Conversational Agents in Software Engineering: Survey, Taxonomy and Challenges

QUIM MOTGER and XAVIER FRANCH, Department of Service and Information System Engineering (ESSI), Universitat Politècnica de Catalunya (UPC), Spain
JORDI MARCO, Department of Computer Science (CS), Universitat Politècnica de Catalunya (UPC), Spain

The use of natural language interfaces in the field of human-computer interaction is undergoing intense study through dedicated scientific and industrial research. The latest contributions in the field, including deep learning approaches like recurrent neural networks, the potential of context-aware strategies and user-centred design approaches, have brought back the attention of the community to software-based dialogue systems, generally known as conversational agents or chatbots. Nonetheless, and given the novelty of the field, a generic, context-independent overview on the current state of research of conversational agents covering all research perspectives involved is missing. Motivated by this context, this paper reports a survey of the current state of research of conversational agents through a systematic literature review of secondary studies. The conducted research is designed to develop an exhaustive perspective through a clear presentation of the aggregated knowledge published by recent literature within a variety of domains, research focuses and contexts. As a result, this research proposes a holistic taxonomy of the different dimensions involved in the conversational agents’ field, which is expected to help researchers and to lay the groundwork for future research in the field of natural language interfaces.

CCS Concepts: • General and reference → Surveys and overviews; • Human-centered computing → Natural language interfaces.

Additional Key Words and Phrases: conversational agents, chatbots, systematic literature review

1 INTRODUCTION

Conversational agents, chatbots, dialogue systems and virtual assistants are some of the terms used by scientific literature [122] to describe software-based systems which are capable of processing natural language data to simulate a smart conversational process with humans [23]. These conversational mechanisms are built and driven by a wide variety of techniques of different complexity, from traditional, pre-coded algorithms to emerging adaptive machine learning algorithms [89]. Usually deployed as service-oriented systems, they are designed to assist users to achieve a specific goal based on their personal needs [38]. To this end, they autonomously generate natural language messages to interact and communicate with users by emulating a real human being [43, 108].

Although the interest in both industry and research has dramatically increased in recent years [79], the study of natural language communication between human beings and machines is indeed not a novel concept. ELIZA [133], which has been historically considered as the first chatbot, was designed and developed by the Massachusetts Institute of Technology (MIT) more than half a century ago, between 1964 and 1966. Alongside successive chatbots like PARRY [21] and A.L.I.C.E. [130], these innovative systems laid the groundwork for specialized research in the field of human-computer interaction (HCI), focusing on the social and communicative perspectives and their impact on the design and development of these systems. Recent advances in the field of artificial intelligence have brought back attention to the potential of conversational agents, especially with the emergence of machine and deep learning techniques [10]. Furthermore, specialized research fields such as natural language understanding (NLU), natural language generation (NLG) and
dialogue stage tracking (DST) have become disruptive areas by introducing innovative, efficient and accurate solutions to machine cognitive problems [132].

Consequently, several literature reviews and surveys have been conducted in recent years [56, 70], with an especial emphasis on the implications of the latest innovations from a technical perspective [7, 27]. Nevertheless, these secondary studies are typically limited in terms of context or domain of application [40], focus of research field [96] or even research method [5]. To the best of our knowledge, no literature review in the field of conversational agents addresses all the research dimensions involved in this field.

All things considered, in this paper, we conduct a state-of-the-art review in the field of conversational agents, and their impact in the software engineering field, applying a rigorous research method based on Kitchenham et al. guidelines on systematic reviews [69]. The contributions of this work are: to provide an up-to-date, holistic review of the conversational agents’ research field; to present a taxonomy of the main concepts uncovered in the study, which will allow researchers in the field to classify their work and compare to other’s proposals; to identify the main challenges and research directions for future work. The remainder of this article is organized as follows. Section 2 describes the research method. Sections 3, 4, 5, 6 and 7 present the results of the feature extraction process and the conclusions of each of the research questions. Section 8 summarizes the main findings from the research method and recommendations for future surveys. Section 9 evaluates the limitations of the study. Finally, Section 10 summarizes the main conclusions.

2 RESEARCH METHOD

We conducted our survey as a tertiary study over the scientific literature in the domain. A tertiary study is a systematic literature review used to find and synthesize the knowledge in a wide scope of research that is scattered in existing reviews. We followed Kitchenham et al.’s approach to perform tertiary studies [69] with a few modifications, remarkably:

- In addition to automated search selection, we conducted a reference snowballing process to avoid missing relevant papers in the field of conversational agents (see Section 2.4.3).
- We extended the quality assessment criteria list according to the latest practices in the software engineering field (see Section 2.4.4).

The detailed research method and the generated artefacts are collected and published in a replication package¹, which allows the auditability and the replicability of the research process.

2.1 Research questions

This paper addresses the following general research question (GQ):

- **GQ. What is the current state of research in the field of conversational agents?**

This general question is refined into a sub-set of specific scientific questions (SQs). Each SQ is supported through the feature extraction process by categorizing the sub-set of studies addressing the topic covered by each SQ (see Section 2.5). These SQs are defined as follows:

- **SQ1. What research has been published in the field of conversational agents?**
- **SQ2. Which HCI features have a relevant impact on user experience with conversational agents?**
- **SQ3. Which technical methods and technologies are used for the design and implementation of conversational agents?**
- **SQ4. Which methodological approaches and resources are used for training, testing and evaluating conversational agents?**
- **SQ5. Which research challenges are reported by conversational agents’ literature?**

¹Available at: https://zenodo.org/record/4973650
2.2 Search string

Based on our general research question, two key major concepts are identified: state of the art and conversational agents. Regarding the former, and since we have adopted tertiary studies as our methodological approach, we decided to include literature reviews performed according to a rigorous process. To guarantee that the search exhaustively covers all relevant literature in the field, we use a set of synonyms as proposed by Kitchenham et al. [69] which encompass the majority of structured literature reviews and secondary studies (e.g., literature reviews, mapping studies, literature surveys...). Regarding the latter, we carried on an exploratory review of the current state of research through recent systematic literature reviews in the field to include an extensive selection of alternative terms used by the literature that refer to conversational agents (e.g., chatbots, dialogue systems, conversational entities...). The search string is depicted in detail in the replication package.

2.3 Data sources

We selected the following digital libraries as well-known, broadly used repositories in literature reviews from the computer science and software engineering fields: Web of Science, Scopus, ACM Digital Library, IEEE Xplore, Science Direct and Springer Link.

2.4 Study selection strategy

In this section, we depict the study selection strategy, whose results are summarized in Figure 1.

Fig. 1. Summary of the study selection process.

2.4.1 Study search. The study search is designed as the following set of semi-automatic tasks:

1. Customize the search string to the syntactic requirements of each data source.
2. Query the adapted string in each data source engine using the filter by title, abstract and authors’ keywords. No restriction regarding the date of publication was applied.
3. Export the bibliographic references of the search results into a shared spreadsheet document.
4. Integrate all references using a common format based on all shared fields.
5. Filter out and remove all duplicated references.

This process resulted in an initial data-set of 136 studies.

---

2 The search was conducted on the 31st January, 2021
2.4.2 **Inclusion and exclusion criteria.** After the extraction of the initial data-set of studies from the study selection process, we applied a set of inclusion and exclusion criteria to evaluate each of the selected studies. For the inclusion criteria (IC), we include those studies matching all the following:

- **IC1.** The paper is a literature review as defined by Kitchenham et al. [69].
- **IC2.** The focus of the research is on conversational agents, chatbots or any of the synonyms included in the exploratory analysis.

Regarding the exclusion criteria (EC), we exclude all studies matching any of the following:

- **EC1.** Research results, contributions and conclusions cannot be applied to the computer science, software engineering or artificial intelligence research fields.
- **EC2.** Research focuses exclusively on embodied conversational agents, virtual reality or 3-D modelling features of conversational agents.
- **EC3.** The paper is a previous study of a more recent publication which has already been included in the initial data-set.
- **EC4.** The paper is a work in progress.
- **EC5.** Full-text is not available and it has not been delivered after request to their authors.
- **EC6.** Full-text is not published in English.

From the initial data-set of 134 studies, only 66 matched both ICs. Concerning the most common ECs, 39 studies were not focused on analysing and reporting research conclusions and results directly related to the computer science domain (EC1). Some examples include analysis on healthcare rhetorical structure theory [51], digital health coaching programs [118] or behavioural change in patients using medical chatbots [40]. On a secondary basis, up to 11 studies matching EC2 were focused on embodied conversational agents [77, 117] or physical modelling features [16]. Finally, we still excluded 14 additional studies applying the rest of ECs.

As a result, we identified a sub-set of 22 studies fulfilling all ICs and none of the depicted ECs.

2.4.3 **Reference snowballing.** To extend the scope of research and to minimize the risk of missing relevant studies, we integrated an additional reference search strategy into the study selection strategy. The snowballing reference search process is designed as depicted by Wohlin [135] based on iterative backward and forward reference searches. In each iteration, we conducted a backward and forward snowballing process of all secondary studies included in our data-set. Each reference fitting the goal of this research and passing the IC/EC evaluation (i.e., matching all IC and none of the EC defined in Section 2.4.2) was added to the original data-set, and a new iteration was applied to the enlarged data-set. We conducted three complete iterations until reaching saturation (see Figure 1). This process resulted in 8 new studies, which led to a total amount of 30 studies.

2.4.4 **Quality assessment.** The study selection step was concluded with a quality assessment to evaluate the quality of the research strategies depicted and used in the selected secondary studies. Using relevant tertiary studies from the software engineering field as reference [17, 75, 139], the quality assessment plan was designed based on the quality assessment criteria (QAs) included in the DARE-5 extended proposal from the York University Centre for Reviews and Dissemination [35] as refined by Kitchenham et al. in systematic reviews [68] and tertiary studies [69]. We define these QAs as follows:

- **QA1.** Does the search protocol presumably cover all relevant data sources?
- **QA2.** Are inclusion and exclusion criteria adequately described?
- **QA3.** Does the search protocol include a quality assessment evaluation?
- **QA4.** Are primary studies adequately identified and described?
- **QA5.** Is there a synthesis process to summarize and support research results and conclusions?
We use this QAs to compute a global quality score for each paper, which we use to remove all secondary studies with a quality score of 0 (details on QAs evaluation and filtering are available in the replication package). After applying this QA filter, we filtered out 5 studies, which gave us a final data-set of 25 studies.

