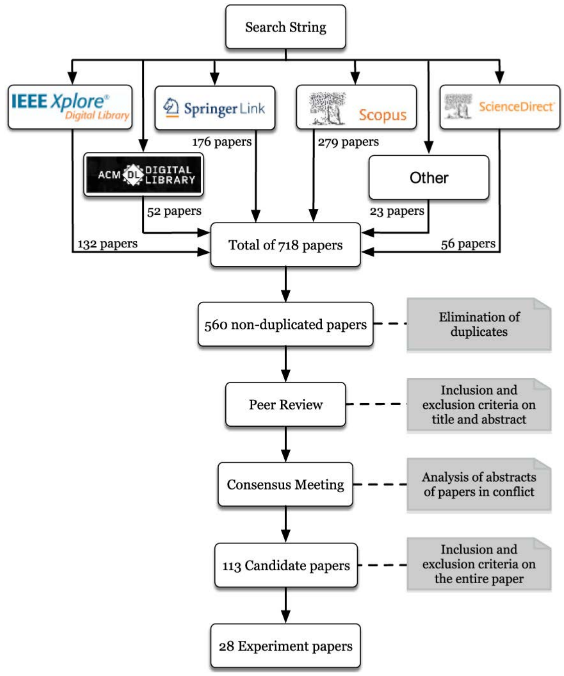

FIGURE 1. Diagram of the steps for the selection of experiment papers.

TABLE 3. Number of studies remaining after filtering the database results.

DBs	Retrieved	Non-Duplicate Retrieved	Candidates	Experiments
IEEE Xplore	132	90	15	2
ACM Digital Library	52	49	13	5
SpringerLink	176	148	14	2
Scopus	279	217	59	11
ScienceDirect	56	40	2	1
Other	23	16	10	7
Total	**718**	**560**	**113**	**28**

TABLE 4. Agreement matrix for nominal variable.

		Researcher 2		
		Accepted	Rejected	**Total**
Researcher 1	Accepted	104	28	**132**
	Rejected	44	384	**428**
	Total	**148**	**412**	**560**

For the second assessment, k = 0.66. According to [40], this is indicative of substantial agreement.

IV. RESULTS

This section reports the results of the SMS and responses to the research questions.

RQ1: What is the state of the art of chatbot usability experimentation?

To answer this research question, we analyzed 28 papers. They are mostly quantitative, and they performed controlled experiments by comparing chatbots with extended

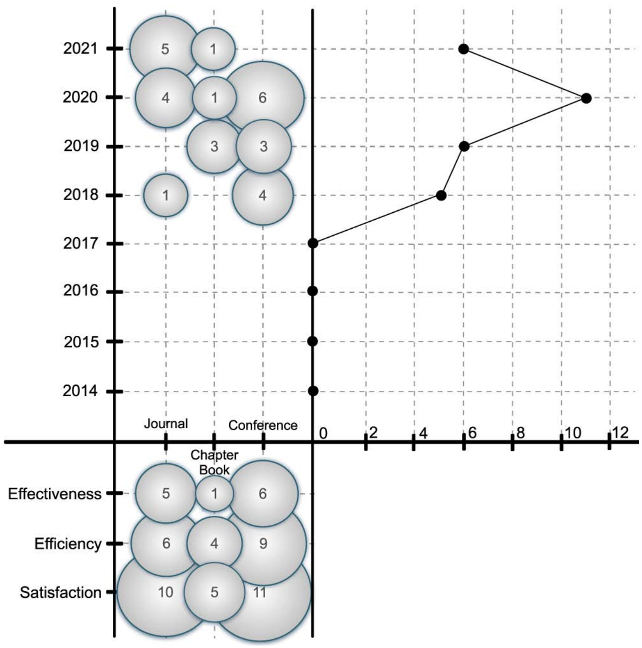

FIGURE 2. Mapping showing the experiment papers according to usability characteristics, including publication type and year.

versions of chatbots or other software with similar functions. Fig. 2 presents an overview of the identified primary studies.

As shown in Fig. 2, the results have been segmented into two areas. The left-hand side consists of two scatter (XY) plots (top and bottom) with bubbles at the junctions of the year-type of publication categories (top left-hand side) and usability characteristic-type of publication categories (bottom left-hand side). The types of publications were conferences, journals, and book chapters. The size of each bubble was determined by the number of experiment papers that had been classified into each category. The right-hand side of Fig. 2 illustrates the number of primary studies published per year. As the top right-hand side of Fig. 2 shows, the interest in chatbot usability experimentation is increasing and is very recent, with the earliest papers dating from 2018. Considering that the search end date was June 2021, the number of papers identified by our SMS for 2021 is rather high. Satisfaction is the most widely used usability characteristic (bottom left-hand side) as it was measured in each and every experiment. Note that the number of papers at the bottom of Fig. 2 does not match the number of papers at the top. The reason is that the same paper can discuss several usability characteristics.

Table 5 indicates the publication source of selected papers and type of publication (J = journal, C = conference, B = book chapter). In terms of the type of publication, 46.4% (13) of publications are conference papers, 35.7% (10) are journal articles, and 17.9% (5) are book chapters.

RQ2: What research questions did chatbot usability experiments investigate?

Table 6 summarizes the research objectives of the selected papers, including information like the references, the goals of the experiment, the stated or modified research questions and hypotheses of the experiment, the respective responses, whether the experimental raw data were provided and chatbot types.

Note that some papers defined the research question implicitly or stated multiple research questions. In the first case, we opted for the research question addressed by the

TABLE 5. Publication source.

Study	Publication Source	J	C	B
[PS1]	International Conference on Affective Computing and Intelligent Interaction (ACII)		X	
[PS2]	International Journal of Human Computer Interaction	X		
[PS3]	User Modeling and User-Adapted Interaction	X		
[PS4]	Lecture Notes in Electrical Engineering			X
[PS5]	Journal of Genetic Counseling	X		
[PS6]	Lecture Notes in Computer Science			X
[PS7]	Lecture Notes in Computer Science			X
[PS8]	Evaluation and Assessment in Software Engineering (EASE)		X	
[PS9]	Conference on Human Information Interaction and Retrieval (CHIIR)		X	
[PS10]	Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)		X	
[PS11]	Americas Conference on Information Systems: Digital Disruption (AMCIS)		X	
[PS12]	Conference on Human Factors in Computing Systems (CHI)		X	
[PS13]	International Conference on Intelligent User Interfaces Companion (IUI)		X	
[PS14]	Patient Education and Counseling	X		
[PS15]	IFIP TC13 Conference on Human-Computer Interaction (INTERACT) / Lecture Notes in Computer Science			X
[PS16]	Australian Conference on Human-Computer Interaction (OzCHI)		X	
[PS17]	International Conference on Information and Communication Technology Convergence (ICTC)		X	
[PS18]	SN Computer Science	X		
[PS19]	IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)		X	
[PS20]	International Journal of Human-Computer Studies	X		
[PS21]	International Journal of Medical Informatics	X		
[PS22]	Journal on Multimodal User Interfaces	X		
[PS23]	JMIR Medical Informatics	X		
[PS24]	Asia Pacific Journal of Information Systems	X		
[PS25]	Diversity, Divergence, Dialog / Lecture Notes in Computer Science			X
[PS26]	International Conference on Computing and Networking Technology (ICCNT)		X	
[PS27]	Conference on Human Factors in Computing Systems (CHI)		X	
[PS28]	Conference on Conversational User Interfaces (CUI)		X	

TABLE 6. Summary of research questions.

Primary Study	Goal	Research Questions	Hypothesis	Raw data	Type of chatbot
[PS1]	Compared with baseline conditions, they aim to evaluate the usability of the situationally aware adaptation of the chatbot with qualitative feedback during the real-world experiment.	(RQ1*) Compared with baseline conditions, does the chatbot with situationally aware adaptation have a positive impact on usability?	No	No	Situationally- and emotionally aware conversational agent
	Answers to research questions or hypothesis. In the first session, participants were exposed to four experimental conditions of adaptation (real-life situations in a simulated environment) while interacting with a conversational agent. In the second session, researchers used role-playing methodology for the user study. Researchers asked the participants to imagine that they are in a particular situation and interact with the chatbot in four different emotions. Finally, the researchers used the Affective Slider scale to assess pleasure and arousal and 5-point Likert scale questions to evaluate the overall user experience. In comparison to baseline conditions, the results show that the chatbot's situationally aware adaptation increases usability and elicits superior user experience, which is, in general, preferred by the users. However, it still needs to be improved to reach the level of personalization desired by the users.				
[PS2]	Compared with the remote-control unit (RCU), they aim to evaluate TV viewers' user experience (UX) of the conversational agent (CA)-assisted interactions while watching TV in terms of pragmatic quality (PQ), hedonic quality (HQ), and attractiveness (ATT).	(RQ2*) Compared with RCU, what is the TV viewers' UX of the CA-assisted interactions while watching TV in terms of PQ, HQ, and ATT?	No	No	Personal assistant when watching TV
	Answers to research questions or hypothesis. The researchers adopted physiological measurements for objective data and self-report questionnaires to comparatively analyze the user experience with the CA interface and RCU interface. The results of this study show that UX qualities vary according to interaction (CA and RCU). CA was inferior to RCU for all attributes except "human" in the pragmatic quality dimension. All hedonic qualities were higher in CA than in RCU. Attractiveness was also higher in CA than in RCU. In summary, the participants claimed that they felt that the CA's interactions were more interesting and attractive, although they felt frustrated and confused.				
[PS3]	They conducted an experiment to investigate the accuracy of conversational recommender systems (CoRSs) and compare the interaction efficiency of different interaction modes (completely natural language interface, completely button interface, and mixed interface).	(RQ3.1) To what extent can conversational interfaces support the music recommendation? (RQ3.2) What is the best conversational interface in terms of cost of interaction? (RQ3.3) What is the best conversational interface in terms of recommendation accuracy? (RQ3.4) Is the disambiguation step particularly strenuous for the user of a conversational music recommender system?	No	No	Recommender
	Answers to research questions or hypothesis. Responding to the first research question (RQ3.1), the researchers investigated the behavior of each interaction mode in terms of interaction time, time spent per question, query density, and the number of questions during the interaction. The result shows that the interaction cost is drastically reduced through an interaction mode completely based on natural language. They concluded that CoRSs is a valid alternative for a music recommender system. In response to the second research question (RQ3.2), the researchers determined that the best conversational interface in terms of cost of interaction is based on natural language supported by buttons when the user has to choose among multiple options. To answer the third research question (RQ3.3), the researchers adopted an accuracy ratio and mean average precision measure accuracy variable. The result identified that the NL-based interface supported by buttons has the best performance in terms of accuracy since users can express their preferences more effectively through the mixed interaction. To answer the fourth research question (RQ3.4), the researchers summarized that a particularly strenuous step for the user of a CoRS is confirmed to be disambiguation. When the user has to choose among multiple options, buttons associated with the different options make this task easier.				
[PS4]	They aim to examine the usability of the Tianmao jingling chatbot in terms of the effects of continuous conversation and task complexity on interaction with an AI-infused conversational agent in a simulated smart home environment.	(RQ4*) Does the continuous conversation (on vs. off) and task complexity (simple vs. complex) affect the usability of the Tianmao jingling chatbot?	No	No	Personal assistant in a smart home
	Answers to research questions or hypothesis. The participants were asked to complete four experimental tasks with continuous conversation on or off. These tasks were designed based on typical home environment activities in the use of conversational agents. The author administered a paper-based questionnaire to collect participants' responses to perceived system usability metrics to assess the usability. The researchers found that continuous dialog usually encouraged interaction between people, and as the continuous dialog was on, the number of queries per task would increase if the wake word could be eliminated when multiple commands were executed.				

usability experiment based on their experiments (identified in Table 6 with an asterisk). In the second case, we selected the research questions related only to usability.

The raw data (fifth column, Table 6) were poorly reported. We found that only one paper provided access to experiment raw data and three provided some of the experiment raw data as textual records. The chatbot types are listed in the sixth column of Table 6. We found that many chatbots are used in a number of real-life scenarios: 67.9% of the chatbots reported in our primary studies are deployed as personal assistants

TABLE 6. ***(Continued.)*** **Summary of research questions.**

Primary Study	Goal	Research Questions	Hypothesis	Raw data	Type of chatbot
[PS5]	This study evaluated the efficacy of a virtual conversational agent (VCA) interface, a new innovative approach for collecting data in health care, to collect family health histories (FHx) by comparing it with the standard interface.	(RQ5*) Compared with the standard interface, does the VCA interface positively affect the interface workload, usability, preference, and satisfaction when collecting FHx?	No	No	Personal assistant in the healthcare domain
	Answers to research questions or hypothesis. To address the research question, descriptive statistics were used to analyze the time taken to complete the task, the number of errors made, the results from the TAM scale, IBM-CSUQ scale, NASA-TLX test scores, the final preference questionnaire, and the think-aloud method was used to identify the most frequent and common responses by the participants. Through quantitative and qualitative feedback, participants reported lower levels of mental demand, temporal demand and effort and overall workload in completing the tasks using the VCA interface, while perceiving this interface as useful and easy to use. The researchers concluded that the VCA interfaces performed markedly better across all measures assessed except time.				
[PS6]	They compared mobile interfaces (Twitter and SMS) to each other and against a web-based embodied conversational (ECA) agent to assess the usability of a mentoring conversational agent.	(RQ6) How usable are the mobile conversational agent interfaces compared to the web-based ECA interface?	H.6 The mobile conversational agent interfaces will have better perceived usability than the web-based ECA.	No	Personal assistant
	Answers to research questions or hypothesis. Before the experiment was executed, the researchers introduced the conversational agent development process, including a virtual mentoring system and interfaces. Participants were instructed to use all three interfaces one by one, then to complete an online survey immediately after they were finished using an interface. User experience was assessed concerning usefulness, usability, credibility, desirability, accessibility, value, and a System Usability Scale (SUS) score. A repeated-measure multilevel linear model was used to compare the variables among the three interfaces. To answer the research question (RQ6), the researchers first explained demographic data and then pointed out that all three interfaces met the participants' expectations during their interaction. Although the belief that the ECA was valuable was the most frequent response, there was no apparent saturation of value for ECAs. The results show that the SMS and Twitter interfaces have a significantly higher mean score than the web-based ECA interface.				
[PS7]	They evaluate the user interaction when using the chatbot ROB to screen for symptoms associated with attention deficit hyperactivity disorder (ADHD) by comparing with an original paper version of the ADHD Self-Report Scale (ASRS).	(RQ7.1*) Compared to the original paper version of ASRS, does the use of a chatbot ROB positively affect user interaction quality with respect to the participants' scores on the ASRS and the time for completion? (RQ7.2*) What are the different user experiences between the two versions of ASRS?	No	No	Personal assistant in the healthcare domain
	Answers to research questions or hypothesis. In the analysis of the first research question (RQ7.1), they used the scores from the adult ASRS questionnaire and the time of completion for each screening. The results indicate that the average completion time of the paper version was faster than in the chatbot version. The mean sum scores of the chatbot were about the same as for the paper version. The researchers conducted a semi-structured interview to explore the participants' experiences with ROB and the original version (RQ7.2). Compared with the paper-based version, participants were satisfied when using the chatbot ROB because contextual information can easily be added to the response. Besides, most of the participants reported that they preferred the conversational interface over the paper version.				
[PS8]	To evaluate the usability of the chatbot SOCIO by comparing it to the web tool Creately with respect to effectiveness, efficiency, and satisfaction from the point of view of users, and the quality of the resulting class diagrams.	(RQ8) Compared to Creately, does the use of SOCIO positively affect the efficiency, effectiveness, and satisfaction of the users when making class diagrams, and the quality of class diagrams?	H.8.1 There is no difference in efficiency between SOCIO and Creately when making a class diagram. H.8.2 There is no difference in effectiveness between SOCIO and Creately when making a class diagram. H.8.3 There is no difference in satisfaction between SOCIO and Creately when making a class diagram. H.8.4 There is no difference in the quality of the class diagram made with SOCIO or Creately.	Yes	Collaborative tool
	Answers to research questions or hypothesis. In this study, they evaluated the usability of SOCIO based on four aspects: efficiency, effectiveness, satisfaction, and quality. Regarding efficiency, teams using SOCIO finished earlier than those using Creately. For collaboration, the fluency of the teams using SOCIO was high, and they had an interaction-cost advantage over those using Creately. For effectiveness, SOCIO and Creately performed similarly in terms of completeness. For satisfaction, SOCIO satisfied users to a greater extent than Creately with respect to the results of the adapted SUS score. More users expressed that they preferred SOCIO over Creately. On quality, SOCIO outperformed Creately in terms of precision, while solutions with Creately had better recall and perceived success. In summary, the usability of SOCIO had a positive effect on most aspects against the Creately baseline.				

TABLE 6. ***(Continued.)*** **Summary of research questions.**

Primary Study	Goal	Research Questions	Hypothesis	Raw data	Type of chatbot
[PS9]	Compared with a chatbot that communicates in modern English, the e-commerce chatbots apply a Shakespearean language style that affects customer experience and their attitude towards a presented product.	(RQ9) How does adding a language style to an e-commerce chatbot affect user satisfaction, user interest in a product, the perceived product value, and user engagement?	No	No	E-commerce chatbot
	Answers to research questions or hypothesis. This study applied and compared two chatbots to execute an experiment: one chatbot communicated in modern English (the baseline chatbot), whereas another chatbot communicated in a Shakespearean manner (the research chatbot). Subjects were randomly assigned one of the chatbots and asked to use the chatbot service to book a ticket for a play at a fictional Shakespeare theatre. Before and after the booking, the participants were asked to fill in a questionnaire. The researchers measured chatbots with the post-questionnaire on aspects of user satisfaction, user interest in a product, perceived product value, and user engagement. From the quantitative analysis, user satisfaction was found to be higher for the baseline chatbot, whereas user engagement was deemed to be higher for the Shakespeare chatbot. The qualitative analysis results showed that more users described the baseline chatbot as easy to use, while the Shakespearean chatbot was more often described as fun to use. Although the Shakespearean chatbot achieved a lower user rating, the ticket price that users stated they would pay was on average higher.				
[PS10]	Comparing two chatbot interfaces (Audio-only FarmChat and Audio+Text FarmChat) with differing interaction modalities to understand the usability of the FarmChat system.	(RQ10.1*) What is the acceptability level of FarmChat as an information system to satisfy farmers' information needs? (RQ10.2*) How is the usability in interacting with conversational interfaces? (RQ10.3*) What is the preference between the two variants of the conversational interfaces—Audio+Text versus Audio-only—and how does it differ among different user populations?	No	No	Personal assistant for farmers
	Answers to research questions or hypothesis. In this study, the researchers proposed a novel speech-based conversational system, FarmChat, to meet the low-literacy rural Indian farmer's information needs. To explore these research questions, they relied on demographic information, log data, Likert-scale ratings, user study notes by the study facilitator, and audio transcriptions of the user study and post-study interviews. Since the researchers found that farmers enjoyed using FarmChat because it provided immediate responses to their queries and constant access to farming-related knowledge, FarmChat was generally accepted by the farmers as an information source to satisfy their farming information needs (RQ10.1). Although it was the first time that any participants had interacted with a chatbot, they generally found the system to be usable (RQ10.2), as they gave relatively high Likert ratings. In the Audio+Text versus Audio-only comparison (RQ10.3), the results suggest that users' preferences were highly dependent on the participants' literacy level, digital-literacy level, and other individual factors, like profession, physical, and environmental factors. However, the majority of the users preferred the Audio-only interface.				
[PS11]	This research aims to study the differences in system satisfaction between a chatbot system and a website system and what factors determine satisfaction based on self-determination theory.	(RQ11.1) What factors affect system satisfaction in a chatbot system? (RQ11.2) Does the level of system satisfaction differ between a website system and a chatbot system?	H.11.1 Chatbots with a natural language processing interface will lead to a lower level of perceived autonomy/cognitive effort than websites with a menu-based interface. H.11.2 Perceived autonomy has a positive effect on perceived competence/performance satisfaction/process satisfaction. H.11.3 Cognitive effort has a negative effect on perceived competence/performance satisfaction/process satisfaction. H.11.4 Perceived competence is positively associated with process satisfaction. H.11.5 Performance satisfaction/process satisfaction has a positive effect on system satisfaction.	No	Personal assistant in tourism
	Answers to research questions or hypothesis. This experimental result has yet to be reported. This research applied self-determination theory and HCI literature to understand the factors influencing user satisfaction with natural language processing-based systems. The researchers hypothesized that a process research model, if supported, would reveal how factors, such as perceived autonomy, competence, and cognitive effort, jointly influence user satisfaction with the interaction process, task performance, and ultimately the systems itself.				

TABLE 6. *(Continued.)* **Summary of research questions.**

Primary Study	Goal	Research Questions	Hypothesis	Raw data	Type of chatbot
[PS12]	They measure and compare the Convey chatbot and the default chatbot in terms of the time taken to complete the task, the total number of words input by the user, and the total number of user actions (browsing and zooming in on a particular product).	(RQ12*) Compared with the default chatbot, does the Convey chatbot perform better in terms of the time taken to complete the task, the total number of words input by the user, and the total number of user actions (browsing and zooming in on a particular product)?	No	No	E-commerce chatbot
	Answers to research questions or hypothesis. In this paper, the researchers proposed a context view called Convey on top of the chatbot interface to help users understand the mental state of the chatbot during conversation (helping users and chatbot mesh) while sustaining the familiarity of the text-based messaging interface. To answer the research question (RQ12), they conducted a within-subjects user study with two treatments: default chatbot and Convey chatbot. After analyzing the survey results and log data, they found that participants preferred using the chatbot with Convey and found it easier to use, less mentally demanding, faster, and more intuitive than a default chatbot without Convey.				
[PS13]	They aim to examine how users' understanding affects perceptions and experiences of using a CA, specifically Apple Siri.	(RQ13*) To what extent does the personal experience of using a CA and the technical knowledge about a CA's system model affect how people feel about the CA?	No	No	Personal assistant (Apple Siri)
	Answers to research questions or hypothesis. Using two factors: (1) the personal experience of using a CA and (2) the technical knowledge about a CA's system model, the researchers conducted two-way ANOVAs on perceived usability measured by the Software Usability Measurement Inventory (SUMI) questionnaire. The results showed that inexperienced users with technical knowledge and experienced users without technical knowledge had more positive perceived usability. In addition, technical knowledge allowed the users to perform a user analysis of the CA for sense-making and adapt their use.				
[PS14]	The study assesses: 1) the effect of perceived similarity between the MyPAL robot and an avatar on children's friendship toward the avatar, and 2) the effect of this friendship on the usability of a self-management application containing the avatar and children's motivation to play with it.	(RQ14.1*) How does the MyPAL app perform on similarity, friendship, motivation, and usability? (RQ14.2*) What are the relationships between similarity, friendship, motivation, and usability with the MyPAL app performance?	No	No	Personal assistant in the healthcare domain
	Answers to research questions or hypothesis. The PAL project developed a conversational agent with a physical (robot) and a virtual (avatar) embodiment to support children's diabetes self-management ubiquitously. To answer the first research question (RQ14.1), the researchers analyzed data related to the variables, and they found that there was no significant increase in similarity, motivation, or usability scores from T1 to T2 and the SUS score indicated that the MyPAL app was rated as average. The participants' feelings of friendship toward the physical robot were significantly higher than the feeling of friendship toward the avatar. With regard to the second research question (RQ14.2), the results showed that there was an increase in friendship with the avatar as perceived similarity with the avatar increased; a positive effect of friendship with the avatar on motivation to play with the MyPAL app; and the feelings of friendship by the child toward the avatar are positively correlated with the usability of MyPAL. Overall, children stated that the physical robot was more (inter)active, more present, and capable of doing different things. Children felt stronger friendship towards the physical robot than toward the avatar. When children perceived the robot and its virtual counterpart as the same agency, they felt a stronger friendship with the avatar. In addition, the closer the friendship the children struck up with the avatar, the more user-friendly they perceived the MyPAL app to be, and the more they were motivated to play with it.				
[PS15]	They aim to ascertain whether the relatively inexpensive approach of using real-time head pose measurements as a proxy for user attention is a suitable alternative to using a wake-up word.	(RQ15) Is the advanced chatbot version that requires a head pose more usable and likable than the chatbot version that requires a wake-up word to signify that users are addressing the assistant?	No	No	Astrophysics assistant
	Answers to research questions or hypothesis. In this paper, the researchers explored an enhanced chatbot version that helps users explore data about exoplanets. To quantify the usability of the chatbot relative to that of an alternate variant of the chatbot that required a wake-up word (RQ15), they created two variants of the chatbot (Condition A: Users were required to use a wake-up word to signify that they were addressing the assistant; Condition B: Users merely needed to look at the display to signify that they were addressing the assistant) that were nearly identical to conduct a within-subjects crossover experiment. The findings showed that: (1) the head pose system (Condition B) is preferred; (2) on average, Condition B has better usability scores than Condition A, (3) the perceived discernment of the two variants was essentially the same, and adequate.				
[PS16]	They aim to explore how to communicate service offers as part of chatbot interaction, and user preference for such service offers.	(RQ16) How can service offers be communicated to users during conversational interactions at different levels of proactivity, and how are user preferences for such offers?	No	No	Personal assistant in financial services
	Answers to research questions or hypothesis. In this paper, the researchers proposed four approaches (the reactive approach, the intermediate–reactive approach, the intermediate–proactive approach and the proactive approach) to communicate available service offers, reflecting different levels of chatbot proactivity. To evaluate the user perception for these four approaches, they gathered feedback on user preference through interviews. The results showed that proactivity in the communication of service offers was found to be potentially valuable, provided that the offer is relevant to the conversation, does not compromise conversational efficiency, and is easy to discard. However, proactive communication of service offers may also entail challenges concerning perceptions of privacy and invasiveness, and, hence, needs to be designed with great care.				

TABLE 6. *(Continued.)* Summary of research questions.

Primary Study	Goal	Research Questions	Hypothesis	Raw data	Type of chatbot
[PS17]	They aim to analyze the SUS usability of the speech-only contexts Amazon Echo Dot (3rd generation), Apple HomePod, and Google Nest Mini compared with the graphical user interface (GUI) paradigm.	(RQ17) How does the SUS usability of speech-only contexts Amazon Echo Dot (3rd generation), Apple HomePod, and Google Nest Mini compare to the GUI paradigm?	No	No	Personal assistant (Amazon Echo Dot/Apple HomePod/Google Nest Mini)
	Answers to research questions or hypothesis. The primary research goal of this paper was to investigate if SUS is a valid tool for measuring usability for speech-based systems. As for the primary research goal, their results showed that the original SUS findings are less relevant in the present speech-only context. To address RQ17, they evaluated the SUS results in terms of distribution of SUS scores, psychometric properties and reliability analysis, and exploratory factor analysis. The researchers found that the speech-only context provides a more naturalistic and humanized environment than the GUI systems, where the smart-speakers themselves can support the users in the event of any problems.				
[PS18]	They aim to analyze the Voice Usability Scale (VUS) usability of the speech-only contexts Amazon Echo Dot (3rd generation).	(RQ18) What is the VUS usability of speech-only contexts Amazon Echo Dot (3rd generation) like?	No	No	Personal assistant (Amazon Echo Dot)
	Answers to research questions or hypothesis. This is an extension of [PS17]. The researchers developed VUS in line with SUS to account for the unique aspects of voice-based communication. Their primary twofold objective was to check the suitability of SUS for usability evaluation of voice-assistants and developing a subjective scale in line with SUS that considers the unique aspects of voice-based communication. With respect to their primary research goals, an exploratory factor analysis suggested that SUS has drawbacks for measuring voice usability. With respect to VUS, the most optimal factor structure identifies three main components: usability, effectiveness, and recognizability and visibility. To address RQ18, they evaluated the VUS results and SUS results in terms of tool usability. The finding suggests that both scales are reliable. The results of SUS had been explained in answers to the previous research question (RQ17). With regard to VUS, the tool was found to be acceptable in terms of usability, effectiveness, and recognizability and visibility.				
[PS19]	They aim to evaluate ConveRSE's ability to adapt to an interface based on a social humanoid chatbot in terms of both recommendation accuracy and user experience by comparing the smartphone-based and robot-based interfaces.	(RQ19) Can chatbot ConveRSE be implemented through a social chatbot without losing performance in terms of recommendation accuracy and user experience compared to a chatbot-based interface?	No	Partially	Recommender
	Answers to research questions or hypothesis. In this paper, the researchers introduced ConveRSE as a domain independent framework for developing conversational recommender systems. To investigate the possibility of using a humanoid robot as an interface for ConveRSE, they conducted a within-subjects experiment to evaluate ConveRSE's ability to adapt to an interface based on a social humanoid chatbot. By analyzing two types of questionnaires (ResQue and Weiss et al.'s models), the results provide evidence of ConveRSE's adaptability to a social chatbot-based interface, while preserving accuracy and guaranteeing a good user experience.				
[PS20]	They aim to transcribe the real conditions of interactions with a professional virtual agent to capture as accurately as possible the perceptions and usage behaviors of real users.	(RQ20) How would the expression of intimate behaviors by the chatbot impact the users' perception of virtual intimacy, social presence, and user experience in a real-world situation?	H.20.1 Social presence mediates the effect of perceived virtual intimacy on the user experience. H.20.2 Based on the components of user experience model framework, the perceived virtual intimacy of product perception on user experience has a direct impact on the emotion of user experience.	Partially	Personal assistant in tourism
	Answers to research questions or hypothesis. In this paper, the researchers introduced and used a chatbot that is an expert in tourism and able to express intimacy-related behaviors in verbal and nonverbal communication to evaluate if intimate behaviors would impact the user experience. To answer the research question (RQ20) and test the hypotheses (H.20.1-H.20.2), they conducted an interactive experiment in field conditions in which real tourists interacted with a social virtual counselor. They found that the participants behaved more sociably toward the intimate agent and perceived its honesty and genuineness. Nevertheless, the user experience is not significantly influenced by intimate expression.				
[PS21]	This study aimed to investigate the usability of a short-term mobile-based interactive chatbot Todaki in alleviating attention deficit symptoms.	Not defined	H.21 The chatbot Todaki format would report a comparable acceptability compared to books.	No	Personal assistant in healthcare domain
	Answers to research questions or hypothesis. In this paper, the researchers conducted a randomized, non-blind parallel-group pilot study to evaluate whether a chatbot called Todaki is capable of lowering participants' attention deficit symptoms and improving its usability. By comparing the results for the baseline group (use chatbot only) and control group (read a self-help guidebook only), they specifically evaluated the metrics of the measurement effectiveness and satisfaction. The results show that the chatbot Todaki was effective enough to reduce the overall symptoms related to attention deficits, albeit with less subjective satisfaction of the users.				