2.5 Data extraction

The data extraction process was designed in alignment with the research questions described in Section 2.1. We identified and defined a set of research data features (F), which are intended to exhaustively cover the scope and topics covered by the SQs included in our research while facilitating the dissertation and discussion.

- **F1. Terminology.** Identification and analysis of the terminology used by literature to describe and refer to conversational agents, including synonyms, variants and descriptors.
- **F2. Domains.** Identification and categorization of business domains and areas of applications covered by research in the conversational agents’ field.
- **F3. Goals.** Identification and categorization of the primary goals of integrating conversational agents into software systems.
- **F4. HCI features.** Identification and categorization of HCI features and specifications of conversational agents, including verbal communication, non-verbal communication and appearance features. Evaluation of the impact and relevance of such features in the literature.
- **F5. Design dimensions.** Enumeration and depiction of high-level software architecture characteristics in software systems based on the integration of a conversational agent.
- **F6. Technical implementation specifications.** Enumeration and depiction of technical development features, specifications and technologies for conversational agents.
- **F7. Context integration techniques.** Enumeration and depiction of strategies for the integration of contextual data, personalization and self-evolution of conversational agents.
- **F8. Data-sets and data items.** Enumeration and depiction of data-sets of natural language data used for training, testing and evaluating conversational agents in primary studies.
- **F9. Quality and evaluation methods.** Identification and categorization of the quality features, evaluation methods and metrics used for quality analysis of primary studies.
- **F10. Research challenges.** Enumeration and depiction of future research challenges, trends and potential emerging strategies for the successful integration and evolution of the conversational agents’ research field.

Table 1. Cross-reference between research questions (SQ) and features (F)

| Research Question                  | Features                                      |
|-----------------------------------|-----------------------------------------------|
| SQ1. Published research           | F1. Terminology                               |
|                                   | F2. Domains                                  |
|                                   | F3. Goals                                    |
| SQ2. Impact of HCI research       | F4. HCI features                              |
| SQ3. Technical methods and technologies | F5. Design dimensions                       |
|                                   | F6. Technical implementation specifications   |
|                                   | F7. Context integration techniques            |
| SQ4. Training, testing and evaluation | F8. Data-sets and data items                  |
|                                   | F9. Quality and evaluation methods            |
| SQ5. Future research              | F10. Research challenges                      |
3 PUBLISHED RESEARCH (SQ1)

3.1 Terminology (F1)

Being aware of the use of different synonyms in the literature to denote the concept of conversational agents, we investigated in more depth which are these terms and how are they used. The surveyed publications show a lack of dedicated discussion concerning terminology, and there is no taxonomy of reference which is generally used by researchers. While conversational agent and chatbot are the most common terms, it is not clear that there is an evident difference between these concepts. Radziwill and Benton [103] present a classification of software-based dialogue systems, where conversational agents are a sub-class of dialogue systems, and chatbots and embodied conversational agents are both sub-classes of conversational agents. However, this three-level hierarchy is not representative of the findings of this survey. In fact, Syyänen and Valentini [122] demonstrate that terms like chatbot, virtual assistant, agent, conversational bot or even robot are used without an explicit definition in the majority of studies. And when a definition is provided, they are generally used indistinctly among literature with varied definitions, whose differences are not significant. Moreover, a detailed analysis of the terminology used by the studies included in this research demonstrates that only a few studies use the term chatbot as a different term (i.e., a sub-class) of conversational agent [82, 103, 126]. While in some studies the term conversational agent and its variants are used as descriptors to define chatbots [82, 109], the wide majority either use them indistinctly [10, 56] or they present them as synonyms using a shared definition [27, 94, 98, 107].

If we focus on these definitions and the descriptors used, they are generally described as computer programmes [122], tools [98] or software systems [30, 94] which are capable of simulating a human conversation [103, 107, 126] using natural language [56, 70] or even artificial intelligence techniques [84, 98]. Additionally, they are also occasionally described as voice or text interfaces [39, 98] providing access to a set of services or functionalities [61, 103]. All in all, for the sake of consistency and clarity, we propose to build a concise but exhaustive shared definition of conversational agents and chatbots, which will be used as synonyms in this research.

**Conversational agents** or **chatbots** are software-based dialogue systems designed to simulate a human conversational process by processing and generating natural language data through a text or voice interface to assist users in achieving a specific goal.

3.2 Domains (F2)

Discussion on the different domains in which conversational agents have a relevant presence from a scientific perspective is diverse. Some domain-specific secondary studies offer a detailed dissertation of a specific area (e.g., healthcare [72]), while others report a domain taxonomy for a specific research context (e.g., business applications [24]). Based on our feature extraction process, we suggest the categorization covered in Figure 2 which allows us to identify 6 general domains or research areas:

- **Daily life.** The most common domain among research is the support of daily life activities. It is reported as the most frequent domain by de Barcelos et al. [24], Janssen et al. [61] and Knote et al. [70]. The core of this area of application is the use and integration of conversational agents to achieve user’s personal goals in daily life activities. In addition to general-purpose chatbots like question-and-answering (Q&A) agents [26], this domain includes different
domain-specific application areas like tourism [92], restaurants and food [66], games [71], sports [115] and smart-home assistance [32].

- **Commerce.** Activities related to commercial activities, transactions and support are highlighted as the second most present area of application. While being reported by far as the most frequent domain by Bavaresco et al. [10], it is also the second one reported by Janssen et al. [61]. This domain includes e-customer service support chatbots [42] and e-commerce assistance [127] like on-line shopping assistance [141].

- **Business support.** Conversational agents in this domain are designed as software-based tools for employee support to internal business processes. The main goal is to provide an efficient, intuitive interface for traditional business processes to achieve a semi-automatic performance of such processes. These processes include financial tasks [95], negotiation [142] and team working or work support [15].

- **Technical infrastructure.** Bavaresco et al. [10] and de Barcelos et al. [24] report technical management and technical support chatbots in a sub-set of dedicated sub-domains providing access and assistance management to complex technological infrastructures and technical processes. Chatbots reported by literature that can be matched into this category are typically related to complex technical infrastructure management and software engineering assistance. This domain covers from end-user technical support [119] to advanced, semi-automatic management of smart industrial infrastructures [48], including telecommunication systems like autonomous call-centres [88] and tools for bridging the gap in terms of usability between users and software products or tools [53].

- **Healthcare.** Medical or health-related discussion is a matter of dedicated research in several secondary studies in our data-set [72, 84, 109]. However, the healthcare domain is also significantly present in domain-independent literature reviews [24]. Most common areas of application in this domain include support tools for healthcare professionals in medical activities like medication prescription [6] or vital sign control and monitoring [110], and support tools for patient assistance like therapy management [8].

- **Education.** Similarly to the healthcare domain, dedicated research in the education field is also found in our data-set [50, 98], while it is also highlighted as a relevant area in domain-independent research [24]. Education conversational agents can be found as e-learning tools for student assistance in the learning process [124] and as automated tutoring agents [120].

---

**Fig. 2. Aggregated results of conversational agents’ domains in related literature**
3.3 Goals (F3)

Figure 3 shows the result of the feature extraction process for F3 to synthesize and report a general categorization of 6 major goals fitting the goals reported by the literature in the conversational agents field:

- **User support.** The most common goal among conversational agents is to extend software systems with a support tool by integrating human-like communication features into existing activities and processes to enhance user experience. This goal is present in many contexts of use, e.g: in the commerce domain, as customer and service support tools [15]; in the business domain, as semi-automated interfaces of utility processes and for internal support [22]; in the healthcare domain, in the form of therapy assistance as in well-being psychotherapy [8] and treatment support [13].

- **Information request.** Conversational agents designed as commodity query engines giving easy-to-use access to knowledge databases (either generic or domain-specific) compose the second most common goal category. Conversational agents for information request include Q&A chatbots [26] and utility experts for complex domains like autonomous diagnosis or access to health-related information [100]. They also include daily-use, domain-specific utilities like restaurant recommendation [66] and tourist-experience advisor [91]. The information is either directly queried in a set of data sources and reported to the end-user, or processed through information aggregation algorithms to provide personalized answers based on auxiliary data (e.g., user information, context data).

- **User engagement.** Either as a primary or as a secondary goal, most conversational agents’ research cover at some point user engagement with the conversational process itself or with the activity or process the agent is supporting. This engagement applies to a variety of domains, including customer engagement as in e-commerce [29], motivation of the users for a specific activity like patient engagement in physical activity [73], and user engagement in entertainment-focused conversational agents [18]. Research focusing on this goal is primarily oriented to key HCI factors in terms of user interaction and the agent’s communication style.

- **Action execution.** Conversational agents are also used as text or speech interfaces giving access to a set of features from integrated or third-party software systems through a semi-autonomous communicative protocol. This goal is generally observed as a cross-domain objective, including online sale ordering [62], team collaboration task management [125] and even domain-specific expert assistance like healthcare task productivity support [97]. The common feature among these examples is to perform actions and features using natural language (either text or speech) as the interaction bridge between users and the system.

- **User training.** A wide community of chatbot research focuses on evaluating, training and improving user skills in a domain-specific area. This goal is especially prominent in the education field [98], where learning is identified as a primary goal. But user training is not restricted to education: healthcare and medical research focus on training in the means of social skills training for mental spectrum disorders [123], training of psychotherapies like mindfulness [55] and even using virtual patients as a learning tool [58].