TABLE 6. *(Continued.)* Summary of research questions.

Primary Study	Goal	Research Questions	Hypothesis	Raw data	Type of chatbot
[PS22]	They aim to explore the usability and acceptability of chatbot DynamicDuo in both controlled laboratory-based studies and real-world environments.	Not defined	H.22 Presenters will accept a chatbot as a co-presenter for scientific presentations.	No	Personal assistant in presentation
	Answers to research questions or hypothesis. To test the hypothesis (H.22), researchers conducted usability evaluations on chatbot DynamicDuo in both the laboratory and classroom environments. Under the lab-based content, the result shows that the chatbot presenter was rated as being satisfying to work with. By comparing the results for a given class with or without the chatbot DynamicDuo in classroom evaluation, researchers found that students who co-presented with the chatbot expressed high levels of satisfaction and desire to use the chatbot for future presentations. In conclusion, they determined that presenters in both contexts accepted a virtual agent as a co-presenter for scientific presentations, rating it high on measures of satisfaction.				
[PS23]	The purpose of the experiment is to measure perinatal women's and their partners' perceptions of the utilitarian and hedonic value of medical chatbot Dr. Joy experience.	Not defined	H.23 The chatbot Dr. Joy will produce both utilitarian and hedonic value during the 7-day contextual usability testing period.	Partially	Personal assistant in healthcare domain
	Answers to research questions or hypothesis. To address H.23, the researchers collected and analyzed quantitative data (answer to closed-ended usefulness, satisfaction, and ease of use (USE) questions) and qualitative data (user utterance data and responses to open-ended questions). According to the results of the USE questionnaire, it was found that, in this sample, the mean score of ease of learning was the highest, followed by the ease of use, satisfaction, and usefulness scores. As reflected in the responses to the open-ended question about the strengths of Dr. Joy, participants highlighted not only the hedonic value as represented by fun, pleasure, and enjoyment, but also the utilitarian value as represented by usefulness, speed, ease of use, and convenience. The most frequently reported weak point was that Dr. Joy failed to meet all user intents and to cover a much broader range of content domains.				
[PS24]	The purpose of the study is to compare rule-based and natural language processing-based chatbots in terms of usefulness, usability, searchability, reliability and attractiveness.	(RQ24) Compared to rule-based chatbot Talkjipsa, does natural language processing-based chatbot Samantha perform better in terms of usefulness, usability, searchability, reliability and attractiveness?	No	No	Personal assistant
	Answers to research questions or hypothesis. In this paper, researchers introduced two similar-function chatbots used for mobile shopping: a rule-based chatbot Talkjipsa and a natural language processing-based chatbot Samantha. To answer the research question (RQ24), researchers evaluated and compared Talkjipsa and Samantha in terms of usefulness, usability, searchability, reliability and attractiveness. The results indicate that the rule-based chatbot was superior on searchability and reliability, whereas the natural language processing-based chatbot was superior on usefulness and usability. The difference in the scores for attractiveness were not as big as for the other variables. For both chatbots, the level of reuse intention and recommendation for others stated by participants were not as high as expected.				
[PS25]	This study aims to compare the conversational search user interface (chatbot) of a medical resource center database with its graphical search user interface in terms of user engagement and usability.	(RQ25) How does a conversational search interface compare to a graphical search user interface in terms of user engagement and usability?	H.25 The usage of the chatbot for searching has a positive effect on user engagement.	No	Medical resource center chatbot
	Answers to research questions or hypothesis. To answer RQ25, researchers evaluated and compared the usability of a chatbot and a graphical search user interface (SUI) by: (1) collecting quantitative measures of time, task, task success and overall preference, and (2) conducting a thematic analysis of the qualitative data shared by the participants during the experiment and in a post-study questionnaire. The results show that: (1) there was no substantial evidence to say that the usage of the chatbot results in higher user engagement; (2) on average, participants took approximately two minutes longer to complete the tasks with the chatbot than with the website, and (3) there were three instances where the user was not able to successfully complete the task. These findings indicate that the conversational search interfaces need flawless usability to: (1) be able to provide additional value in information retrieval tasks, and (2) elicit a higher level of engagement compared to their SUI-based counterparts.				
[PS26]	This study aims to compare the usability, usage patterns, and the overall satisfaction received after using the voice assistants (VA) between secondary English language speakers and native English speakers.	(RQ26.1*) What is the experience of secondary English speakers interacting with VA? (RQ26.2*) Are there any differences in the usage patterns and overall perceived satisfaction between native English speakers and secondary English speakers? (RQ26.3*) How does the usability of VA affect secondary English speakers?	No	No	Personal assistant (Amazon Echo Dot/Apple HomePod)
	Answers to research questions or hypothesis. To answer the research questions, researchers evaluated the usability, usage patterns, and the overall perceived satisfaction after using the voice assistants (VA) by secondary and native English speakers. They first conducted a survey (user online questionnaire) with commercially available VA users, including about 45% native English speakers and 55% non-native speakers. Then, they conducted a usability experiment by asking participants to complete 10 frequent tasks with Amazon Echo and Apple HomePod in a home environment. Their survey findings showed that there is no statistically significant difference in the way the two groups of users interact with the VAs and their overall usability. However, their experimental results show that limited non-English language support is a major drawback and a greater cause of concern than English accent while interacting with the VAs.				

TABLE 6. *(Continued.)* **Summary of research questions.**

Primary Study	Goal	Research Questions	Hypothesis	Raw data	Type of chatbot
[PS27]	This study aims to understand what kind of response style has more positive effects on users' evaluations of the agents.	(RQ27.1) How would the three different styles of responses (i.e., avoidance, empathy, and counterattacking) by the agents to users' verbal abuse influence the users' perceptions of the agents' capability? (RQ27.2) How would the three different types of verbal abuse (i.e., insults, threats, and swearing) that users employ influence the users' perceptions of the agents' capability?	No	No	E-commerce chatbot
	Answers to research questions or hypothesis. To answer the research question, they designed an experiment by manipulating 3 verbal abuse types (insults, threats, swearing) as levels of a between-subject factor and 3 agent response styles (avoidance, empathy, counterattacking) as levels of a within-subject factor. Regarding the agent response styles, the results showed that the users felt a more guilty if the agents adopted an empathetic attitude; user assessments of counterattacking agents were conflicting; avoidance was mostly evaluated as harmful. However, verbal abuse type had no significant impact on user perception measurements. In conclusion, user perceptions of agent capabilities were strongly influenced by the response style of the agent rather than the verbal abuse type.				
[PS28]	This study aims to quantitatively compare the cognitive workload and linguistic properties of native (L1) and non-native (L2) English speakers in their interaction with intelligent personal assistants (IPAs).	Not defined	H.28 L2 speakers are likely to experience significantly higher mental workload in IPA interaction compared to L1 speakers.	No	Personal assistant (Google Assistant)
	Answers to research questions or hypothesis. To investigate the hypothesis, the authors assessed the mental effort and task completion of two speaker groups, L1 and L2 English speakers, by asking participants to interact with Google Assistant on two different devices. They found significant differences between the two speaker groups in terms of cognitive demands. Specifically, they found that L2 speakers had a significantly higher mental workload than L1 speakers using both smart speakers and smartphone devices.				

[PS2], [PS4], [PS5], [PS6], [PS7], [PS10], [PS11], [PS13], [PS14], [PS16], [PS17], [PS18], [PS20], [PS21], [PS22], [PS23], [PS24], [PS26], [PS28] and more and more chatbots are being deployed in the healthcare domain [PS5], [PS7], [PS14], [PS21], [PS23], [PS25].

Chatbots have the potential to be at the patients' side anytime and anywhere, which is, obviously, out of the question for doctors and care workers, leading researchers to develop chatbots to support healthcare. Even though the demand for measuring the performance of healthcare chatbots is increasing, the evaluation methods for healthcare chatbots appear to be wide-ranging and arbitrary [17].

Additionally, some chatbots act as e-commerce tools [PS9], [PS12], [PS27], collaborative tools [PS8], emotionally aware conversational agents [PS1], astrophysics assistants [PS15], tourist guides [PS11], [PS20] and recommenders [PS3], [PS19].

RQ3: How do experiments evaluate chatbot usability?

From the perspective of HCI, various usability techniques were employed in these experiments, and it is patent that the most employed usability technique was questionnaires, followed by interviews (Table 7).

Compared with [11], we found that, on top of SUS and ad-hoc methods, a broader range of questionnaires were adopted to investigate chatbot usability. In [PS2], the AttrakDiff2 questionnaire was used, which measures how attractive a product is based on its hedonic and pragmatic qualities. The Likert scale was the most used metric in the questionnaires across the whole range of papers [PS3], [PS6], [PS9], [PS12], [PS21] and [PS23].

TABLE 7. **Usability techniques.**

Usability Techniques	Experiments
Questionnaire	[PS1], [PS2], [PS3], [PS4], [PS6], [PS7], [PS8], [PS9], [PS11], [PS12], [PS13], [PS14], [PS15], [PS16], [PS17], [PS18], [PS19], [PS20], [PS22], [PS23], [PS24], [PS25], [PS26], [PS27], [PS28]
Interview	[PS1], [PS7], [PS10], [PS13], [PS15], [PS16], [PS22], [PS28]
Think-aloud	[PS5], [PS13], [PS19], [PS25]
Direct observation	[PS5], [PS15]

Throughout the usability evaluation process, pre-test and post-test questionnaires were combined for use in [PS5] and [PS10] in order to round out the result of evaluation with demographic information. We also noticed that papers seldom discuss the rationale used to select the technique. It should be noted that the selected technique may have an impact on the effectiveness and reliability of the experimental result.

The columns of Table 8 show the metrics used to evaluate the experiment results, specifying whether the results correspond to a family of experiments (F = "Is the experimentation composed of a family of experiments?"), the number of experiments (ES = experiment size), the experiment sample size (SS = sample size), the types of subjects participating in the experiments (TS = type of subjects), experimental design and procedure, the implemented tasks of the experiment, usability characteristics used to measure the results, measurement instruments (MI), and statistical technique (ST).

Our topic cuts across the fields of HCI and ESE. Therefore, we considered the indicators to measure the experiment from both sides, as the software development process is very dependent on the defined tasks and user skills and characteristics [13], and the task and users matter to the HCI community.

We also observed a growing interest in experimentation and have taken note of recent calls for replication in SE [24]. Thus, we considered investigating replication in chatbot usability experimentation. We mainly followed the reporting structure for SE experiment reports proposed by Jedlitschka and Pfahl [41]. As the defined tasks and user skills and characteristics have a profound impact on the software development process [13], the task and users matter in HCI. For the above related reasons, we decided to use the indicators shown in Table 8 to measure each experiment.

We noticed that chatbot developers always acted as evaluators in these experiments. Only six experiments were conducted by third-party researchers or experts who evaluated the usability of the chatbots [PS8], [PS13], [PS17], [PS18], [PS26], [PS28].

A. THE REPLICATION OF EXPERIMENTS

Of the usability experiments that we reviewed, there is only one study [PS6] that conducted replications of an experiment with a consistent experimental design but different participant region or background. We consider the study reported by Huff-Jr *et al.* as a family of experiments, which uses a within-subjects mixed-method design [PS6] using qualitative contents and a multilevel linear model to analyze data. The total sample size of the replication was 35, although the authors did not report the respective sample size of each replication.

To the best of our knowledge, a family of experiments should include at least three experiments [25], whereas [PS6] replicates a single experiment—that is, this paper reports a set of two experiments. However, since two experiments can aggregate the data to evaluate the effect of chatbots, we classified the two experiments as a family of experiments.

It should be noted that there is a study [PS22] that conducted two different experiments in controlled laboratory-based and real-world environments to comprehensively evaluate the usability of their chatbot. Since the experimental designs are different, we do not consider this study to be a family of experiments.

B. SAMPLE SIZES

Regarding the sample size of experiments (fourth column of Table 8), although we acknowledge that the sample size varies for different usage and developmental phases, the sample sizes of published usability experiments for chatbots are relatively small. Of the experiments, 42.9% (12) included fewer than 30 subjects, 42.9% (12) included between 30 subjects and 80 subjects, and 10.7% (3) contained more than 90 but fewer than 500 subjects. One experiment [PS11] did not detail the sample size.

C. TYPES OF SUBJECTS

In terms of the types of subjects involved in experiments (fifth column of Table 8), 35.7% (10) of the experiments included students, while most of the researchers placed no constraint on academic background and academic program. The remaining experiments included experienced users or experts, company employees, farmers, children, residents, and patients. However, 25% (7) of experiments did not define the subject types. Only two studies compared groups: graduates versus undergraduates [PS8] and native vs. non-native English-speakers [PS26].

D. EXPERIMENTAL DESIGN AND PROCEDURE

Regarding the experimental design and procedure, 53.6% (15) were defined as within-subjects experiments. As the sample sizes of the identified experiments are relatively small, the within-subjects design has better statistical power since it doubles the data points. In SE, experimental design plays a role in controlling for extraneous variables: mature experiments are run with pre-established protocols defining the experimental settings and the set of procedures that must be strictly adhered to during the execution and analysis of the experiments. By contrast, many chatbot usability experiments are set up without any a priori plan or experimental design definition.

Furthermore, prior experience and technical knowledge have an impact on the global usability of conversational agents [PS13], [PS26], while some experiments [PS1] did not appear to measure the pre-user experience or knowledge related to chatbots.

Generally, chatbot usability was rated positively in most experiments, while only one chatbot was given a negative evaluation compared with the control tool [PS6]. Despite this, it was pointed out that chatbots still need to be improved in some respects. The NL interaction was the most frequently mentioned improvement within these experimental results.

The result for [PS3] shows that the performance of NL interaction with the chatbot CoRS is poorer than the button and mixed interfaces. In [PS8], several participants suggested an improvement in natural language processing (NLP) as the chatbot SOCIO does not understand some phrases. Aside from these, researchers also suggested voice-based natural language recognition should be improved to support varieties of English accents [PS26].

Besides, chatbot personalization does not always satisfy all users and experts. In [PS1], the users commented that the automatic adaptation strategies need to be further improved to reach the level of personalization desired by the users compared to manual adaptation.

There are some other problems that remain. In [PS8], the control outperforms the chatbot SOCIO in terms of recall and perceived success. A Shakespearean-styled chatbot increased user engagement as well as perceived product value, but user satisfaction decreased [PS9]. As for chatbot use for online shopping, the researchers found that the participants' expressed re-use intentions and the level of recommendation to others were not as high as expected [PS24].