- **Information collection.** Finally, and primarily devoted to the healthcare domain, collecting and processing information from users (either through the natural language interaction or, in some cases, through contextual data) is also the main objective in some research scenarios. This goal is mainly observed in healthcare-related activities like symptom monitoring [106], diagnosis [49] and even advanced clinical decision or triage support [116]. Implicit and explicit data are combined to build and explore complex knowledge about user profiling to enhance and support other processes.
3.4 Findings of SQ1

Below we summarize the main insights that can be extracted from the features of SQ1.

Unclear terminology, but a predominant use of conversational agents and chatbots as synonyms. Research literature in the field of conversational agents is not built upon a clear taxonomy of terms and synonyms for the term variants used to refer to dialogue software-based systems. Conversational agents and chatbots are reported as the most common terms, and they are used in a majority of contexts as synonyms.

Domain/goal alignment. The proposed taxonomies allow to identify some synergies between the identified goals and domains. User training is exclusively covered by education (e.g., user training) and healthcare (e.g., patient rehabilitation). Information collection is mainly related to healthcare (e.g., symptom monitoring). Action execution is predominant in commerce (e.g., shopping) and business support (e.g., scheduling). User support, information request and user engagement are majorly reported as cross-domain goals.

User-centred goals. User support, user engagement and user training require the active involvement of the user in the design and validation of the system. Consequently, the need for dedicated research in terms of user experience, user engagement and HCI is reinforced by these findings.

User adherence beyond proof-of-concept tools. User engagement to conversational agents integrated as frequently used tools is a major challenge in some areas. Specific domains like healthcare, education and technical infrastructure seem to benefit from the user-oriented features of their disciplines. Consequently, the focus on the design of conversational agents as user-centred tools is key for a successful adherence in specific domains.

4 IMPACT OF HCI RESEARCH (SQ2)

4.1 HCI features (F4)

The analysis and integration of HCI features in conversational agents is subject of dedicated discussion as the main research goal in some literature reviews (e.g., Van Pinxteren et al. [126], Feine et al. [30]). In these studies, the authors propose their own taxonomy of what they refer to as social cues or communicative behaviours, respectively. On the other hand, some literature reviews (e.g., Rheu et al. [107], Janssen et al. [61]) provide some valuable insights in the form of a narrative discussion as an enumeration of features and HCI-related conclusions supported by examples.
In our research, we focus on the concept of HCI features or *social cues* as defined by Feine et al. [30], where these are defined as design features that trigger a social reaction of the user towards the experience of use of the conversational agent. Under this definition, they propose a flat taxonomy of 42 social cues under a 4-class categorization based on communication features:

- **Verbal** cues, including features about the content of the agents’ responses (e.g., small talk) and the communicational style (e.g., dialect).
- **Visual** cues, including kinesics (e.g., movement), proxemics (e.g., background and conversational distance), appearance (e.g., physical attributes of the agent), and computer-mediated communication techniques (e.g., use of buttons).
- **Auditory** cues, including voice qualities (e.g., voice pitch) and vocalizations (e.g., laugh).
- **Invisible** cues, including chronemics (e.g., response time) and haptics (e.g., tactile touch).

On the other hand, Van Pinxteren et al. [126] propose a two-dimension taxonomy of communicational features. The first dimension is a classification based on the type or *modality* of such features, similar to the classification proposed by Feine et al. [30] but less detailed and more limited in the type granularity. The second dimension is based on the intention, the purpose or the *footing* of these features, for which they propose a 3-class categorization:

- **Human similarity**, which encompass the degree to which the user perceives that a conversational agent is relatable as a human being. These features are user-independent. They apply to general HCI rules and they are not submitted to context or individual differences (e.g., congruency of body gestures, physical appearance).
- **Individual similarity**, which encompass the degree to which users relate as individuals with the agent. These features are user-dependent. They apply to the capability of a conversational agent to match and adapt to the individual features of the user’s needs, preferences, goals and personal communication styles (e.g., personality, communication style).
- **Responsiveness**, which encompass design factors that determine how a conversational agent reacts to a specific user interaction. These features can either be user-independent (e.g., politeness, small talk) or user-dependent (e.g., empathy, social praise).

All in all, while Van Pinxteren et al. [126] report a deeper approach in terms of verbal features type classification, Feine et al. [30] offer a broader analysis scope by introducing the footing dimension as a criterion for structuring and describing communicational features. Therefore, we conclude as a task of significant value to cross-reference both taxonomies to compensate the strengths and weaknesses of both approaches and to integrate the knowledge and dissertation they provide.

In addition to these taxonomies, Rheu et al. [107] use a narrative approach to discuss major themes related to perceived trust in the use of conversational agents, including social intelligence of the agent, voice characteristics and communication style, anthropomorphic look of the CA, non-verbal communication, and performance quality. Among the main conclusions, they emphasize human-like similarity including verbal communication and expressions as well as non-verbal features, although they stress out the impact and the effect of the latter, especially in embodied conversational agents. Additionally, they also discuss the importance of individual similarity in terms of personification of the agent that suits the users (e.g., formality, ethnicity, communication metrics). Janssen et al. [61] depict a chatbot taxonomy for design dimensions from three different perspectives: intelligence, context and interaction. Regarding the latter, they mention a subset of 7 design dimensions which can be mapped to either one of the two complete taxonomies previously mentioned. Some examples include interface personification (i.e., disembodied vs. embodied) and user assistance design (i.e., reactive communication vs. proactive communication).

After aligning the taxonomies defined by Van Pinxteren et al. [126] and Feine et al. [30] with the aggregated knowledge obtained from the rest of studies, we harmonized the structure and
definitions of these HCI features. The results of this process are summarized in Figure 4, and the contributions of such results can be summarized as follows:

- A two-dimension taxonomy of **69 HCI features**. The first dimension is the modality or type taxonomy of HCI feature as proposed by Feine et al. [30], while the second dimension is the footing as defined by Van Pinxteren et al. [126].
- A synthesis on the **impact evaluation** regarding how each HCI feature affects the communicational process between the user and the agent using a 5-value scale: 37 with positive impact (+); 3 with negative impact (-); 17 with mixed conclusions, typically depending on the context or scenario of use (+/-); 3 with neutral or undetectable impact (=); and 10 with no clear discussion or with missing empirical evaluation conclusions (?).
• A heat-map of the use and frequency of these HCI features based on their mentions in the secondary studies included in this feature, which allows identifying specific categories from the cross-referenced taxonomy with different levels of maturity and relevance.

4.2 Findings of SQ2

In this section, we summarize the main insights that can be extracted from the results of SQ2.

**Research focuses on verbal content features regarding human similarity and responsiveness.** As reported during the analysis of the design and technical specifications (F6), the building of a consistent and adequate knowledge base for the appropriate natural understanding and dialogue management performance is essential for a successful user experience. Moreover, adaptation in terms of contextualized and personalized verbal content is one of the main sources of adaptation reported in recent literature (F7). All in all, research is clearly positioned on the importance of exploring and analysing the impact of verbal content HCI features. When focusing on verbal communicational style, reported HCI features are less frequent, and literature generally reports mixed conclusions with respect to their impact, especially from the responsiveness dimension of the agent (e.g., a proactive character is not always perceived as a positive feature).

**Absence of verbal content features towards individual similarity.** The only reference is the capability of the agent to express content from past conversations with the users, which on the other hand is reported to be undoubtedly positive in terms of trust and perceived naturalness of the agent. As reported in the context evaluation (F7), the potential and benefits of integrating contextual and historical data not only in terms of past interactions but also as extended knowledge sources from the user environment is a key feature to be further investigated to enhance the user experience and improve user engagement. Consequently, there is room for research focused on content adaptation and personalization towards individual similarity based on users’ unique needs.

**Appearance features are more important towards individual similarity rather than human similarity.** Individual similarity in terms of visual proxemics and appearance are generally reported to have a positive impact on user experience and user engagement. On the other hand, human similarity in terms of kinesic features and appearance report mixed or context-dependent results in terms of impact, as reproducing some human behaviours (e.g., eye monitoring, facial features) can have a negative impact in some contexts in terms of anxiety and performance.

**Computer-mediated communication can have a positive impact on user experience.** These communicational features allow complementing communication messages through alternatives to natural language (e.g., buttons, emoticons). Literature proves that these techniques not only help to achieve human similarity but also can be a useful approach for other considerations like to reduce typing effort and avoid human errors (F6), which consequently have a significant impact on relevant quality characteristics of the field like the robustness of unexpected input (F9).

5 TECHNICAL METHODS AND TECHNOLOGIES (SQ3)

5.1 Design dimensions (F5)

Research literature reports high-level design dimensions for conversational agents in the form of categories and subcategories proposing a taxonomy based on design specifications. By synthesizing this knowledge, in Figure 5 we report a summary of 6 design dimensions based on the taxonomy proposed by Adamopoulou and Moussiades [4] and the aggregated knowledge from the research literature included in our study. We describe these dimensions as follows:
• **Prescriptiveness.** Hussain et al. [56] present two major design categories for classifying conversational agents based on their goals: *task-oriented* and *non-task-oriented*. *Task-oriented* agents are defined as short-conversation agents designed to execute a particular action from a known sub-set of pre-configured tasks triggered by the conversational process. An online shopping chatbot designed to assist users in their shopping process searching products and solving order-related questions [141] is an example of a *task-oriented* agent. *Non-task-oriented* agents aim to simulate a human-conversational process without a specific task or action as the main goal of the user interaction with the agent. Leisure or entertainment agents like Cleverbot [18] fall into this category. With respect to *non-task-oriented* conversational agents, Nuruzzaman and Hussain [94] differentiate between *conversational* and *informative* agents. While *conversational* agents match the definition of *non-task-oriented* by Hussain et al. [56], *informative* agents are defined as a type of *non-task-oriented* agents which do not pursue a specific activity or task to be executed, but the interaction and the conversational process has the purpose of collecting information. Q&A and service support chatbots fall into this category [32, 66, 91]. *Task-oriented* agents typically imply either the agent is integrated as a tool or sub-module of another software system providing a set of features to the user, or the agent requires integration with third-party software services to perform these actions.