TABLE 8. All measured metrics of experiments.

Primary Study	F	ES	SS	TS	Design and Procedure	Tasks	Usability Characteristics	MI	ST
[PS1]	No	1	12	Experienced users	Design: Pair-wise comparison with counter-balanced Latin square. Procedure: Participants were exposed to four experimental conditions of adaptation (real-life situations in a simulated environment) one by one. These experimental conditions were: no adaptation, random adaptation, auto adaptation, and manual adaptation. In each context, the experimenters asked the participants to interact with the agent with respect to four different emotions. Finally, participants evaluated the condition of system adaptation and usability with a questionnaire and qualitative interviews.	Participants were required to perform four tasks when exposed to the four experimental conditions. They were asked to imagine that they were in a specific situation (e.g., at home, at a public cafe, alone, in a group). In each context, the experimenters asked the participants to interact with the agent using four different emotions (neutral, upset, happy, and angry, which were simulated using pre-populated questions). In total, each participant had 16 interactions with the agent.	<u>Satisfaction:</u> Pleasure	Questionnaire, interview	Counting the measured values, Wilcoxon rank sum test
[PS2]	No	1	42	Students	Design: Within-subjects crossover design. Procedure: The experiment was first explained to the participants who then gave their signed consent to participate in the experiment. The participants were required to use each interface one by one. To perform the task, sensors were attached to collect physiological response data. The control method for each interface was then studied and preliminary data were evaluated. Once the participants finished each task, they subjectively rated the AttrakDiff2 items addressing their own feelings.	Participants were required to perform a set of 13 tasks. Task 1: Turn on the TV Task 2: Volume up Task 3: Volume down Task 4: Channel zapping to near channel Task 5: Channel zapping to remote channel Task 6: Menu navigation Task 7: Search & play Task 8: VOD control Task 9: VOD play off Task 10: Search & navigation Task 11: Menu navigation Task 12: Volume control Task 13: Turn off the TV.	<u>Effectiveness:</u> Task completion <u>Efficiency:</u> Task completion time <u>Satisfaction:</u> Pragmatic quality, hedonic quality, attractiveness	Questionnaire, software platform	Counting the measured values, paired *t*-test, linear regression, correlation analysis

TABLE 8. *(Continued.)* **All measured metrics of experiments.**

Primary Study	F	ES	SS	TS	Design and Procedure	Tasks	Usability Characteristics	MI	ST
[PS3]	No	1	110	Students	Design: Within-subjects design. Procedure: After taking the training phase where users freely interacted with the system and became familiar with the different interfaces, the experiment started. Firstly, the system provided an introduction to the experiment, and then the participants were required to perform the task. After they finished the task, they were asked to complete the questionnaire.	Participants were required to perform one task: the system proposed a list of five songs. For each song, users gave feedback (like, dislike, or skip the recommendation without providing feedback). Optionally, users could ask for an explanation of the recommended item. The feedback could also be related to a specific property of the song (e.g., the singer, the producer, the songwriter, the genre). When users enjoyed the first set of recommendations, they could decide to request a new set of recommended songs or stop the experiment.	Effectiveness: Accuracy, precision Efficiency: Task completion time, mental effort, communication effort Satisfaction: Ease-of-use, complexity control, pleasure, want to use again, learnability, adaptability	Questionnaire, software platform	Counting the measured values, MANOVA statistical test, Wilcoxon test
[PS4]	No	1	18	Company employees	Design: Two-factor within-subjects design. Procedure: The written consent and detailed introduction of test procedures were provided to participants before the experiment. Participants were instructed to get familiar with the conversational agent and smart home devices and practice orders that should be used to control smart devices; they were then asked to perform four experimental tasks and complete the questionnaire.	Each participant was required to perform four tasks. In single tasks, participants were asked to use conversational agents to turn on the smart light and adjust illumination first to the brightest setting and then to a warmer tone of their preference. The complex tasks required participants to turn on the smart light, fan, and smart television using conversational agents and adjust these devices until they felt cozy and relaxed.	Efficiency: Task completion time, communica-tion effort Satisfaction: Ease-of-use, context-dependent question, complexity control, pleasure, want to use again, learnability	Questionnaire, software platform (record of interaction)	Counting the measured values, ANOVA statistical test

TABLE 8. *(Continued.)* All measured metrics of experiments.

Primary Study	F	ES	SS	TS	Design and Procedure	Tasks	Usability Characteristics	MI	ST
[PS5]	No	1	15	Residents	Design: Within-subjects design. Procedure: After signing a written consent to participate in the study, the participants completed a pre-test questionnaire and then performed the tasks using each interface one by one. After the participants completed the tasks using each interface, they were given a user satisfaction questionnaire and were asked to share their experience of working on the interface in a retrospective think-aloud session.	Each participant was required to perform five tasks: create a user profile, add the fictional FHx, re-access the platform, edit the information, and share the information with a family member.	Effectiveness: Number of errors/error rate Efficiency: Task completion time, mental effort Satisfaction: Ease-of-use, complexity control	Questionnaire, semi-structured interview, software platform	*t*-tests, Wilcoxon signed-rank test
[PS6]	Yes	2	35	Students	Design: Within-subjects mixed-method design. Procedure: Selected participants were given information on how to contact the chatbot through SMS, Twitter, and a web browser. Participants were instructed to use all three interfaces, and they completed the online survey immediately after they finished using each interface. Finally, participants were given compensation through an online gift card after completing the experiment.	Each participant was required to perform one task, that is, to interact with the chatbot.	Satisfaction: Ease-of-use, valuable, recommended	Questionnaire, software platform (record of code)	Analysis of contents, linear model
[PS7]	No	1	11	Not defined	Design: Within-subjects design. Procedure: The participants were first introduced to the background of ASRS. Next, they were introduced to the task and the procedure. The users interacted with the chatbot ROB in a web browser on a PC that was provided and completed the original paper version of the questionnaire. Finally, a semi-structured interview was conducted to explore the users' experiences with ROB.	Not defined.	Efficiency: Task completion time Satisfaction: Ease-of-use, pleasure, overall user experience	Questionnaire, software platform (record of answer)	Counting the measured values, paired *t*-test, analysis of contents
[PS8]	No	1	54	Graduate and undergraduates	Design: Within-subjects crossover design. Procedure: All participants first received a brief tutorial about the tool they had to use. They were then required to perform each task (build a class diagram) with the tool within a maximum of 30 minutes. At the end of each experimental session, the subjects filled in a modified and validated SUS questionnaire associated with the tool.	Participants were required to complete a total of two tasks. Task 1: Build a class diagram representing a store, including management of products and customers. Task 2: Design a class diagram of a school supporting courses and students.	Effectiveness: Task completion Efficiency: Task completion time, communica-tion effort Satisfaction: Adapted SUS score	Questionnaire, software platform (record of interaction) (record of interaction)	Meta-analysis, linear mixed model, analysis of contents

TABLE 8. *(Continued.)* All measured metrics of experiments.

Primary Study	F	ES	SS	TS	Design and Procedure	Tasks	Usability Characteristics	MI	ST
[PS9]	No	1	169	Not defined	Design: Between-subjects design. Procedure: There is no training or introduction session before performance of the task. Subjects were randomly assigned to one of the chatbots and asked to use the chatbot service to complete a task. Before and after the task, the participants were asked to fill in a questionnaire.	Each participant was required to perform one task: use the chatbot service to book a ticket for a play at a fictional Shakespeare theatre. The booking of the ticket took place in an imaginary setup, and no real booking was made.	Efficiency: Task completion time Satisfaction: Ease-of-use, during use, pleasure, overall user experience	Questionnaire, software platform (record of interaction)	Counting the measured values, unpaired Mann-Whitney U test, unpaired z-test
[PS10]	No	1	34	Farmers	Design: Within-subjects design. Procedure: Before the tasks, participants went through a training task to learn how to use each interface. After completing the training successfully, participants were asked to perform three tasks: a structured task, a semi-structured task, and an unstructured task (in that order). After completing all the tasks with each interface, participants were asked to rate their usage experience in response to eight items on a 5-point Likert scale.	Participants were asked to perform three tasks: (1) Participants were shown paper-printed color images of symptoms related to four common potato pests/diseases for the structured task. Moreover, they were asked if they had seen any of these pests/diseases in their field or in their neighbor's field recently. If they had, they could query FarmChat. (2) Participants were shown paper-printed color images of four major farming practices: buying input seeds, seeding, irrigation, and harvesting (including poor yield). Finally, (3) participants were encouraged to ask any potato farming-related questions on their minds for the unstructured task.	Efficiency: Response quality Satisfaction: Ease-of-use, pleasure, want to use again, learnability	Questionnaire, interview, software platform (record of interaction)	Counting the measured values
[PS11]	No	1	-	Students	Design: Between-subjects design. Procedure: Participants were randomly assigned to one of four treatments, each containing two scenarios in the same category and two user instructions (website vs. chatbot). The students were asked to use the Hipmunk website and the Hipmunk chatbot to complete two information search tasks. At the end of the search tasks, the participants were asked about their search results.	Each participant was required to perform two tasks: (1) search flight tickets and (2) search hotel rooms.	Efficiency: Mental effort Satisfaction: During use	Software platform (record of interaction)	Counting the measured values, ANOVA statistical test, structural modeling, adapt literature

TABLE 8. *(Continued.)* All measured metrics of experiments.

Primary Study	F	ES	SS	TS	Design and Procedure	Tasks	Usability Characteristics	MI	ST
[PS12]	No	1	16	Company employees	Design: Within-subjects user study with two interfaces. Procedure: After a one-minute tutorial video, participants started to complete the task. Finally, they were asked to rate their experience on a 5-point Likert scale.	Each participant was required to perform one of these two tasks: (1) select party footwear for themselves, or (2) select a pair of sports shoes for the opposite gender.	Effectiveness: Task completion Efficiency: Task completion time, mental effort Satisfaction: Ease-of-use, pleasure, want to use again	Questionnaire, software platform (record of interaction)	Counting the measured values, paired t-test
[PS13]	No	1	41	Experienced and inexperienced users, users with and without technical knowledge	Design: 2×2 factorial. Procedure: Participants were asked to self-report whether they considered themselves as experienced CA users and how frequently they used CA in the past three months. Then, each participant used Apple Siri to complete a set of tasks. The whole process of using Siri was also videorecorded as materials for a retrospective think-aloud and an interview after completing the query tasks. Finally, they completed the SUMI questionnaire.	Each participant was required to perform a series of trip planning tasks, for instance, to find an inexpensive hotel in Osaka.	Effectiveness: Experts and users' assessment Satisfaction: Learnability	Questionnaire, interview, video record	ANOVA statistical tests, t-tests
[PS14]	No	1	21	Children	Design: Within-subjects design. Procedure: During a four-day camp, a variety of activities with both a robot and an avatar were organized. On the first day and the last day of the camp, the participants were asked to complete the questionnaires.	Each participant was required to perform six research activities during the camp: Task 1: Plenary talk, robots introduce themselves. Task 2: "Small talk" in small groups. Task 3: Playing with the MyPAL app after dinner. Task 4: Bedtime story by the robots. Task 5: "Robot-rounds": four games; two games (quiz and sorting game) with the robot, and the same two games with the avatar. Task 6: Disco night with a robot dance performance.	Satisfaction: Ease-of-use, pleasure, want to use again, learnability	Questionnaire	Counting the measured values, ANOVA statistical test, paired t-tests, Gabriel's post hoc tests, logistic regression analysis, linear regression

TABLE 8. *(Continued.)* All measured metrics of experiments.

Primary Study	F	ES	SS	TS	Design and Procedure	Tasks	Usability Characteristics	MI	ST
[PS15]	No	1	8	Students	Design: Within-subjects crossover design. Procedure: Each subject was paired with a research assistant who played the role of a conversation partner. A conference room-style setting was used with a large display on which a camera was mounted. The subjects were given a sheet listing the commands and given an opportunity to study it for a few minutes prior to their interaction with the assistant. Once the interaction began, it consisted of an interleaved conversation with the research assistant (who would explain and/or suggest specific commands) and the automated assistant, to whom the subject would issue commands. Each user interacted with both Condition A chatbot and Condition B chatbot (see conditions A and B in Table VI). In order to reduce any bias that might result from the order in which they were exposed to these variants, half of the population were shown Condition A first while the other half were shown Condition B first. Following the interaction with each variant, they asked users questions and followed up with an interview. During the experiment, the subjects were encouraged to think out loud.	Not defined	<u>Efficiency:</u> Task completion time <u>Satisfaction:</u> Ease-of-use, user experience, helpfulness, attentiveness	Questionnaire, video record	Counting the measured values, video record, Fisher's exact test, Wilcoxon-Mann-Whitney test, analysis of contents (categorized qualitive answers into "Headpose" and "Wake Word" systems)
[PS16]	No	1	17	Not defined	Design: Latin square design. Procedure: A moderator guided the user through interactions with the prototype and gathered feedback in the form of interview-sessions at predefined points during the evaluation session. As such, the procedure resembled a cooperative evaluation but with user feedback gathered through demarcated interviews at different points in time in the evaluation protocol. The evaluation sessions lasted about 1 hour. The first 30 minutes of each session approximately concerned the four approaches to communicate service offers. All participants tried out all four approaches sequentially. After the scenario presentation, the moderator observed the participants as they interacted with the chatbot prototype for the specific approach.	Participants were required to perform four tasks: for each of the four approaches, the moderator presented a simple scenario where the context of use was exemplified through a persona in a specific situation. One such scenario could be that the persona wanted more information on life insurance, browsed the service provider webpage for such information, and then evoked the chatbot.	<u>Efficiency:</u> Task completion time, communication effort response quality <u>Satisfaction:</u> User experience, during use, valuable	Video record	Analysis of contents (thematic analysis)

TABLE 8. *(Continued.)* All measured metrics of experiments.