• **Knowledge base.** A major design key factor is the scope of the knowledge base that conversational agents support in their conversations, as well as the level of granularity. This design feature has an impact on the data sources (in terms of number, variety and complexity) required to train and build the knowledge base. Adamopoulou and Moussiades [4] differentiate between *generic* and *domain-dependent* knowledge bases. Regarding the former, *generic* knowledge bases offer information and conversation topics from any domain. An example of this category is a self-adaptive crowd-powered agent designed to learn and adapt to new contexts and topics based on user interaction [54]. On the other hand, *domain-dependent* agents include open or *cross-domain* (i.e., integrating multiple or several knowledge bases and domain data sources [112]) and *closed-domain* (i.e., focusing on a single, expert knowledge base [74]), which are the two main categories covered by Hussain et al. [56] and Nuruzzaman and Hussain [94]. *Generic* and *cross-domain* chatbots typically require auxiliary NLP techniques to process and contextualize user input into a specific topic or domain, like co-reference resolution techniques [4]. This is a consequence of the broader scope of the conversation and user intents and entities. On the other hand, *closed-domain* chatbots are more likely to be tightened to a specific, well-known subset of intents and entities.

• **Service.** The service dimension is defined as the type of relationship established between the conversational agent and its users, based on the needs fulfilled by the agent. Nuruzzaman and Hussain [94] differentiate between *interpersonal* and *intrapersonal* agents. *Interpersonal* agents do not build a personal relationship with the user and are focused to provide a specific service based on a general user categorization. An example of this category is a conversational agent for restaurants’ recommendation and reservation [66]. On the other hand, *intrapersonal* agents are focused on personal scenarios where the chatbot helps users fulfilling personal tasks in their personal life, and therefore there is a personification of the service based on user needs. For instance, smart home assistants for the elderly [32] fall into this category. While *interpersonal* agents are typically context-independent and user-independent, *intrapersonal* agents are designed with user profiling techniques and user configuration mechanisms to fit and adapt to the user’s needs and personal goals.

• **Response generation.** This design dimension relates to the mechanism used by the conversational agent to generate an adequate response message and/or response action (if any) based on the user natural language input messages. State of the art primarily differentiates
two major categories of solutions: deterministic mechanisms and AI-based mechanisms. Deterministic algorithms process user input messages to extract some kind of structure, interpreted knowledge, and apply some kind of deterministic strategy to link this structured data to a specific output message or action. On the other hand, more recent strategies are exploring the potential of AI-based strategies, which integrate the use of machine learning and deep learning models to process user input and build output messages based on the knowledge sources and training data. Mainly, there are two types of AI-based strategies: retrieval-based and generative-based. As defined by Adamopoulou and Moussiades [4], retrieval-based systems use ML/DL models and techniques to predict the most accurate response from a closed set of responses using an output ranked list of possible answers. On the other hand, generative-based systems focus on using Deep Learning models to synthesize and build the reply to a specific user input, rather than selecting it from a closed data-set of responses.

- **Interaction.** This dimension defines the communication mechanism used by the conversational agent to process user information and to generate responses to the user. The core of related research is focused on the design and development of natural language interfaces supported by some kind of natural language pre-processing pipeline or natural language understanding module (see Section 5.2) to extract interpretable knowledge from natural language messages. To this end, conversational agents use text interfaces, voice recognition or a combination of both, potentially adapting the medium to the needs and the context of the user at each stage of the dialogue [107]. Typically, agents integrating voice interaction with their users introduce some kind of speech-to-text and text-to-speech or automatic speech recognition systems as a top layer of the conversational process to support both speech and text and integrate the processing techniques of natural language data. This is the case of well-known commercial voice assistants like Alexa or Cortana [52]. In addition, some studies introduce complementary interaction mediums to overcome the limitations of natural language using alternative data formats. Adamopoulou and Moussiades [4] report image processing as a valuable mechanism for user interaction. Some agents introduce image recognition as a feature to support and extend the limitations of natural language communication, supported by other commercial, popular agents like Siri or XiaoIce [113].

Fig. 5. Aggregated results of conversational agents’ design dimensions in related literature
• **Human aid.** Depicts the degree of autonomy in which the conversational agent can be handled, whether it is designed as a *human-mediated* or an *autonomous* agent. As depicted by Adamopoulou and Moussiades [4], *human-mediated* refers to agents which require from human computation at some point in the conversational process to be operated [74]. On the other hand, *autonomous* agents are fully operated autonomously by users without human-assistance in the loop. While information processing speed might become critical in *human-mediated* agents due to the need of integrating human-aided steps, some contexts of use like crowd-sourcing business services [54] require from human workers to adapt and extend the agent’s knowledge base at runtime.

### 5.2 Technical implementation specifications (F6)

Analysis of the methods, techniques and technologies used for the implementation of conversational agents is one of the most frequent subjects of dedicated research in our study data-set. Some secondary studies integrate technical specifications as a minor goal of their research, providing a flat enumeration of techniques and technologies aligned with relevant examples [82]. Contrarily, we also found dedicated surveys for the analysis of technical aspects presenting taxonomies for technical development, including domain-specific [109] or cross-domain [61]. The focus of the technical research is on the natural language processing and knowledge interpretation tasks regarding the smart communicational process with the user. We identify four major topics involved in the implementation of conversational agents: natural language processing (NLP), natural language understanding (NLU), natural language generation (NLG) and dialogue stage tracking (DST). Figure 6 summarizes the methods, techniques and technologies reported by research literature.

Concerning NLP techniques, given that natural language is the essential core of the conversational process of the majority of chatbot interfaces, these are typically included as a top layer between the user input interface (e.g., chat interface, voice recognition) and the natural language interpretation module responsible for interpreting these messages. Enumerations of NLP techniques are present in research literature, but they are frequently presented as a simplified, preliminary stage [12] or even as a secondary task [2]. We suggest classifying these techniques into three categories: **baseline techniques**, **complementary techniques** and **advanced knowledge interpretation**.

- **Baseline techniques.** Used commonly as preliminary tasks to any NLP pipeline, base techniques are used to structure, clean and assign basic grammatical annotations to the user input messages [12]. They are generally integrated with independence of the integration of a NLU module. Base techniques include tokenization, part-of-speech (POS) tagging, sentence boundary disambiguation and named entity recognition.

- **Complementary techniques.** As an extension of base techniques, complementary techniques are typically applied before advanced NLU techniques to obtain extended grammatical and semantic knowledge [82]. Complementary techniques include co-reference resolution, lemmatization, dependency parsing and semantic role labelling.

- **Advanced knowledge interpretation.** Advanced knowledge structures like the use of vector-space models are the bridge between pure NLP techniques and the integration of NLU modules. This advanced interpretation allows transforming text corpus data into a knowledge base for the conversational agent communicational process [80]. These techniques include vector-based representations like TF-IDF and skip-gram models like Word2Vec.

Concerning NLU, NLG and DST techniques, and as discussed in Section 5.1 (F5), secondary studies discussing the technical development of the NLU and NLG mechanisms agree on a general categorization of two main approaches: **rule-based** approaches and **AI-based** approaches. We depict each of these methods and techniques as follows:
• **Rule-based.** Based on the definition in Section 5.1, we identify four technique categories: *fixed input, pure NLP, vector-space and pattern-matching.*
  
  – **Fixed input.** Fixed input agents restrict the conversational process to a closed set of possible inputs to select at each stage of the dialogue. Despite being only explicitly mentioned by Safi et al. [109], and even though pure fixed input does not require a natural language module, Janssen et al. [61] present hybrid NLU approaches in which fixed input mechanisms at a specific stage of the dialogue are used to reduce typing effort and human errors by simplifying the user interaction. Some examples include interactive elements (e.g., buttons) [60] or multiple-choice selection of a dynamically updated list [14].
  
  – **Pure NLP.** Some approaches integrating basic text understanding techniques in conversational agents integrate the use of base and complementary natural language processing techniques. These techniques are complemented alongside search algorithms for proper response selection through pre-programmed rules triggering specific answers when finding specific keywords [131] or more advanced techniques like ontology-based matching through dependency parsing [41].
  
  – **Vector-space.** Advanced knowledge interpretation techniques can extend the NLP module to build interpreted domain knowledge through vector-space modelling techniques. These models can be used as mapping modules between user inputs and chatbot outputs in a variety of contexts. For instance, TF-IDF models can be used to build expert knowledge bases [90], and vector-space models can also be exploited by more complex information retrieval (IR) algorithms for complex domain-specific queries [31].
  
  – **Pattern matching.** The most representative approach for rule-based agents is the development of a pattern matching approach through popular, standardized technologies like Artificial Intelligence Markup Language (AIML) [36], ChatScript [134], Cleverscript [78], or RiveScript [99]. These are examples of open-source interpreted languages defining the syntax to build a set of templates to identify patterns of user input messages and their link with patterns of output responses. These syntaxes are designed using well-known data-interchange formats like XML (i.e., AIML), JSON (i.e., Cleverscript) or simplified custom formats (i.e., ChatScript, Rivescript). Despite their conceptual, syntactical and technical differences, all of them allow to define a knowledge base through a set of dialogue management rules. Through input processing, they apply pattern recognition techniques to find the most suitable template and proper response selection for a specific user input. These approaches offer an easy-to-develop, efficient and effective method to manually define limited knowledge bases, perfectly suitable for closed-domain and specific-purpose chatbots where a limited conversation scope is expected. These contexts of use include examples like automated medical chatbots [105], academic advisors for student assistance [76] or a website-based chatbot for e-commerce support [45].

• **AI-based.** Based on the definition in Section 5.1, we also identify four technique categories: *ensemble learning, I/E classifiers, web services and neural networks.*

  – **Ensemble learning.** In the context of conversational agents, ensemble learning models like decision trees or random forests are used for input text classification through a voting approach to determine the most suitable response through the tree paths definition and the user answers to the chatbot questions. These decision trees are built and can be dynamically adapted by adding new nodes (i.e., new user inquiries) through new user interaction, and consequently tracing new paths and adapting to unknown conversational patterns. Ensemble learning models are used in the context of Q&A chatbots for domain-specific contexts like education advisors [85] and autonomous diagnostic agents [65]. Given the significant amount of conversational paths a conversation might undertake, they are
Conversational Agents in Software Engineering

complex to develop and maintain [20]. However, in some contexts, they have proven to provide a higher accuracy than alternative methods like fuzzy logic approaches [12].

- **Intent/entity (I/E) classifier.** A standardized approach for developing NLU modules is the use of machine learning classifiers for automated, adaptive intent and entity recognition. In this domain, intents are defined as the user intentions, while entities are the topics these intents refer to. The use of pre-trained models through manual annotation and model evolution through runtime user interaction allows processing user input messages and to predict with a certain probability the most accurate response based on the intents and entities predicted from that message. I/E classification can be developed using a wide variety of classifiers, including Support Vector Machine (SVM) [86] and fuzzy logic (clustering) [20].

- **Web services.** Given the exponential growth of conversational agents in a wide variety of domains and business cases (as reported in 3.2), several web-service solutions for their development, deployment and maintenance have emerged in recent years. They are designed as web-based service providers for designing and training as well as operating chatbots in an easy-to-use, decoupled and scalable environment. As reported by Safi et al. [109], these solutions typically allow the combination of machine learning (i.e., I/E classifiers) and rule-based approaches for NLU tasks. Some examples include commercial solutions like Google Dialogflow[^3], Amazon Lex[^4], IBM Watson[^5] and Microsoft Luis[^6]. Despite being proprietary software solutions, all of them include some kind of free-tier service (e.g., limited by the number of features or the amount of data that can be processed). Among open-source solutions, the most popular is Rasa[^7], which has become a technical reference in the conversational AI community.

- **Neural networks.** Artificial intelligence neural networks are the latest technical contribution to the conversational agents’ field. As described by Adamopoulou and Moussiaides [4], user input messages are transformed into vector representations through word embedding techniques [83]. These vectors are used as input features of the neural networks, which can be used to either predict (i.e., retrieval-based) or create (i.e., generative-based) the response. A special type of neural networks has become more popular among conversational agent developers due to its capability of integrating contextual knowledge with respect to previous conversations. Recurrent Neural Networks (RNNs) define an architectural loop in which output knowledge data is fed as input features to the network cells continuously. Among RNNs, Long Short-Term Memory neural networks (LSTM) allow conversational agents to differentiate between long and short-term memory. Long-term memory refers to the general knowledge data of the model for conversation prediction and generation, while short-term memory refers to knowledge data that is only valid and relevant for a specific time window in which context is defined by recent interaction between the user and the agent [80]. This time window, which defines how cells behave and how they manage its contextual knowledge, can be configured according to the needs and nature of the agent. RNNs and LSTM are frequently used in generic or open-domain agents. Some examples include a highly accurate chatbot of frequently asked questions for customer service based on a retrieval-based approach [87] and a proposal of a generative-based, knowledge-grounded chatbot model [67]. When focusing on generative-based approaches, the use of Sequence-to-Sequence (Seq2Seq [121]) models is becoming predominant. Seq2Seq models

[^3]: https://cloud.google.com/dialogflow/docs
[^4]: https://docs.aws.amazon.com/lexv2/latest/dg/what-is.html
[^5]: https://www.ibm.com/watson
[^6]: https://azure.microsoft.com/en-us/services/cognitive-services/language-understanding-intelligent-service/
[^7]: https://rasa.com/
use LSTM neural networks as layers to map the user input sequence into a target sequence, which is the generated response of the agent. As representative examples of context-aware and generative-based conversational agents, we highlight a generic emotionally aware chatbot for user engagement through open-conversation [143] and an open-domain question-answering chatbot for user support [102]. Complementary, some approaches use convolutional neural networks (CNN) not only for intent/entity classification [104], but also in the context of image processing [113] as a complementary interaction mechanism with users for an extension of the knowledge base of the conversational agent.

5.3 Context integration techniques (F7)
Integrating contextual data is reported as a key feature and as a future research challenge at some point by most of the secondary studies in our data-set. While being limited, research literature discussing context integration focus on how conversational agents integrate context into their systems to provide a personalized, self-adaptive and context-aware user experience as defined by Bavaresco et al. [10] and Janssen et al. [61].

Deep learning strategies like RNNs and LSTM introduce a context-adaptation layer based on a specific conversation-related context through recent user messages, which allows topic-centred analysis and runtime feedback to increase personalization and accuracy. But context integration strategies are manifold. Knote et al. [70] define context data as the knowledge that can be extracted from physical (e.g., location, temperature, humidity) and logical (e.g., calendar entries, application’s portfolio) environments. According to the feature extraction analysis from secondary studies, there is no formal analysis or categorization on how to integrate context into conversational agents. Instead, we find a summary enumeration of general techniques or specific examples.

Bavaresco et al. [10] define three focused research question concerning self-learning, personalization and generative-based methods in conversational agents which they address through a narrative approach commenting examples of their primary studies data-set. These examples are described emphasizing the data sources of contextual knowledge and the mechanisms used for context integration. Some examples include user personalization features and user’s past interactions with the agent for knowledge base adaptation to domain-specific queries [119], user historic feedback regarding the agents’ responses to re-train the generative models [140], and user historical data (e.g., user’s purchase in an e-commerce website) from the software system in which

![Fig. 6. Aggregated results of methods, techniques and technologies in related literature](image-url)
the agent is integrated or from third-party services for recommendations [111]. Kocaballi et al. [72] present an exhaustive enumeration of examples related to a list of personalized contents and context-adaptation purposes or goals. For the former, some examples include progress towards the goals set and communication style adaptation, as well as the use of reminders, warnings and alerts sent by the agent to this end [114], and multimedia adaptation like customized activity graphs based on user activities as well as monitoring questions on these activities [73]. For the latter, some examples include improving user engagement, delivering adaptive training [55] and support self-reflection activities like analysis on self-reported symptoms of depression [57].

Knote et al. [70] enumerate a few examples of context changes reactions, focusing on the mechanism and the object of these reactions or changes on assisting users on a specific task like semi-automatic purchases [128] or adaptive text or voice responses [25]. Finally, Janssen et al. [61] simply limit to report that self-adaptive personalized, as well as context-aware agents are a significant minority in the data-set of chatbots they cover, with a significant presence of short-term relationship chatbots which do not consider context data. Despite being a recent contribution, they argue that this is a direct consequence of the rather complex design properties and technical specifications involved in these dimensions.

All in all, we identified 4 different dimensions in which context integration can be described: the purpose or main goal of the context integration strategy; the source or trigger of the adaptation; the mechanism used for integrating context into the conversational experience; and finally the object of adaptation, which is the entity or process affected and adapted by the context-aware knowledge. A categorization of the examples covered by the previous studies allows us to build and depict a general approach based on how context adaptation and personalization are integrated into conversational agents. Figure 7 is the result of such harmonization process.

5.4 Findings of SQ3
In this section, we summarize the main insights that can be extracted from the results of SQ3.

From a technical perspective, there are two major categories of conversational agents. We propose a classification of two major types or generations of conversational agents: deterministic and AI-based approaches. We define each of these categories through a sub-set of 6 major characteristics based on the feature extraction process. These characteristics are summarized in Table 2.
Deterministic conversational agents are the first generation of dialogue systems. This type of conversational agents are designed as a fixed knowledge conversational base, which allows a deterministic mapping between user input and responses output from a closed data-set of responses. They are developed using standardized rule-based tools like pattern matching techniques, which are reported as predominant in literature. These strategies, while representing an early, first generation of conversational agents, provide an important advantage in terms of their simplicity, their maturity, their efficiency of development and the lack of need for training data.

AI-based conversational agents are the second generation of dialogue systems. This type of conversational agents are designed as an adaptive knowledge conversational base, which is capable of processing and interpreting user input and either predict or create original generative output as a response to that user input. These approaches are an emerging trend that serve from up-to-date machine learning and deep learning techniques (e.g. RNN, LSTM, Seq2Seq) which benefit from in-dialogue context integration and self-adaptation to provide an adaptive conversational smart process. They require large data-sets and expert knowledge to be developed while offering a broader and more personalized communication experience with a higher specialization degree.

| Characteristic   | Deterministic approaches | AI-based approaches |
|------------------|--------------------------|---------------------|
| Behaviour        | Deterministic            | Non-deterministic   |
| Output           | Static                   | Dynamic             |
| Knowledge base   | Fixed                    | Adaptive            |
| Complexity       | Low                      | High                |
| Maturity         | High                     | Low                 |
| Pervasiveness    | High                     | Low                 |

Table 2. General categorization of conversational agents design approaches based on F5 and F6

Deterministic and AI-based are complementary approaches. The advantages and disadvantages of each approach can be used for choosing the best solution for a specific scenario, or even for the design and deployment of hybrid approaches. Hybrid design can serve from rule-based mechanisms for deterministic response generation in combination with generative capabilities and self-adaptive mechanisms through context-awareness to increase the degree of precision and accuracy of the conversational process. Consequently, an accurate and more precise conversational process can be integrated while increasing the satisfaction of the user when interacting with the agent through additional characteristics like human or individual similarity (F4) and improving the performance in terms of quality evaluation in terms of task or functional effectiveness (F9).