Primary Study	F	ES	SS	TS	Design and Procedure	Tasks	Usability Characteristics	MI	ST
[PS17]	No	1	61	Not defined	Design: Not defined. Procedure: As per the SUS recommendation, clear instructions are given to each of the participants before starting the experiment. Then, the participants had to ask the smart-speakers questions that were framed in a manner to identify the high-level features and functionalities provided by these devices. After getting back the responses from the smart-speakers based on their subjective opinions the participants had to rate the 10 items of the SUS scale. At the end, the participants were asked one additional question based on their overall experience.	There is one task that contains various questions they must ask. Question Themes: Semantic Intelligence, Clarity/details, Recognition over Recall, Interactive Feedback/Guidance, Mapping and Recovering from Errors.	Effectiveness: Number of errors/error rate Efficiency: Response quality Satisfaction: Learnability, semantic intelligence, clarity/details, recognition over recall, mapping	Questionnaire	Counting the measured values, data distribution, correlation analysis, Cronbach's alpha, Exploratory Factor Analysis (EFA), unweighted least-squares factor analysis (ULS), maximum likelihood factor analysis (ML); Kaiser-Meyer-Olkin (KMO) Test; Bartlett's Test of Sphericity, linear regression, analysis of contents (adjective rating scale)
[PS18]	No	1	61	Not defined	Design: Not defined. Procedure: Firstly, the general procedure of the experiment was explained to the participants. Then each participant was provided with a script that contains a variety of questions which they needed to ask the voice-assistants. Besides, the participants were free to retry completing any tasks as many times as they liked.	1. There is one obligatory task that contains various questions they must ask: General/usability, Affective recognition & visibility, Pragmatic, Errors & frustration, and Guidance & help. 2. The participants could interact with the chatbot freely in addition to asking necessary questions.	Effectiveness: Number of errors/error rate Efficiency: Response quality Satisfaction: Ease-of-use, pleasure, learnability, user experience	Questionnaire	Counting the measured values, data distribution, assumptions of multivariate analysis, EFA, ULS, ML, Parallel analysis, KMO Test, Bartlett's Test of Sphericity

TABLE 8. ***(Continued.)* All measured metrics of experiments.**

Primary Study	F	ES	SS	TS	Design and Procedure	Tasks	Usability Characteristics	MI	ST
[PS19]	No	1	20	Not defined	Design: Within-subjects design. Procedure: 1) They randomly divided the participants into two groups. 2) They explained the objective of the study to the participants and instructed them on how to use the systems. 3) Each participant was given the chance to try out the systems and learn how to interact with them before the real experiment began. 4) The participants were required to complete the task for both interfaces. 5) The users were asked to fill out two questionnaires, one for each interface.	Participants were asked to perform one task, including insert their movie preferences, request a recommendation, and evaluate five recommended movies, using the functions described previously.	Effectiveness: Accuracy precision Efficiency: Communication effort Satisfaction: Ease-of-use, want to use again/intent to use, user experience	Questionnaire	Counting the measured values
[PS20]	No	1	60	Visitors to the tourist office	Design: Between-subjects design. Procedure: The participants were briefly introduced to the nature of the experiment and signed a consent form. The participants were randomly assigned to either the intimate condition or the control condition. After instruction about the device and the agent's expertise, the subjects interacted with the agent. Finally, the participants completed an online survey and received a noncommercial gift.	Participants were asked to perform one task: interact with chatbot by asking as many questions as they liked.	Effectiveness: Number of errors/error rate Satisfaction: User experience	Questionnaire, software platform (record of interaction)	Counting the measured values, Kolmogorov-Smirnov tests, *t*-test, linear regressions, mediation analyses
[PS21]	No	1	46	Patients	Design: Parallel-group design. Procedure: The participants received monetary compensation for their participation. They were randomly assigned to either the chatbot group or the control group. Participants were required to complete baseline and post-intervention assessments after using the chatbot or an informative book for 4 weeks.	Participants in the baseline group were asked to perform one task: use the chatbot for 4 weeks freely. Participants in the control group were asked to perform one task: read a paperback titled "My Brain Still Needs Glasses: ADHD in Adolescents and Adults" for 4 weeks. The book contains self-help information that can be practiced by people with attention deficit but does not only target ADHD.	Efficiency: Task completion time, communication effort Satisfaction: Pleasure, learnability, user experience, pragmatic quality, valuable	Questionnaire, software platform (record of interaction)	Counting the measured values, post-hoc analysis, Pearson's correlation analysis, *t*-test, chi-square test, mixed effects models

TABLE 8. *(Continued.)* All measured metrics of experiments.

Primary Study	F	ES	SS	TS	Design and Procedure	Tasks	Usability Characteristics	MI	ST
[PS22]	No	2	12+ 10	Students and professionals	Design: Within-subjects counterbalanced design. Procedure: Participants rehearsed and delivered one 7-minute presentation using prepared PowerPoint slide decks and notes in each session, once with DynamicDuo and once without. In each session, participants were given 30 minutes to rehearse before delivering their videotaped presentation.	Participants were asked to perform one task: present slides with the chatbot.	<u>Satisfaction:</u> User experience	Questionnaire, interview	Counting the measured values, analysis of contents
[PS23]	No	1	15	People in pregnancy preparation or different pregnancy stages were enrolled	Design: Not defined. Procedure: Before the experiment, participants were required to add Dr. Joy as a friend on KakaoTalk to ensure ready access to the chatbot during the experiment. All the participants were given the daily tasks of asking Dr. Joy at least 3 questions and then giving the chatbot feedback with emoji, using at least one feature of the obstetrics chatbot, and finally sending a facilitator all the screenshots for the history of the day's use via KakaoTalk before midnight. After completing the usability testing, all participants were asked to fill out a questionnaire containing demographic characteristics, closed-ended questions, and open-ended questions.	Participants were asked to perform one task: ask chatbot Dr. Joy at least 3 questions freely every day.	<u>Satisfaction:</u> Ease-of-use, pleasure, learnability, user experience (usefulness)	Questionnaire, software platform (record of interaction)	Counting the measured values, Shapiro-Wilk normality test, Spearman correlation, analysis of contents (thematic analysis)
[PS24]	No	1	79	Students	Design: Within-subjects design. Procedure: The subjects were randomly divided into 2 groups for alternating sequence. Each group used two chatbots to perform 2 tasks in different order. Group 1 used Talkjipsa first and Group 2 used Samantha first. Each group was given a total of 20 minutes to finish two tasks. After completing each task, the subjects were asked to fill in a questionnaire.	Participants were asked to perform one task: select a desired product using the specified chatbot.	<u>Efficiency:</u> Task completion time, response quality <u>Satisfaction:</u> User experience, attractiveness, want to use again / intent to use, recommended	Questionnaire, manual record	Correlation coefficients, mean standards deviation between the variables, Varimax rotation, Cronbach Alpha value, Kaiser-Mayer-Olkin, Bartlett's chi-squared, linear regression

TABLE 8. (Continued.) All measured metrics of experiments.

Primary Study	F	ES	SS	TS	Design and Procedure	Tasks	Usability Characteristics	MI	ST
[PS25]	No	1	10	Not defined	Design: Within-subjects design. Procedure: After answering basic demographic questions, each user interacted with an interface twice, completing two tasks with each interface using a think-aloud protocol. After they had completed each task, they were required to fill out the User Engagement questionnaire. After they had completed all tasks, they were required to fill out a self-report questionnaire.	Participants were asked to perform four tasks: sleep disturbance, cognitive impairment, biomarkers and mobile health. Tasks have been formulated as "simulated work tasks" in the following way: "You have a friend who needs help with a school project where he needs to explore [topic]. He asks you to send him some easy-to-understand material about the topic, so you decide to use the Progress in Mind platform to search for resources. Use the [search interface] to search for publications and find at least 3 diseases that may be linked to [topic]/where [topic] can be applied. When you read a publication, please also decide whether or not you would send it to your friend to help him with his project."	Effectiveness: Task completion Efficiency: Task completion time Satisfaction: Overall user experience	Questionnaire, software platform (record of interaction)	Counting the measured values, t-test, analysis of contents
[PS26]	No	1	52	Native and non-native English-speakers	Design: Not defined. Procedure: The researchers first conducted a survey to gather participants' feedback on usability, usage pattern and the usefulness of the VAs. They were then. required to complete several tasks using Apple's HomePod and Amazon's Echo in a home environment for one week. Once they completed each task, they were asked to fill in a post-test SUS questionnaire.	The participants were asked to perform 10 tasks, including: 1. Use the VA to check the current weather. 2. Ask the VA to get directions to commute between two places. 3. Use the VA to check daily news headlines. 4. Use the VA to play songs. 5. Use the VA to set up an alarm for a specific time. 6. Ask the VA about a famous person. 7. Ask the VA to adjust its volume. 8. Ask the VA to tell some jokes. 9. Use the VA to read the audiobook. 10. Use the VA to read and write e-mails.	Satisfaction: User experience	Questionnaire	Counting the measured values, Pearson's correlation analysis, paired t-test, analysis of contents

TABLE 8. *(Continued.)* **All measured metrics of experiments.**

Primary Study	F	ES	SS	TS	Design and Procedure	Tasks	Usability Characteristics	MI	ST
[PS27]	No	1	94	Students	Design: 3x3 mixed factorial design. Procedure: The participants were introduced to the task, were asked to read the abuse script assigned to them and signed an informed consent. The participants were asked to perform the task. Each participant repeated the same process three times, as they interacted with three different CAs in turn. Each participant filled out a questionnaire at the end of each interaction session. Finally, each participant received about $12 in return for participation.	Each participant was required to perform the same task three times: they were asked to follow the common procedure that people normally go through for refunds such as greetings, order number confirmation, and refund requests, through the CA. When the agent informed them that a refund was not possible, the subjects were told to start the abuse session. Participants were asked to use in any order the 8 abusive phrases provided in the assigned script and to speak naturally.	<u>Satisfaction:</u> Agent likability Anthropo-morphism, perceived intelligence, clarity	Questionnaire	Counting the measured values, analysis of contents
[PS28]	No	1	32	Student and university staff	Design: Within-subject design. Procedure: Participants were briefly introduced to the experiment and signed an informed consent. They then completed a demographic questionnaire and were asked to perform 6 tasks (by converting a pictogram into required verbal command) with each device (a smartphone or a smart speaker). After interacting with each device, they were required to complete the NASA-TLX questionnaire. After finishing all tasks with both devices, participants then completed an interview. Finally, each participant received a 10€ voucher in return for participation.	Participants were asked to look at 6 practice pictograms, convert the information into verbal commands and perform 6 tasks twice (with different devices), including: 1. Playing music 2. Setting an alarm 3. Converting values 4. Asking for the time at a particular location 5. Controlling device volume 6. Requesting weather informa-tion.	<u>Effectiveness:</u> Task completion <u>Efficiency:</u> Mental effort	Questionnaire, interview, software platform (record of interaction)	Counting the measured values, ANOVA, linear mixed-effects models (LMM)

E. TASKS

Regarding the implemented experimental tasks, 35.7% (10) of the experiments contained between two and six tasks, 42.9% (12) contained only one task, three experiments required participants to perform 10 tasks, 12 tasks and 13 tasks respectively [PS26], [PS28], [PS2], and two experiments did not specify the number of tasks [PS13], [PS15].

F. USABILITY CHARACTERISTICS

In terms of usability characteristics used to measure the results, the surveyed experiments measured usability based on effectiveness, efficiency, and satisfaction. Of the experiments, 32.1% explored all three aspects [PS2], [PS3], [PS5], [PS8], [PS12], [PS17], [PS18], [PS19], [PS25], 32.1% explored efficiency and satisfaction [PS4], [PS7], [PS9], [PS10], [PS11], [PS15], [PS16], [PS21], [PS24], 25% explored only satisfaction [PS1], [PS6], [PS14], [PS22], [PS23], [PS26], [PS27], and three studies explored effectiveness and satisfaction [PS13], [PS20], [PS28].

We follow the definition of satisfaction given in ISO/IEC 25010 [16]: *"The degree to which users' needs are satisfied when a product or system is used in a specified context of use."* Satisfaction is the usability characteristic of most concern to researchers since it was evaluated most often. The measures of satisfaction primarily include ease-of-use, context-dependent questions (or inconsistency), satisfaction before and during use, complexity control, the physical discomfort of the interface, pleasure, the willingness to use the chatbot again (or intent to use the chatbot again), and enjoyment and learnability. Of the measures of satisfaction, ease of use, pleasure, and willingness to use the chatbot again were the most frequently measured, as shown in Table 9.

We found that in recent years more chatbot designers are inclined to evaluate the usability of the chatbot in order to put it into use in real life or in industry rather than for research or scholarly purposes. In [PS3], chatbot developers wanted to know if participants have a willingness to pay or know the price they are willing to pay. In [PS3], [PS9] and [PS12], they investigated whether participants intend to use their chatbot in real life.

It is obvious that more chatbots can afford more complex functions: the chatbot in [PS3] is equipped with a sentiment analyzer as it discovers items that best fit users' needs. Effectiveness is defined as the accuracy and completeness with which users achieve specified goals in the HCI field [16], [42]. From Table 10, we find that task completion and error rate are the effectiveness measures of most concern.

Efficiency is defined as the resources expended in relation to the accuracy and completeness with which the users achieve their goals in the HCI field [16], [42]. From Table 11, we find that more research focuses on measuring task completion time.

In [PS3], they measure detailed time spent per question and the number of concepts the user can introduce for each message from the chatbot. We discovered that the hedonic quality of conversation is relevant to the chatbot's efficiency since the effort required for users to understand and answer a chatbot request is frequently measured. In conclusion, it is clear that researchers have sought to understand chatbot reaction time and clarity of speech.

TABLE 9. Measures of satisfaction.

Measures of Satisfaction	N	Experiments
Overall user experience / SUS score	15	[PS7], [PS8], [PS9], [PS15], [PS16], [PS17], [PS18], [PS19], [PS20], [PS21], [PS22], [PS23], [PS24], [PS25], [PS26]
Ease of use	13	[PS3], [PS4], [PS5], [PS6], [PS7], [PS9], [PS10], [PS12], [PS14], [PS15], [PS18], [PS19], [PS23]
Pleasure	11	[PS1], [PS3], [PS4], [PS7], [PS9], [PS10], [PS12], [PS14], [PS18], [PS21], [PS23]
Learnability	9	[PS3], [PS4], [PS10], [PS13], [PS14], [PS17], [PS18], [PS21], [PS23]
Want to use again / Intent to use	7	[PS3], [PS4], [PS10], [PS12], [PS14], [PS19], [PS24]
Complexity control	3	[PS3], [PS4], [PS5]
During use	3	[PS9], [PS11], [PS16]
Valuable	3	[PS6], [PS16], [PS21]
Attractiveness	2	[PS2], [PS24]
Pragmatic quality	2	[PS2], [PS21]
Recommended	2	[PS6], [PS24]
Semantic intelligence/ Perceived intelligence	2	[PS17], [PS27]
Clarity/details	2	[PS17], [PS27]
Adaptability	1	[PS3]
Helpfulness	1	[PS15]
Context-dependent question	1	[PS4]
Hedonic quality	1	[PS2]
Attentiveness	1	[PS15]
Recognition over recall	1	[PS17]
Mapping	1	[PS17]
Anthropomorphism	1	[PS27]
Agent likability	1	[PS27]

TABLE 10. Measures of effectiveness.

Measures of Effectiveness	N	Experiments
Task completion	5	[PS2], [PS8], [PS12], [PS25], [PS28]
Number of errors/error rate	4	[PS5], [PS17], [PS18], [PS20]
Precision	2	[PS3], [PS19]
Accuracy	2	[PS3], [PS19]
Expert and user assessment	1	[PS13]

TABLE 11. Measures of efficiency.

Measures of Efficiency	N	Experiments
Task completion time	13	[PS2], [PS3], [PS4], [PS5], [PS7], [PS8], [PS9], [PS12], [PS15], [PS16], [PS21], [PS24], [PS25]
Communication effort	6	[PS3], [PS4], [PS8], [PS16], [PS19], [PS21]
Response quality	5	[PS10], [PS16], [PS17], [PS18], [PS24]
Mental effort	5	[PS3], [PS5], [PS11], [PS12], [PS28]

G. MEASUREMENT INSTRUMENTS

Measurement instruments refer to the instrument used to measure the experiment result quantitatively. Of the

TABLE 12. Descriptive statistics representation.