Context-awareness is yet to be fully explored. There is a lack of synthesized, structured knowledge regarding context-awareness strategies for full-personalized conversational agents. The surveyed studies and the results of F7 can be used to identify 3 major research lines in terms of context-awareness strategies: (1) integrating additional data sources (e.g., user profiles, third-party services); (2) exploring and categorizing adaptive mechanisms (e.g., adaptive knowledge base, runtime feedback collection), including HCI feature adaptation (e.g. communicational style, character/mood of the agent); and (3) extend the scope of adaptation beyond the agent’s response verbal content (e.g., connected smart objects, third-party applications).
6 TRAINING, TESTING AND EVALUATION (SQ4)

6.1 Data-sets and data items (F8)

Despite being a critical issue for training, testing and evaluating conversational agents, information and discussion on data repositories, data-sets and data items is missing from almost all secondary studies included in this research. Given that dissertation is limited, and no formal categorization is provided, we propose a narrative approach to report and comment on this feature.

Dsouza et al. [27] analyse and categorize data sources according to the domain-specific primary studies covered by their healthcare-focused research. Mainly, they identify three data sources: medical knowledge repositories, user information databases and conversation scripts. Medical knowledge repositories store domain-specific information data (in this case, healthcare-focused) related to the area of application of the conversational agent. These include private or internal data repositories storing medical information like brain and facial images collected through traditional medical activity [129], and public or online available sources like disease-specific documents and information (e.g., asthma) from Wikipedia or other public disease repositories [63]. User information databases are reported to be used especially for context adaptation and personalization tasks (as depicted in 5.3), but no further insights are reported. Some examples include the use of demographic data [63] and patient’s electronic health record [81]. Finally, a minority of studies use publicly available conversation scripts like call-centre and frequently asked question internal logs [1, 137]. It is relevant to highlight that a significant amount of primary studies do not mention the type of data used for either training, testing or evaluation.

On the other hand, Safi et al. [109] identify and report the data-sets used by primary studies for evaluation analysis. Although most examples are healthcare-related, they also report commerce and generic-purpose conversational agents. In alignment with Dsouza et al. [27], healthcare-focused primary studies use public or private medical knowledge repositories like scrapped data from medical-related forums [11] or disease-symptom mapping knowledge databases [104]. Additionally, two additional insights can be extracted from these examples. First, for generic design approaches, conversational scripts like Telegram [28], public chat repositories [101] or the use of public, specific-purpose repositories like DBpedia [9] or Stanford CoreNLP [44] which provide academic data-sets for training, testing and evaluating in domain-specific contexts like commerce transactions [46] or any closed-domain chatbot design [93]. And second, as briefly mentioned by Adamopoulou and Moussiades [4], the use of publicly available corpora reported by primary studies might be used for inferring training features and evaluation strategies in domain-related conversational agents.

6.2 Quality and evaluation methods (F9)

Alongside technical implementation specifications (F6), analysis on how research literature addresses the evaluation of conversational agents is one of the most frequent subjects of dedicated research, from a general overview of evaluation methods [72] to full-dedicated discussion [19]. We classify the discussion based on two different approaches: analysis on quality characteristics (i.e., what is evaluated) and on evaluation methods and metrics (i.e., how they are evaluated).

For quality characteristics, we propose a feature extraction process based on the ISO/IEC 25010 software product quality model [59]. We extract and compare the quality characteristics discussed and reported in the secondary studies covered by these features, and we align them with the ISO/IEC 25010 quality characteristics and sub-characteristics. This feature extraction process allows (1) to organize the evaluation of quality characteristics following a standardized, well-known framework, and (2) to exhaustively compare traditional software quality characteristics with the most frequent and relevant characteristics in the conversational agents’ field, which allows us to detect and reflect on those quality characteristics which are reported as most important by the literature. Figure 8
summarizes the results of this alignment analysis. For simplicity, we only include those quality characteristics and sub-characteristics from ISO/IEC 25010 to which a match was found among the discussion reported by secondary studies.

- **Functional suitability.** The focus on functional suitability lies on the degree to which the interaction with the agent provides accurate responses with the required level of precision, as well as the degree to which these responses facilitate the achievement of users’ goals and the tasks exposed by the agent. Regarding functional correctness, the most common term in literature is effectiveness. Casas et al. [19] differentiate between functional effectiveness, which includes objective measures like command interpretation accuracy and speech synthesis and generation performance, and human effectiveness, which relates to the human similarity footing dimension described in Section 4.1. Milne-Ives et al. [84] identify the process of service delivery as a general quality characteristic involving both task and communication correctness. Therefore, whether explicitly separated or not, the accuracy and correctness of the conversational process is a relevant quality sub-characteristic. Regarding functional appropriateness, relevancy of content or content evaluation is the most mentioned quality characteristic. As defined by Kocaballi et al. [72], personalized content adaptation (e.g., topic suitability and verbal HCI features) is evaluated to achieve appropriateness in terms of users’ goals achievement to match each user’s specific needs. Milne-Ives et al. [84] specifically refer to appropriateness of content and the generated responses, and it is reported as one of the most frequently used in the evaluation of conversational agents.

- **Performance efficiency.** Concerning performance efficiency as defined by ISO/IEC 25010, few examples are mentioned which might be aligned to its quality sub-characteristics. The most common, shared term in this quality characteristic refers to the time behaviour sub-characteristics, to which research generally refers to as performance efficiency [19, 80]. In terms of resource utilization, Milne-Ives et al. [84] report as a major quality characteristic cost-effectiveness in the means of the relation between the cost (i.e., the resources) and the effectiveness characteristic depicted before. It is relevant to notice that Casas et al. [19] and Radziwill and Benton [103] use the concept of efficiency in the context of usability in ISO standards as defined by Abran et al. [3]. Therefore, their insights and comments on this quality attribute must not be considered under this category. Regarding capacity, in Radziwill and Benton [103] report as an example of a quality attribute the appropriate escalation of channel services for assisting user traffic requests.

- **Usability.** Radziwill and Benton [103] propose to align the quality characteristics collected from primary studies with the concept of usability reported in ISO 9241 [3] in terms of efficiency, effectiveness and satisfaction. Regarding the latter, user satisfaction or simply satisfaction, which relates to the degree to which users acknowledge the usefulness of the conversational agent to assist their needs (i.e., appropriateness recognizability), is reported as a major quality characteristic by several secondary studies [19, 72, 84]. Additionally, other studies refine the concept of satisfaction. As defined by Kocaballi et al. [72], emotional awareness of the agent is also identified as a quality attribute to be measured and which contributes to overall satisfaction. Hobert [50] reports further psychological factors related to this satisfaction characteristic like user enjoyment, which is supplemented by Pérez et al. [98] with other HCI-related attributes like humanity, affection or friendliness of the agent. On a secondary basis, learnability or teachability of the conversational agent is also a literal concept from the ISO/IEC 25010 covered by research literature, including synonyms or related approaches like acceptability as defined by Milne-Ives et al. [84]. Remaining sub-characteristics reported in Figure 8 are not generally discussed among secondary studies.
Sub-characteristics like operability, user interface aesthetics and accessibility are superficially discussed and covered by the concept of usability as defined in some studies (e.g., Hobert [50]). Finally, it is relevant to highlight a special mention by Radziwill and Benton [103] concerning user error protection sub-characteristic, which is reported as robustness of unexpected input and identified as the capacity of the agent to avoid inappropriate utterances and apply damage control techniques when manipulation or unexpected input messages are sent.

- **Security.** Finally, security sub-characteristics like confidentiality and integrity are partially covered by the major category as reported by Milne-Ives et al. [84], which is generally summarized into safety/privacy/security concerns. These terms, which are sometimes referred to as synonyms during evaluation, relate to the more restricted concept of trustworthiness (from a security point of view) reported by Pérez et al. [98]. The confidentiality characteristic is emphasized on both subjective perceptions and objective evaluations on privacy concerns, while integrity focus on data operation and management (e.g., conversations with the agent, user private data, manipulation and modification of such data).

For the analysis of quality methods and metrics, we suggest a categorization between qualitative and quantitative analysis methodologies. Figure 9 summarizes different method implementations covered by secondary studies, as well as a list of frequently used metrics for each of these general categories.

- **Qualitative analysis.** We identify two main methodologies for qualitative evaluation: interviews and qualitative questionnaires.
  - Interviews. As defined by Hobert [50], qualitative interviews allow evaluators to get detailed feedback from the participants of an experiment with the conversational agent in regard to its impact and the effects of user interaction with the agent. Based on Kocaballi et al. [72], we observe two different patterns for conducting interviews, based on the primary studies covered in their research: focus group interviews, where participants might be grouped into small groups to facilitate discussion and the exchange and contrast of opinions and impressions [106]; and individual custom interviews, which are designed and carried on separately for each individual based on their experience with the agent and their personal
user experience as in open-ended, semi-structured interviews [114]. For measuring and analyzing the responses of the participants, the most frequent metric is the use of a Likert scale measure for objective questions (e.g., measuring the level of satisfaction of a specific response [114]), alongside discourse analysis on interview transcripts.

- **Questionnaires.** Two different dimensions are extracted from research literature for evaluation through qualitative questionnaires. The first one relates to purpose of the questionnaire, for which in the results reported by Kocaballi et al. [72] we can differentiate between goal oriented questionnaires and user satisfaction questionnaires. Goal oriented questionnaires are designed to measure and monitor features or qualities to which the interaction with the agent has a significant impact. For instance, in the healthcare domain, questionnaires can be used for measuring disease symptoms like depression or anxiety before and after interaction with a conversational agent focused on dealing and treating these pathologies [37]. On the other hand, user satisfaction questionnaires cover the topics related to the usability quality characteristics reported in Figure 9, including examples like emotional awareness, learning and relevancy of content [34]. Additionally, Maroengsit et al. [80] differentiate two levels for user satisfaction evaluation: session-level questionnaires, where users are asked to evaluate an entire conversation session with the agent based on a different set of factors (e.g., appropriateness, empathy, helpfulness [138]); and turn-level questionnaires, where users must evaluate each response from a conversational agent, which might be used for an average rating based on user or expert evaluation [64]. As qualitative interviews, the use of a 5/10-point Likert scale is the most frequent metric used in questionnaires.