Descriptive Statistics Representation	N	Experiments
Descriptive statistics table / Frequency distribution table	15	[PS2], [PS3], [PS4], [PS14], [PS15], [PS17], [PS18], [PS19], [PS20], [PS21], [PS23], [PS24], [PS25], [PS27], [PS28]
Textual description	14	[PS9], [PS13], [PS15], [PS16], [PS17], [PS18], [PS19], [PS20], [PS21], [PS22], [PS23], [PS24], [PS25], [PS27]
Box plot	8	[PS1], [PS2], [PS6], [PS8], [PS12], [PS14], [PS20], [PS28]
Scatter plot	5	[PS2], [PS7], [PS17], [PS20], [PS21]
Line chart	4	[PS2], [PS13], [PS21], [PS24]
Bar chart	4	[PS3], [PS5], [PS10], [PS19]
Histogram	4	[PS15], [PS17], [PS18], [PS26]
Scree plot	2	[PS17], [PS18]

experiments, 92.9% (26) adopted questionnaires to measure chatbot usability, and almost all the usability questionnaires have undergone some type of psychometric evaluation [43], 57.1% (16) adopted software platforms to record participants' interaction or input information objectively, 21.4% (6) adopted interviews to record participants' answers to open-ended questions, and three experiments used video recording. Of the measurement instruments, questionnaires and software platforms were most frequently combined. We also observed that one experiment used the questionnaire without recording quantitative data. It is important to note that usability is not a one-dimensional software property: usability is a concept that includes effectiveness, efficiency, and satisfaction.

Usability techniques are different from measurement instruments. Measurement instruments are methods to measure and collect experimental data, whereas usability techniques refer to HCI techniques used in the usability evaluation process to raise the usability level of the software product. They could be methods of inspection, inquiry, or testing.

H. STATISTICAL TECHNIQUES

The statistical techniques used in the experiments are categorized from four perspectives: descriptive statistics, inferential statistics, a general linear model (GLM), and qualitative research. Descriptive statistics (Table 12) are representation methods that visually integrate multiple datasets to contextualize the data and improve reader understanding.

Of the 28 experimental results on chatbot usability, descriptive statistics tables and textual description were the most used presentation formats. Descriptive statistics tables and frequency distribution tables were used to understand the collected data in numerical form. Textual description always reports the effect size and confidence interval. Box plots were used to report the sample dispersion and skewness [23] (e.g., task completion rates of two compared tools [PS2]). There is one experiment that has not yet been executed [PS11].

Inferential statistics (Table 13) were used to analyze 18 experiment results. Inferential statistics deals with the process of using data analysis to deduce properties of an underlying probability distribution [44]. Inferential statistics methods are classified into parametric statistics and nonparametric statistics.

TABLE 13. Inferential statistics methods.

Type of Inferential Statistics Method	N	Inferential Statistics Method	Experiments
Parametric tests	12	Pearson correlation	[PS2], [PS24], [PS26]
		Paired t-test	[PS2], [PS7], [PS12], [PS14], [PS26]
		t-test	[PS5], [PS13], [PS20], [PS21], [PS25]
		z-test	[PS9]
		Kaiser-Meyer-Olkin Bartlett's chi-squared	[PS24]
Nonparametric tests	9	Wilcoxon rank sum tests	[PS1]
		Wilcoxon signed-rank test	[PS5]
		Wilcoxon test	[PS3], [PS19]
		Mann–Whitney U-test	[PS9]
		Wilcoxon-Mann-Whitney test	[PS15]
		Fisher's exact test	[PS15]
		Principal components analysis (PCA)	[PS17]
		Kolmogorov-Smirnov tests	[PS20]
		Spearman correlation	[PS23]

In general, parametric statistical tests, like Pearson correlation, paired t-test, and z-test, assume that some of the parameters are normally distributed. The Pearson correlation analysis is conducted in order to describe how a measurement of A is related to a measurement of B [23]. As the result of the analysis, the researchers in [PS2] claim that there was a partial correlation between the results of the physiological measurements and the UX quality evaluation results. In most cases, the z-test is an inference on a population of known variance, while the t-test is adopted if variance is not known. Nonparametric tests, such as the Wilcoxon and Mann–Whitney tests, were used in six experiments when experiments have one factor and two treatments. Note that the authors of [PS13] and [PS3] did not specify which of the t-tests and Wilcoxon tests they used, respectively.

I. LINEAR MODELS

The GLM category of methods are parametric tests used to describe the concept of the model. A GLM ensures that the estimated values provide the best possible linear fit to the data, minimizing the error with the least square method [45]. The analytical methods are ANOVA and regression, which are variations of GLM [46]. In terms of regression, 5 studies in Table 14 used linear regression (e.g., the Durbin–Watson test), and logistic regression and mixed effect models were each used once. ANOVA and MANOVA were also used in 7 experiments. In [PS3], they conducted a MANOVA statistical test on all the accuracy and cost of interaction metrics of usability experiments, whereas two-way ANOVAs were used to investigate the interaction effect between prior experience and technical knowledge on overall chatbot usability in [PS13].

TABLE 14. General linear model.

General Linear Model	N	Experiments
ANOVA/MANOVA	7	[PS3], [PS4], [PS11], [PS13], [PS14], [PS27], [PS28]
Linear regression	5	[PS2], [PS14], [PS17], [PS20], [PS24]
Mixed effect models	2	[PS21], [PS28]
Logistic regression	1	[PS14]

Qualitative research (Table 15) was conducted in 39.3% of experiments. The researchers analyzed the contents, specifically recording interviews and answers to open-ended questions, whereas [PS16], [PS23] and [PS27] adopted thematic analysis to analyze recorded interviews and user utterance data, respectively.

TABLE 15. Qualitative research.

Qualitative Research	N	Experiments
Analysis of contents	11	[PS6], [PS7], [PS8], [PS15], [PS16], [PS17], [PS22], [PS23], [PS25], [PS26], [PS27]

Additionally, most researchers did not explain the motivation behind technique adoption or indicate the challenges or advantages of adopting the technique. Some analysis decisions within chatbot usability experiments were affected or driven by previous examples from other researchers and personal preferences [PS2].

V. DISCUSSION

The mind map in Fig. 3 shows a summary of the five main aspects associated with chatbot usability experimentation, which are identified in the literature of our SMS: (i) measures, (ii) types of chatbots, (iii) usability techniques, (iv) descriptive statistics representation, and (v) inferential statistics methods.

The center of Fig. 3 corresponds to our research topic (Level 0 of the mind map). Five branches that point away from the center of the mind map symbolize the five above-mentioned aspects (Level 1). Another three hierarchical values—*effectiveness*, *efficiency*, and *satisfaction*— (Level 2) associated with the measured values branch off. At Level 3 of the mind map, values correspond to each item of the immediately preceding branch.

Continuing with our example, the measured values of *effectiveness* are *accuracy*, *expert* and *user assessment*, *number of errors/error rate*, *precision*, and *task completion*. Finally, at Level 4 of the mind map, experiment papers report the characteristics of the previous branch. Continuing with our example, the experiment reported in paper [PS3] corresponds to *accuracy*.

When conducting an SMS, the search strings should provide a broad overview of the research area [34]. Considering that chatbot usability experimentation is a relatively small field, we chose search strings that consisted of two components —synonyms of the terms ''chatbot'' and ''usability''— that helped to identify as many relevant papers as possible. We experimented with more than one synonym of the terms that formed different search strings to choose the best search string. Although our goal is to conduct an analysis of chatbot usability experimentation, we noticed that the interfaces of most current chatbots take the form of a NL dialog: the development of chatbots has become standardized because many platforms built for different goals and usages (e.g., Google's NLP platform and Dialogflow) have been widely used [PS1], [PS6], [PS10].

Of the initial 718 papers selected from well-known electronic research databases, 28 studies were selected following a rigorous screening process during which disagreements found during the selection process were resolved. The comparison of two or more treatments and the randomization of the subjects were key points for identifying whether the study described an experiment [23] when we reviewed each paper.

Regarding theoretical models, a few of the 28 experiments discussed theories that inspired research questions. In paper [PS11], we learned that self-determination theory was used to propose a research model to study the factors that affect chatbot satisfaction. On the other hand, most usability questionnaires assessed usability at the end of a study [43]. Self-designed questions [PS12], [PS14], [PS15], [PS18], [PS24] and standardized questionnaires are the two main usability scales used. However, the adoption of usability scales varies a lot in chatbot usability experiments, because it mainly depends on the research goal and chatbot type.

Researchers developed multiple questions for self-designed questionnaires according to their research topics and measurements. For example, researchers developed, based on the SUS questionnaire, the Voice Usability Scale for speech-based systems, as SUS does not comprehensively account for several characteristics that are unique to a voice environment [PS18]. Most experiments used standardized usability scales (like Affective Slider [PS1], ResQue model [PS3], [PS19], SUS [PS6], [PS8], [PS14], [PS17], SUMI [PS13], Adjective Rating Scale [PS18], Usefulness, Satisfaction, and Ease of use questions [PS23], User Engagement Scale [PS25]), whereas scales cited in national and international standards (SUS and SUMI) were adopted in only five experiments.

The chatbot usability experiment correlates to chatbot development. In general, evaluations of chatbot usability were considered as a part of the software development process. However, there are two experiments related to a usability experiment on an advanced or modified version of a chatbot [PS12], [PS15].

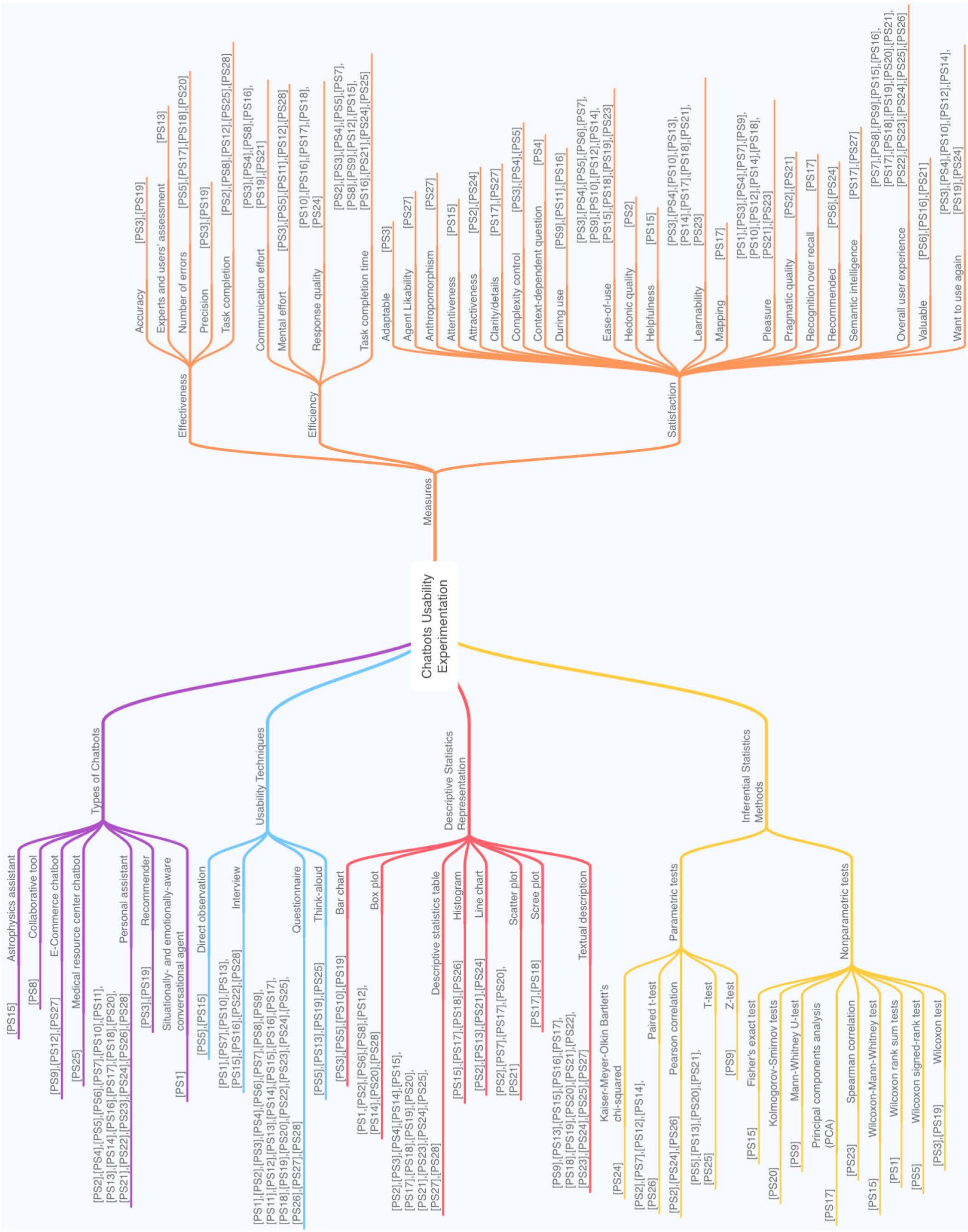

FIGURE 3. Main aspects of the research on chatbot usability experimentation.

For experimental results to be reliable, all the treatment aspects (except for factor manipulation) should be similar across all groups, as irrelevant variables pose a threat to validity.

We found that many studies did not clearly state extraneous variable control in their experimental designs. For example, they did not discuss the possible learning effects between different sessions [PS6], [PS10], if there was a short break between the different experiment sessions to avoid participant fatigue [PS1], [PS3], or whether the experimental environments were consistent in different sessions [PS4], [PS5].

We observed that most chatbot experiments were based on some specificities—including relatively small sample sizes, subjects coming from a specific background, preset tasks, and whether it was the users' first contact with a chatbot—as the expansion of the experimental results to an industrial setting was very limited. Besides, there was research that had not published the experimental results as of our search date. The proposed experimental setting in [PS11] included the procedure, type of subject, measurements, and analysis methods, but the sample size and experiment result were not provided.

VI. THREATS TO VALIDITY

The first threat to validity of this research is the bias in the paper selection process. Although the selection criteria and results have been double-checked and accepted by other authors, the publications were evaluated and classified based on our criteria and experience, and other researchers may have evaluated the publications differently. To improve the inter-rater reality, we provide percentage agreement and Cohen's Kappa statistics to evaluate disagreement between researchers (see Section 3).

The second point is related to the type of studies included in this investigation. We expanded the search scope by using search strings that identified a wider range of publications: the paper retrieval steps were as shown in Fig. 1 and the selected papers were grouped according to different dimensions as shown in Table 8. On the one hand, this systematic study was developed using five popular databases (IEEE Xplore, ACM Digital Library, SpringerLink, Scopus and ScienceDirect), as they are regarded as the most complete and most used databases in SE. On the other hand, this search only includes papers written in English. Nonetheless, the final number of studies focusing on exploring chatbot usability is relatively small—relevant papers produced by additional databases or resources or written in other languages or using other synonyms of chatbot could have been overlooked.

VII. CONCLUSION AND FUTURE WORK

This section reports the final conclusions of the study based on the research questions stated above.

RQ1: What is the state of the art of chatbot usability experimentation?