- **Quantitative analysis.** We identify two main methodologies for quantitative evaluation: dialogue tracking and quantitative surveys.

  - **Dialogue tracking.** Automatic or semi-automatic monitoring and analysis of the user-agent communicational process (i.e., the natural language messages sent by the user and the responses generated by the chatbot) is considered a key evaluation method for both usability and functional suitability quality evaluations [50, 72, 82]. Message monitoring techniques for meta-data features regarding the user-agent interaction is one of the most common methods, given its simplicity and the relevance of the data in terms of analysing how users interact with agents, which guides the evaluation of quality attributes like user engagement [37]. In addition to explicit natural language feedback metadata features, more advanced methods like transcript discourse analysis, technical log files analysis and content evaluation are used to get deeper insights both into the technical correctness of the agent and its performance. These automatic methods are typically measured through standard numerical forms including prediction quality measures (e.g., accuracy, precision, recall and F-measure) or information retrieval measures (e.g., mean average precision and mean reciprocal rank). These metrics are especially relevant for functional evaluation in context-specific or Q&A conversational agents, for which expert manual annotation process of the agent’s responses can be used for the evaluation measure [22]. Additionally, conversation statistics (e.g., number of messages sent by user, length of these messages and number of chatbot responses rated as positive) and task completion measures like statistical data about the number of tasks successfully achieved are also used to complement evaluation based on the conversational process [98, 103].

  - **Surveys.** With the focus on functional suitability, quantitative surveys allow involving users in the annotation process for evaluating the quality of the responses generated by the agent and the level of achievement of user’s tasks and goals. As reported by Kocaballi et al. [72], the use of in-app feedback questions is a relevant example to progressively measure the task completion ratio or the number of messages evaluated as errors or bad quality
The use of pre or post quantitative surveys can be used to the same end, with the ability to provide quantitative layouts before and after the session, which can later be used to compute and measure the level of task or goal achievement, e.g., as in the learning success of an educational chatbot for a specific skill or state of knowledge [50]. Finally, Maroengsit et al. [80] and Fitrianie et al. [33] highlight the use of expert content evaluation surveys to facilitate the annotation process for expert evaluation of the response or even the ranked set of response candidates of a conversational agent, which is used to report prediction quality metrics like top-N response accuracy or task completion measures [102].

6.3 Findings of SQ4

In this section, we summarize the main insights that can be extracted from the results of SQ4.

There is a research gap on data-set and data-items dissertation. While building knowledge bases for deterministic conversational agents has been traditionally addressed through the dedicated work of domain experts (e.g., elaboration of pattern matching templates), state-of-the-art solutions involving AI-based techniques require a large amount of data to train the models used in these solutions. Consequently, dimensions such as type, format, quality and amount of data items must be researched from a general perspective to address the challenge of gathering data for the training of conversational agents’ knowledge bases. To this end, we suggest to explore and follow the examples of primary studies using public, open-source data repositories of NLP data-sets for dialogue tasks. Research should primarily focus on data-driven approaches in the context of novel technical solutions like the use of recurrent neural networks and context-aware conversational solutions, which face specific challenges in terms of type and amount of available data.

Quantitative evaluation is mainly covered by dialogue tracking methods. These techniques are limited to evaluation methodologies for low-level sub-tasks of the conversational process, including intent and entity classification and natural language response generation correctness. On the other hand, qualitative analysis evaluation is predominantly oriented towards quality characteristics focused on the capacity of conversational agents to accurately match and suit the specific goals and needs of their users. Furthermore, quality sub-characteristics like appropriateness recognizability (e.g., user satisfaction, emotional awareness) or learnability (e.g., teachability,
acceptance/adherence) involving the perception of the user are widely addressed and incorporated into most quality evaluation strategies.

**Qualitative evaluation is designed as a user-oriented method based on subjective metrics.** Concepts like satisfaction or awareness are repeatedly discussed through literature to address and propose different strategies for capturing the degree to which users perceive conversational agents as helpful tools, whether we refer to usability, performance suitability, performance efficiency or even security. However, secondary studies do not provide further details on the design criteria to design and carry on these types of experiments. Given the relevance and the impact that HCI features have demonstrated to have on the user experience with the conversational agent, it remains obscure how these HCI features can be captured and evaluated into an interview or a questionnaire-based evaluation in order to process, model, interpret and compare conversational agents and human behaviours.

7 FUTURE RESEARCH (SQ5)

7.1 Research challenges (F10)

We identify and report 6 major challenges in the field of conversational agents, which we envisage that will guide future research:

1. **Adoption and application of domain-specific and target-specific requirements.** Advanced research in domain-specific scenarios has deeply explored the implications of the design specifications in terms of HCI features and technical development. De Barcelos et al. [24] unveil a recent trend in integrating conversational agents into healthcare (e.g., vital signs monitoring, mental health), education or daily-life scenarios. However, they highlight that the articles show that most people only use conversational agents for simple tasks. A similar conclusion is reported by Syvänen and Valentini [122], as they report as a major future work among primary studies to improve the proof-of-concept conversational agents to become real tools for their respective domains. Similar considerations are made regarding specific target audiences (e.g., elderly, functional diversity) [72, 126]. Research literature lays the groundwork for successfully integrating conversational agents solutions into their software systems. Fully exploiting the potential of some specific scenarios and the characteristics of specific target audiences in real environments remains to be thoroughly explored.

2. **Improve user engagement through perceived quality.** Whereas user communication and user experience is the core of the conversational process, analysing and adapting HCI features to improve user engagement is still a major key challenge. In fact, Syvänen and Valentini [122] claim that the concept of engagement is understudied, and therefore the consequences and the impact on user engagement concerning the agent’s communicational capabilities must be researched with further details. De Barcelos et al. [24] focus on user engagement through the need for improving the perceived quality in terms of usability of the agent and user satisfaction, but also in terms of the perceived security and privacy of the data shared with the agent. Highly advanced taxonomies provide a theoretical background for the categorization of relevant HCI features and their impact on these features. Based on the knowledge extracted from these taxonomies, users’ experience must be investigated from the user’s perception to improve the perceived quality through the conversational process.

3. **Extend verbal content and non-verbal HCI features towards individual similarity.** Concerning HCI features, the aggregated knowledge depicted in Figure 4 allows us to identify relevant research gaps in terms of HCI features according to the modality and the footing of such features. Van Pinxteren et al. [126] report the need for improving communicative behaviours, with special emphasis on non-verbal (e.g., speech rate) and physical (e.g., gestures)
features. In alignment with the integration of context-aware and personalization techniques (F7), individual similarity can be further investigated to improve user engagement through the adaptation of the communicational process. To be specific, response content customization and the adaptation of non-verbal features like auditory or appearance features are some areas where we suggest that research should focus in the near future.

(4) **Achieve full personalization through user profiling and context-awareness mechanisms.** The emerging of recent AI-based technologies like RNNs and recent examples of context-aware conversational agents through the extension of the conversation scope are the fundamental base for future research to achieve full personalization. Bavaresco et al. [10] report a few examples of how chatbots address self-learning, personalization and context-awareness in domain-specific scenarios, which can be generalized to high-level techniques like user profiling or user intentions analysis. De Barcelos et al. [24] also report personalization, user profiling, context understanding and user goals (i.e., intentions) automatic recognition. We envisage that future research will require the design and definition of clear, structured general practices integrating user profiling techniques for autonomous user intention’s detection, the adaptation of the communicational process (as suggested in SQ2) and the integration of contextual data.

(5) **Expand the applicable knowledge of required data in generic and domain-specific scenarios.** The lack of aggregated and discussed data reports a significant gap in terms of shared, structured knowledge in regard to data sources and data items for the training and evaluation of conversational agents. Bavaresco et al. [10] explicitly report that some domains have very limited knowledge base data sources for building and training the models in AI-based solutions. A clear taxonomy of available open-source data repositories based on a set of classification criteria (e.g., domain, purpose, size, type, format, reported evaluation quality, average meta-data features like messages length or speech rate) should help researchers focusing on these tasks as a starting point to evaluate and discuss the expectations of data availability and data suitability for their own research.

(6) **Extend qualitative analysis methodologies to improve perceived quality and appropriateness quality attributes.** User’s perceived quality is an essential characteristic in conversational agents’ quality evaluation. These quality characteristics focusing on perception and appropriateness (i.e., functional appropriateness, appropriateness recognizability) are thoroughly discussed among literature. Milne-Ives et al. [84] claim the need for extending qualitative evaluations to clarify and evaluate user quality perception, and consequently user engagement. Additionally, they highlight the need for identifying all the structural, physical and psychological barriers in these evaluation methodologies to identify and recognize how to improve user engagement and the penetration of conversational agents. Consequently, we suggest that further research in the field of evaluation should be primarily focused on standard practices for qualitative analysis towards perceived quality.

## 7.2 Findings of SQ5

In this section, we summarize the main insights that can be extracted from the results of SQ5.

**User perceived quality is a key research challenge.** Research challenges (1), (2), (3) and (6) are related to the degree to which users perceive conversational agents as useful tools. The future evolution of the pervasiveness of these systems is mainly conditioned by the insights and contributions of future research in terms of improving user experience through an advanced, efficient and accurate smart conversational process.
Context-awareness, personalization and data-driven research requires from structured, synthesized knowledge. Research challenges (4) and (5) are related to the degree to which conversational agents are capable of building customized and personalized knowledge bases with the required precision to provide fully-personalized experiences. As reported by the surveyed studies, there is a lack of structured, synthesized knowledge in these areas. Consequently, identifying and exploiting data sources and data items is a key challenge not only for the design of these systems, but also for its evolution and personalization.

8 RESEARCH METHOD FINDINGS

As a closure of our research, below we report a summary of the main findings of the research method depicted in this survey. The purpose of this report is two-fold. First, to validate and reinforce the validity of tertiary studies as well as the details on the depicted research method in Section 2. And second, given the highly evolving nature of the covered research area, we envisage that future research will require from up-to-date surveys in the field, for which we expect that the lessons learned from this research can serve as an input for future researchers.