Chatbot development usability testing is not a new concept, but chatbot usability experimentation has emerged recently. Our SMS found that researchers started to evaluate the usability of chatbots through experimentation in 2018 (Fig. 2).

Several usability techniques have been used to collect usability data: questionnaires, interviews, think-aloud and direct observation. Of these techniques, questionnaires (applying various scales and types) are the most used technique. With regard to publication venue, half of the reviewed papers in our SMS were published at conferences.

In summary, chatbot usability experimentation tends to have the following characteristics: (i) very few raw data were provided (see Table 6); (ii) there was a range of chatbot types due to their usage scenarios, where a total of 67.9% of the experiments investigated chatbots pertaining to personal assistants, especially in the healthcare domain (Table 6); (iii) most experiments did not clearly define the research questions, hypotheses, or provide original data (Table 6), that is, they did not apply ESE methods to set up the experimental design [23], which may lead to weak experiment replicability; (iv) satisfaction, efficiency, and effectiveness were the main evaluated chatbot usability measures in most experiments (Tables 9, 10 and 11), and (v) parametric tests were the inferential statistics commonly used to analyze the experimental results in most studies (see Table 13).

RQ2: What research questions did chatbot usability experiments investigate?

Table 6 lists all the research questions used in the selected studies from five perspectives: (i) the goals of the experiment, (ii) the stated, selected or supplemented research questions, (iii) experiment hypotheses, (iv) answers to respective research questions and hypotheses, (v) provision of the experimental raw data, and (vi) chatbot types.

Regarding the treatment applied in these experiments, we found that control tools are commonly applied in experiments, and relatively few studies used the web or a real-life product [PS1], [PS2], [PS5], [PS8], [PS17], [PS18], [PS26], [PS28]. To determine whether the chatbot was able to provide a similar experience to the user, some developed different versions of chatbots with different functions or expression [PS3], [PS9], [PS10] to identify user preferences and how to operate differently depending on different user populations.

In general, most studies investigate not only usability factors but also the quality of the interaction or chatbot performance [PS3], [PS7], [PS8], [PS10], [PS28] in order to understand chatbot usability comprehensively. Also, some studies investigated the relationships between usability and other factors (e.g., acceptability, interface workload, and similarity) [PS5], [PS10], [PS14].

Most experiments did not provide access to raw data. The raw data may be withheld from the public domain either because they are confidential or because the researchers want to continue publishing data analyses sometime in the future [25]. However, this situation prevents rigorous peer review and stops third-party researchers from reanalyzing data using aggregation techniques that may be better suited than the original method [25].

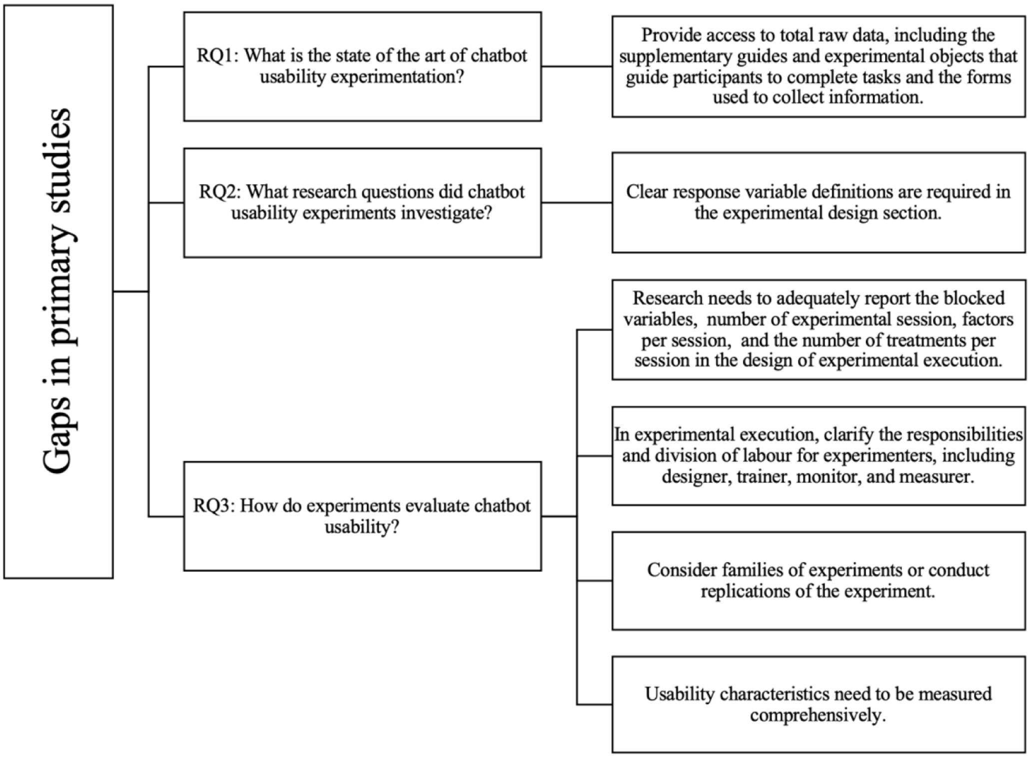

FIGURE 4. **Research gaps and future direction.**

RQ3: How do experiments evaluate chatbot usability?

As for chatbot usability experiments, we analyzed the evaluation metrics from nine perspectives, shown in Table 8: (i) whether the experiment is part of a family of experiments; (ii) the number of experiments; (iii) the experiment sample size; (iv) the types of subjects participating in the family; (v) experimental design and procedure; (vi) the implemented experimental tasks; (vii) usability characteristics used to measure the results; (viii) measurement instruments; and (ix) statistical techniques.

After reviewing the chatbot usability evaluations, we found that: (i) the families of experiments have seldom been used in this field so far since we found only one experiment replication; (ii) within-subjects experiments are generally the most popular design in chatbot usability experimentation; (iii) a total of 42.9 per cent of the experiments included a small sample size (under 30 subjects) and subjects are mostly students, and (iv) the number of tasks is relatively small, as most of the experiments applied fewer than six tasks.

The evaluation results revealed some common problems that existed within these chatbots. NL interaction (or natural mode of interaction) was the most cited problem. In general, chatbots satisfied and surprised users in basic interactions by using NL. However, chatbots required more effort from users in complex or flexible interaction and cannot yet compete with to human–human interaction. Chatbot personalization was the second issue mentioned, especially with respect to chatbots designed to target people with special needs, like students who require special mentorship or children with a specific disease. These chatbots should be highly adjustable, efficient, attractive in appearance and even have a physical embodiment. The experimental results show that personalization still needs improvement.

In terms of usability characteristics, satisfaction is of more concern than efficiency and effectiveness. The overall user experience, ease of use, and pleasure are the most frequent metrics used to measure satisfaction. Various studies assessed different aspects of satisfaction, complicating direct comparison. Some of this variation (e.g., adaptability [PS3], helpfulness [PS15], context-dependent question [PS4], and hedonic quality [PS2]) may be due to the individual characteristics of chatbot implementations and their distinct use cases [47]. On the other hand, task completion and number of errors/error rate are the effectiveness characteristics of most concern and have been measured a total of 9 times. With regard to efficiency, task completion time was measured most frequently.

Questionnaires and software platforms were the most popular measurement instruments. Questionnaires were commonly used for opinion polls [23], and software platforms were employed to record information for statistical analysis. Then, the collected information could be arranged in a quantitative or qualitative manner [23], and most researchers counted measurable values or analyzed the contents (e.g., the record of the interview, the answers to the open-ended questions), or ran parametric (e.g., t-tests) or nonparametric (e.g., Wilcoxon, Mann–Whitney tests) statistical tests depending on the experimental design type.

The research gaps shown in Figure 4 are used to identify experimental features associated with chatbot usability. They include defining each response variable clearly during the process of experimental design. In order to clearly report the execution of a chatbot usability experiment, factors like blocked variables, number of experimental sessions, factors per session, and the number of treatments in each session need to be properly specified in the design of the experimental execution. The clarification of responsibilities and division of labor for experimenters also helps in understanding the experimental process. Further, we encourage the measurement of usability characteristics whenever possible in order to gain a comprehensive understanding of chatbot usability. With regard to the data analysis and aggregation process, not enough raw data are provided, and families of experiments or experiment replications have seldom been reported to date. In view of this, we encourage future researchers to: (i) provide access to full raw data to guarantee the replicability of the experiment and the transparency of results to promote a better measurement of usability characteristics and a greater understanding of chatbot usability; (ii) clearly indicate the required characterization of chatbot usability experiments by including effect sizes, operationalization, design of the experimental execution, the experimenters; (iii) consider families of experiments or the possibility of conducting replications of the baseline experiment to consolidate the experimental results and increase the statistical power, and (iv) measure as many usability characteristics as possible to provide a thorough understanding of chatbot usability.

Future research may use and include the results of this SMS, especially the characteristics of chatbot usability experiments identified in this investigation, as a basis for conducting more studies to investigate this topic. Considering that the research is limited by search date, databases, and search strings, this study could be replicated in a future study. This is certainly an open research problem that requires further investigation. Based on the result of this research, we plan to conduct a family of experiments to evaluate the usability of a chatbot with an advanced version to fill the gaps and explore the topic further.

APPENDIX A
PRIMARY STUDIES

This appendix lists the references of the primary studies used for the mapping study described in this paper.

[PS1] S. Katayama, A. Mathur, M. Van den Broeck, T. Okoshi, J. Nakazawa, and F. Kawsar, "Situation-aware emotion regulation of conversational agents with kinetic earables," in *Proc. 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII'19)*, Cambridge, UK, 2019, pp. 725–731.

[PS2] S. Lee, H. Ryu, B. Park, and M. H. Yun, "Using physiological recordings for studying user experience: Case of conversational agent-equipped TV," *International Journal of Human Computer Interaction*, vol. 36, no. 9, pp. 815–827, Feb. 2020.

[PS3] F. Narducci, P. Basile, M. de Gemmis, P. Lops, and G. Semeraro, "An investigation on the user interaction modes of conversational recommender systems for the music domain," *User Modeling and User-Adapted Interaction*, vol. 30, pp. 251–284, Mar. 2020.

[PS4] J. Guo, D. Tao, and C. Yang, "The effects of continuous conversation and task complexity on usability of an AI-based conversational agent in smart home environments," in: S. Long, B. Dhillon (Eds.), *Man–Machine–Environment System Engineering,* MMESE'19, (pp. 695–703). Lecture Notes in Electrical Engineering, vol 576. Springer, Singapore, 2020.

[PS5] A. Ponathil, F. Ozkan, B. Welch, J. Bertrand, and K. C. Madathil, "Family health history collected by virtual conversational agents: An empirical study to investigate the efficacy of this approach," *Journal of Genetic Counseling*, pp. 1–12, Mar. 2020.

[PS6] E. W. Huff-Jr, N. A. Mack, R. Cummings, K. Womack, K. Gosha, and J. E. Gilbert, "Evaluating the usability of pervasive conversational user interfaces for virtual mentoring," in: P. Zaphiris, A. Ioannou (Eds.), *Learning and Collaboration Technologies. Ubiquitous and Virtual Environments for Learning and Collaboration*, HCII'19 (pp. 80–98). Lecture Notes in Computer Science, vol. 11591, Springer, Cham, 2019.

[PS7] R. Håvik, J. D. Wake, E. Flobak, A. Lundervold, and F. Guribye, "A conversational interface for self-screening for ADHD in adults," in: S. Bodrunova *et al.* (Eds.), *Internet Science*, INSCI'19 (pp. 133–144). Lecture Notes in Computer Science, vol. 11551. Springer, Cham, 2019.

[PS8] R. Ren, J. W. Castro, A. Santos, S. Pérez-Soler, S. T. Acuña, and J. de Lara, "Collaborative modelling: Chatbots or on-line tools? An experimental study," in *Proc. Evaluation and Assessment in Software Engineering (EASE'20)*, Trondheim, Norway, 2020, pp. 260–269.

[PS9] E. Elsholz, J. Chamberlain, and U. Kruschwitz, "Exploring language style in chatbots to increase perceived product value and user engagement," in *Proc. 2019 Conference on Human Information Interaction and Retrieval (CHIIR'19)*, Glasgow, Scotland, UK, 2019, pp. 301–305.

[PS10] M. Jain, P. Kumar, I. Bhansali, Q. V. Liao, K. N. Truong, and S. N. Patel, "FarmChat: A conversational agent to answer farmer queries," in *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT'18),* vol. 2, no. 4, 2018, pp. 170:1–170:22.

[PS11] Q. N. Nguyen, and A. Sidorova, "Understanding user interactions with a chatbot: A self-determination theory approach," in *Proc. 24th Americas Conference on Information Systems: Digital Disruption (AMCIS'18)*, New Orleans, LA, USA, 2018.

[PS12] M. Jain, R. Kota, P. Kumar, and S. N. Patel, "Convey: Exploring the use of a context view for chatbots," in *Proc. 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*, Montreal, QC, Canada, 2018, pp. 468:1–468:6.

[PS13] M.-L. Chen, and H.-C. Wang, "How personal experience and technical knowledge affect using conversational agents," in *Proc. 23rd International Conference on Intelligent User Interfaces Companion (IUI'18)*, Tokyo, Japan, 2018, pp. 53:1–53:2.

[PS14] C. Sinoo, S. van der Pal, O. A. B. Henkemans, A. Keizer, B. P. B. Bierman, R. Looije, and M. A. Neerincx, "Friendship with a robot: Children's perception of similarity between a robot's physical and virtual embodiment that supports diabetes self-management," *Patient Education and Counseling*, vol. 101, pp. 1248–1255, Jul. 2018.

[PS15] R. R. Divekar, J. O. Kephart, X. Mou, L. Chen, and H. Su, "You talkin' to me? A practical attention-ware embodied agent," in: D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, and P. Zaphiris (Eds), *Human-Computer Interaction*, INTERACT 2019, (pp. 760-780). Lecture Notes in Computer Science, vol 11748. Springer, Cham, 2019.

[PS16] A. Følstad and R. Halvorsrud, "Communicating service offers in a conversational user interface: An exploratory study of user preferences in chatbot interaction," in *Proc. 32nd Australian Conference on Human-Computer Interaction (OzCHI'20)*, Sydney, NSW, Australia, pp. 671–676, 2020.

[PS17] D. S. Zwakman, D. Pal, T. Triyason, and V. Vanijja, "Usability of voice-based intelligent personal assistants," in *Proc. International Conference on Information and Communication Technology Convergence (ICTC'20)*, Jeju, Korea (South), pp. 652–657, 2020.

[PS18] D. S. Zwakman, D. Pal, and C. Arpnikanondt, "Usability evaluation of artificial intelligence-based voice assistants: The case of Amazon Alexa," *SN Computer Science*, vol. 2, article 28, 2021.

[PS19] A. Iovine, F. Narducci, M. De Gemmis, and G. Semeraro, "Humanoid robots and conversational recommender systems: A preliminary study," in *Proc. IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS'20),* Bari, Italy, pp. 1–7, 2020.

[PS20] D. Potdevin, C. Clavel, and N. Sabouret, "A virtual tourist counselor expressing intimacy behaviors: A new perspective to create emotion in visitors and offer them a better user experience?," *International Journal of Human-Computer Studies*, vol. 150, article 102612, 2021.

[PS21] S. Jang, J. J. Kim, S. J. Kim, J. Hong, S. Kim, and E. Kim, "Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study," *International Journal of Medical Informatics*, vol. 150, article 104440, 2021.

[PS22] T. Bickmore, E. Kimani, A. Shamekhi, P. Murali, D. Parmar, and H. Trinh, "Virtual agents as supporting media for scientific presentations," *Journal on Multimodal User Interfaces*, vol. 15, pp. 131–146, 2021.

[PS23] K. Chung, H. Y. Cho, and J. Y. Park, "A chatbot for perinatal women's and partners' obstetric and mental health care: Development and usability evaluation study," *JMIR Medical Informatics*, vol. 9, no. 3, 2021.

[PS24] Y. Lim, J. Lim, and N. Cho, "An experimental comparison of the usability of rule-based and natural language processing-based chatbots," *Asia Pacific Journal of Information Systems*, vol. 30, no. 4, pp. 832–846, 2020.

[PS25] T. Fergencs, and F. Meier, "Engagement and usability of conversational search – A study of a medical resource center chatbot," in: K. Toeppe, H. Yan, and S.K.W. Chu (Eds.), *Diversity, Divergence, Dialogue*. iConference 2021 (pp. 328-345). Lecture Notes in Computer Science, vol 12645. Springer, Cham, 2021.

[PS26] D. Pal, C. Arpnikanondt, S. Funilkul, and V. Varadarajan, "User experience with smart voice assistants: The accent perspective," in *Proc. 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT'19)*, Kanpur, India, pp. 1-6, 2019.

[PS27] H. Chin, L. W. Molefi, and M. Y. Yi, "Empathy is all you need: How a conversational agent should respond to verbal abuse," in *Proc. 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*, Honolulu, HI, USA, pp. 1-13, 2020.

[PS28] Y. Wu, J. Edwards, O. Cooney, A. Bleakley, P. R. Doyle, L. Clark, D. J. Rough, and B. R. Cowan, "Mental workload and language production in non-native speaker IPA interaction," in *Proc. 2nd Conference on Conversational User Interfaces (CUI'20)*, Bilbao, Spain, pp. 1-8, 2020.

REFERENCES

[1] J. Guichard, E. Ruane, R. Smith, D. Bean, and A. Ventresque, "Assessing the robustness of conversational agents using paraphrases," in *Proc. IEEE Int. Conf. Artif. Intell. Testing*, Newark, CA, USA, 2019, pp. 55–62.

[2] Nielsen Norman Group. (2020). *The User Experience of Chatbots*. [Online]. Available: https://www.nngroup.com/articles/chatbots/

[3] M. Jain, R. Kota, P. Kumar, and S. N. Patel, "Convey: Exploring the use of a context view for chatbots," in *Proc. Conf. Hum. Factors Comput. Syst.*, Montreal, QC, Canada, 2018, pp. 1–6.

[4] S. Katayama, A. Mathur, M. Van den Broeck, T. Okoshi, J. Nakazawa, and F. Kawsar, "Situation-aware emotion regulation of conversational agents with kinetic earables," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact.*, Cambridge, U.K., 2019, pp. 725–731.

[5] Q. N. Nguyen and A. Sidorova, "Understanding user interactions with a chatbot: A self-determination theory approach," in *Proc. 24th Amer. Conf. Inf. Syst. Digit. Disruption (AMCIS)*, New Orleans, LA, USA, 2018, pp. 1–5.

[6] K. Panetta. (2016). *Gartner's Top 10 Strategic Technology Trends for 2017.* [Online]. Available: https://itango.eu/gartners-top-10-strategic-technology-trends-for-2017/

[7] S. Lee, H. Ryu, B. Park, and M. H. Yun, "Using physiological recordings for studying user experience: Case of conversational agent-equipped TV," *Int. J. Human Comput. Interact.*, vol. 36, no. 9, pp. 815–827, Feb. 2020.

[8] F. Narducci, P. Basile, M. de Gemmis, P. Lops, and G. Semeraro, "An investigation on the user interaction modes of conversational recommender systems for the music domain," *User Model. User-Adapted Interact.*, vol. 30, pp. 251–284, Mar. 2020.

[9] R. Ren, J. W. Castro, A. Santos, and J. de Lara, "Collaborative modelling: Chatbots or on-line tools? An experimental study," in *Proc. Eval. Assessment Softw. Eng.*, Trondheim, Norway, 2020, p. 260 269.

[10] DriftTM. (2020). *Why are Chatbots Important Chatbot Learning Center*. [Online]. Available: https://www.drift.com/learn/chatbot/why-are-chatbots-important/

[11] R. Ren, J. W. Castro, S. T. Acuña, and J. de Lara, "Evaluation techniques for chatbot usability: A systematic mapping study," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 29, no. 11n12, pp. 1673–1702, Nov. 2019.

[12] S. Greenberg and B. Buxton, "Usability evaluation considered harmful (some of the time)," in *Proc. 26th Annu. CHI Conf. Hum. Factors Comput. Syst.*, 2008, Art. no. 111120.

[13] A. Seffah, M. C. Desmarais, and E. Metzker, "HCI, Usability and software engineering integration: Present and future," in *Human-Centered Software Engineering—Integration Usability in the Software Development Lifecycle* (Human-Computer Interaction Series), vol. 8, A. Seffah, J. Gulliksen, M. C. Desmarais, Eds. Dordrecht, The Netherlands: Springer, 2005, pp. 37–57.

[14] K. Curcio, R. Santana, S. Reinehr, and A. Malucelli, "Usability in agile software development: A tertiary study," *Comput. Standards Interface*, vol. 64, pp. 61–77, May 2019.

[15] *Ergonomic Requirements for Office Work With Visual Display Terminals (VDTs)—Part 11: Guidance on Usability*, Standard ISO 9241-11 1998.

[16] *Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—System and Software Quality Models*, Standard ISO/IEC 25010, 2011.

[17] A. Abd-Alrazaq, Z. Safi, M. Alajlani, J. Warren, M. Househ, and K. Denecke, "Technical metrics used to evaluate health care chatbots: Scoping review," *J. Med. Internet Res.*, vol. 22, no. 6, Jun. 2020, Art. no. e18301.

[18] S. Hobert, "How are you, chatbot? Evaluating chatbots in educational settings - Results of a literature review," in *Proc. Gesellschaft Informatik*, N. Pinkwart, J. Konert, Eds. Phocis, Greece: DELFI, 2019, pp. 259–270.

[19] A. Rapp, L. Curti, and A. Boldi, "The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots," *Int. J. Hum.-Comput. Stud.*, vol. 151, Apr. 2021, Art. no. 102630.

[20] G. Tariverdiyeva and S. Borsci, "Chatbots' perceived usability in information retrieval tasks: An exploratory analysis," M.S. thesis, Behavioural, Manage. Social Sci., University of Twente, Enschede, The Netherlands, 2019.

[21] N. M. Radziwill and M. C. Benton, "Evaluating quality of chatbots and intelligent conversational agents," 2017, *arXiv:1704.04579*.

[22] K. Seaborn and J. Urakami, "Measuring voice UX quantitatively: A rapid review," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Yokohama, Japan, vol. 416, May 2021, pp. 1–8.

[23] C. Wohlin, P. Runeson, M. Hést, M. C. Ohlsson, B. Regnell, and A. Wesslen, *Experimentation in Software Engineering*, Berlin, Germany: Springer, 2012.

[24] N. Juristo, "Once is not enough: Why we need replication," in *Perspectives on Data Science for Software Engineering*, T. Menzies, L. Williams, and T. Zimmermann Eds. Burlington, MA, USA: Morgan Kaufmann, 2016, pp. 299–302, doi: 10.1016/B978-0-12-804206-9.00054-4.

[25] A. Santos, O. Gomez, and N. Juristo, "Analyzing families of experiments in SE: A systematic mapping study," *IEEE Trans. Softw. Eng.*, vol. 46, no. 5, pp. 566–583, May 2020.

[26] E. Fernéndez, O. Dieste, P. Pesado, and R. Garcéa, "The importance of using empirical evidence in software engineering," in *Proc. Comput. Sci. Technol. Ser.*, G. Simani, and H. Padovani, Eds. Provincia de Buenos Aires, Argentina: Universidad de la Plata, 2011, pp. 181–189. [Online]. Available: https://digital.cic.gba.gob.ar/bitstream/handle/11746/4016/11746_4016.pdf-PDFA.pdf?sequence=1&isAllowed=y

[27] V. R. Basili, F. Shull, and F. Lanubile, "Building knowledge through families of experiments," *IEEE Trans. Softw. Eng.*, vol. 25, no. 4, p. 456 473, Jul/Aug. 1999.

[28] G. Biondi-Zoccai, *Umbrella Reviews: Evidence Synthesis With Overviews of Reviews and Meta-Epidemiologic Studies*. Cham, Switzerland: Springer, 2016.

[29] H. Cooper and E. A. Patall, "The relative benefits of meta-analysis conducted with individual participant data versus aggregated data," *Psychol. Methods*, vol. 14, no. 2, pp. 165–176, Jun. 2009.

[30] T. P. A. Debray, K. G. M. Moons, G. van Valkenhoef, O. Efthimiou, N. Hummel, R. H. H. Groenwold, J. B. Reitsma, and G. M. R. Group, "Get real in individual participant data (IPD) meta-analysis: A review of the methodology," *Res. Synth. Methods, vo.*, vol. 6, no. 4, p. 293 309, Aug. 2015.

[31] G. H. Lyman and N. M. Kuderer, "The strengths and limitations of meta-analyses based on aggregate data," *BMC Med. Res. Methodol.*, vol. 5, no. 1, pp. 1–7, Apr. 2005.

[32] L. A. Stewart, M. Clarke, M. Rovers, R. D. Riley, M. Simmonds, G. Stewart, J. F. Tierney, and P.-I. D. Group, "Preferred reporting items for a systematic review and meta-analysis of individual participant data: The PRISMA-IPD statement," *JAMA*, vol. 313, no. 16, p. 1657 1665, 2015.

[33] B. A. Kitchenham, D. Budgen, and P. Brereton, *Evidence-based software engineering and systematic reviews*, vol. 4. Boca Raton, FL, USA: CRC Press, 2016.

[34] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *Proc. 12th Int. Conf. Eval. Assessment Softw. Eng.*, Swindon, U.K., 2008, pp. 68–77.

[35] R. P. Reyes, O. Dieste, E. R. Fonseca, and N. Juristo, "Publication bias: A detailed analysis of experiments published in ESEM," in *Proc. Eval. Assessment Softw. Eng.*, Trondheim Norway, 2020, pp. 130–139.

[36] A. Ampatzoglou, S. Bibi, P. Avgeriou, and A. Chatzigeorgiou, "Guidelines for managing threats to validity of secondary studies in software engineering," in *Contemporary Empirical Methods in Software Engineering*, M. Felderer and G. Travassos, Eds. Cham, Switzerland: Springer, 2020, pp. 415–441.

[37] C. E. Anchundia and E. R. Fonseca, "Resources for reproducibility of experiments in empirical software engineering: Topics derived from a secondary study," *IEEE Access*, vol. 8, pp. 8992–9004, 2020.

[38] K. L. Gwet, *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Gaithersburg, MD, USA: Advanced Analytics, 2014.

[39] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[40] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.

[41] A. Jedlitschka and D. Pfahl, "Reporting guidelines for controlled experiments in software engineering," in *Proc. Int. Symp. Empirical Softw. Eng.*, Noosa Heads, QLD, Australia, 2005, pp. 95–104.

[42] K. Hornbák, "Current practice in measuring usability: Challenges to usability studies and research," *Int. J. Hum.-Comput. Stud.*, vol. 64, no. 2, p. 79 102, Feb. 2006.

[43] J. Sauro and J. R. Lewis, "Standardized usability questionnaires," in *Quantifying the User Experience*, 2nd ed. Boston, MA, USA: Morgan Kaufmann, 2016, pp. 185–248. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128023082000084?via%3Dihub

[44] G. Upton and I. Cook, *A Dictionary of Statistics*. New York, NY, USA: Oxford, 2008, doi: 10.1093/acref/9780199541454.001.0001.

[45] J. Sauro and J. R. Lewis, "An introduction to correlation, regression, and ANOVA," in *Quantifying User Experience*, vol. 10. Boston, MA, USA: Morgan Kaufmann, 2016, p. 277 320. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128023082000102

[46] A. Rutherford, *ANOVA and ANCOVA: A GLM Approach*. Hoboken, NJ, USA: Wiley, 2011.

[47] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Mach. Learn. Appl.*, vol. 2, Oct. 2020, Art. no. 100006.

RANCI REN received the M.S. degree in ICT research and innovation from the Universidad Autónoma de Madrid (UAM), Spain, in 2019, where she is currently pursuing the Ph.D. degree in software engineering. Her main research interests include experimental software engineering, human–computer interaction, and chatbots. She is a member of the ACM.

MIREYA ZAPATA received the Ph.D. degree in electronic engineering from the Universitat Politècnica de Catalunya, in 2017. She is a member of the Institute for Research, Development and Innovation, and an Associate Professor at the Universidad Tecnológica Indoamérica, Ecuador. She has participated in three Spanish national research projects with the Integrated Smart Sensors and Health Technologies Group, Department of Electronic Engineering, Universitat Politècnica de Catalunya. Her research interests include digital VLS and FPGA design, bioinspired/neuromorphic implementations, and interactive systems applied to education.

JOHN W. CASTRO received the M.S. degree in computer science and telecommunications, specializing in advanced software development, and the Ph.D. degree from the Universidad Autónoma de Madrid, in 2009 and 2015, respectively. He has 15 years of experience in the area of software systems development. He is currently an Assistant Professor at the University of Atacama, Chile. He worked as a Research Assistant at the Universidad Politécnica de Madrid. His research interests include software engineering, software development process, and the integration of usability in the software development process.

OSCAR DIESTE received the B.S. and M.S. degrees in computing from the Universidade da Coruña and the Ph.D. degree from the Universidad de Castilla-La Mancha. He is a Researcher with the Universidad Politécnica de Madrid (UPM)'s School of Computer Engineering. He was previously with the University of Colorado Colorado Springs, Colorado Springs, as a Fulbright Scholar; the Universidad Complutense de Madrid; and the Universidad Alfonso X el Sabio. His research interests include empirical software engineering and requirements engineering.

SILVIA T. ACUÑA received the Ph.D. degree from the Universidad Politécnica de Madrid, in 2002. She is currently an Associate Professor of software engineering at the Universidad Autónoma de Madrid's Computer Science Department. Her research interests include experimental software engineering, software usability, software process modeling, and software team building. She has coauthored the book *A Software Process Model Handbook for Incorporating People's Capabilities* (Springer, 2005), and edited the books *Software Process Modeling* (Springer, 2005) and *New Trends in Software Process Modeling* (World Scientific, 2006). She is a member of the IEEE Computer Society and the ACM. She is the Deputy Conference Co-Chair on the Organizing Committee of ICSE 2021.