Bibliometrics. Based on the final data-set of selected publications, we analyse two bibliometric metadata features: distribution of secondary studies per year of publication (Figure 10), and distribution of the number of primary studies covered by each secondary study (Figure 11). Year of publication distribution reports an incipient trend in the conversational agents’ field, which is reflected in an increasing growth of the publication of secondary studies in recent years. Only one of the studies [2] was published before 2016, while the remaining 24 are comprised between 2017 and 2020. On the other hand, Figure 11 reports the distribution of the number of primary studies covered by the included literature reviews. This increasing tendency of published literature reviews and the significant amount of primary studies included by these studies demonstrate that conversational agents are a subject undergoing intense research. Additionally, it also reinforces our selection of tertiary studies as the research method for conducting this survey.

Search strings. We used the search string depicted in the surveyed secondary studies when explicitly depicted (18/25) as proxy for this analysis. As a result, we find a clear evidence on the prominent use of conversational agent and chatbot as synonyms in research strategies, which is reinforced by the findings on the terminology (F1). Both terms are predominantly included in the majority of secondary studies’ search strings (11/18, 61%), and almost all of them include at least
one of these terms (17/18, 94%). Therefore, while typically researchers focus on using a single term in their dissertation for consistency, the majority of research strategies covered by this study include both chatbot and conversational agent in their research method. Alternative terms like chatterbot, conversational interface or dialogue system are also frequently used, but they have a significant minor presence. Below this frequency rate, up to 41 synonyms are used in two or fewer studies, including chatbot variants (e.g., chat-bot, chatterbox), agent variants (e.g., embodied agent, personal attendant) or conversation variants (e.g., conversational system, conversational assistant). This exhaustive analysis of the search strings is useful for researchers performing search-based literature reviews in the domain, as it allows to define the search string more accurately.

**Digital libraries.** Regarding the digital libraries (DL) used for automated search (explicitly covered by 17/25 secondary studies), the most common database is ACM Digital Library, followed by IEEE Xplore, Science Direct, EBSCOhost, Google Scholar, Scopus, Web of Science and EMBASE. With a frequency rate below 4 studies we mostly observe domain-specific databases (e.g., PubMed) or publisher databases whose results are generally covered by other databases like Scopus (e.g., Springer Link). Details on the use of most common digital libraries are depicted in Figure 14. The use of less popular digital libraries like EBSCOhost and EMBASE can be justified by the relevance of related literature in the field of healthcare and education, as well as dedicated research on the social implications of the use of chatbots.

**Research purpose.** We inductively categorize the secondary studies under six classes, based on the focus of the research. The most common research goal is (1) the analysis of design and technical implementation of software systems based on the integration of conversational agents (9/25, 36%). The level of detail on these technical specifications is diverse, from a summarized report on major design-key technical specifications [61] to a technical-oriented report, categorization and depiction of design and development specifications [109]. Discussion on (2) the role of conversational agents and their design features in a domain-specific field (e.g., health care [72], educational [98], business [10]) is the second most common goal (8/25, 33%). These studies are characterized for reporting the domain-specific implications and relevant features of the integration of conversational agents in their traditional software systems or business processes. The remaining research goal categories are less frequent and include: (3) research on quality evaluation characteristics, methods and metrics, (4) research trends and challenges, (5) social and communicative features analysis and (6) a taxonomy proposal for classifying conversational agents based on high-level design dimensions. The proposed classification for the research goals of the surveyed studies is exhaustively covered by the scientific research questions and the features of this research. Specifically: design and technical development
research (SQ3), role in domain-specific field (SQ1), evaluation research (SQ4), research trends and challenges (SQ5), social role analysis (SQ2) and a taxonomy for classification (SQ1, SQ2, SQ3, SQ4).

![Fig. 14. DLs used by secondary studies](image)

![Fig. 15. Secondary studies research purposes](image)

9 LIMITATIONS

We evaluate the limitations of this research based on the taxonomy of threats to validity as defined by Wohlin et al. [136].

**Construct and internal validity.** There are potential threats driven by the defined research strategy and the execution of the study selection process that might have led to miss or exclude relevant studies for the review. Concerning missing secondary studies with our research strategy, the data sources and additional reference strategies have been carefully designed and extended according to high quality assessment standards in terms of studies' coverage [35] and in adaptation to the software engineering field to mitigate as much as possible the risk of missing relevant studies. The search strings used for conversational agent synonyms have been updated through a preliminary literature analysis and validated by the research method findings and the set of synonyms used by the secondary studies. Concerning the threat of excluding relevant studies, inclusion criteria have been designed using broad, flexible definitions of the terms conversational agent and literature review. On the other hand, domain-specific exclusion criteria (i.e., EC1, EC2) have been designed as strong restrictions to reduce the probability of excluding studies whose conclusions could be applied at any level to our research field. Finally, as part of the threats raised by the research method, the risk of personal bias during the inclusion/exclusion criteria evaluation, the quality assessment or even during the feature extraction is also a possible threat. The first two processes have been reviewed and discussed among the authors to reduce personal bias by reaching a common agreement through discussion on the threats and limitations of the decisions taken at each evaluation. Concerning the feature extraction process, dedicated collaborative sessions with all the co-authors were carried out for each feature in alignment with each research questions to discuss the quality, reliability and potential threats of the results.

**External validity.** Missing relevant studies also apply to external validity for those papers whose full-text was not available even under explicit request to the authors (EC5), for which we did not receive any response. For those studies that were finally included in the study selection data-set, there is also a potential threat in terms of incomplete or inaccurate research information and conclusions reported by those secondary studies, especially for those with a low quality assessment score. Given that a conservative strategy was adopted to mitigate the construct validity threat of missing relevant studies, it was necessary to carefully inspect and trace all secondary study results which have been reported and integrated into our feature extraction and discussion results.
Conversational Agents in Software Engineering

Consequently, all features, examples, categories, methodologies and techniques have been linked to specific references to mitigate the risk of replicating inaccuracy on the results.

**Conclusion validity.** Personal bias in inclusion/exclusion evaluation, quality assessment and feature extraction processes might also lead to a validity threat in terms of the discussion and conclusions reported in this research, for which the same mitigation strategies can be applied. In terms of replicability of the study and its conclusions, the main threat is the restricted time span of the study search, which we have reduced by removing all time restrictions from the research strategy except for the date in which the search was executed. Despite this, assuming the inevitability of the time dimension consequences, the replication package attached to this study aims to provide an exhaustive, detailed summary of the entire research method execution to provide a clear, transparent overview of the research process and how it was conducted.

### 10 CONCLUSIONS

The integration of conversational agents in traditional software-based systems as intuitive, easy-to-use interfaces between humans and machines is expected to be a breakthrough in the future of HCI. As an emerging research field, academic research in the last years has focused on evaluating the specifications and the impact of conversational agents in a wide variety of contexts of use and domains. But complementarily to scientific research, commercial solutions are already exploring and exploiting the benefits of these software-based dialogue systems through successful solutions like smart-home assistants, autonomous call-centres or customer-service chatbots.

Research literature demonstrates the interdisciplinarity of conversational agents not only in terms of application domains, but also in terms of dedicated scientific areas contributing to the state of the art of the field. Published research in scientific areas like healthcare and psychology and in social sciences like education have significantly contributed to extend the applicable knowledge to the HCI field. And while laying the basis for an effective, successful design of a smart conversational process in terms of user satisfaction and adherence, they also contribute to guide future research for the next challenges that will require full attention in terms of user engagement, individual personalization or context-awareness strategies.

To this end, the latest contributions in the field benefit from the potential of deep learning solutions. These techniques allow effective integration of immediate, conversation-related knowledge using past interactions with the user through the use of recurrent neural networks. But adapting and personalizing the communicational process based on past interactions is just the beginning of full context-aware, personalized conversational agents. A holistic design of the data sources and adaptation mechanisms in a software system that can be integrated into a conversational agent is essential for designing customized solutions exploiting as much as possible the available knowledge for a better user experience. Consequently, effective design of data-driven solutions for the exploitation of contextual data in the conversational process is a key factor for future research.

All in all, this survey aims to provide a holistic, clear overview of the state of the art of the conversational agents’ research field through a set of distributed taxonomies for the research disciplines covered by each SQ. These contributions are intended not only to demonstrate and summarize the previously reported discussion and conclusions, but also to build a clear exposition which might help to synthesize existing knowledge in the field and to those for which conversational agents are a scientific topic outside their research areas. Finally, we expect that these contributions also help to guide future research towards a new era of HCI through the successful integration of conversational agents.
ACKNOWLEDGMENTS

With the support from the Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia and the European Social Fund. The corresponding author gratefully acknowledges the Universitat Politècnica de Catalunya and Banco Santander for the financial support of his predoctoral grant FPI-UPC.

REFERENCES

[1] A. Abashev, R. Grigoryev, K. Grigorian, and V. Boyko. 2017. Programming Tools for Messenger-Based Chatbot System Organization: Implication for Outpatient and Translational Medicines. BioNanoScience 7 (2017).

[2] S A Abdul-Kader and J C Woods. 2015. Survey on Chatbot Design Techniques in Speech Conversation Systems. International Journal of Advanced Computer Science and Applications 6, 7 (2015).

[3] Alain Abran, Adel Khelifi, Witold Suryn, and Ahmed Seffah. 2003. Usability Meanings and Interpretations in ISO Standards. Software Quality Journal 11 (11 2003).

[4] Eleni Adamopoulou and Lefteris Moussiaides. 2020. Chatbots: History, technology, and applications. Machine Learning with Applications 2 (Dec. 2020).

[5] Nahdatul Akma Ahmad. 2018. Review of Chatbots Design Techniques. International Journal of Computer Applications 181 (08 2018).

[6] N S Ahmad, M H Sanusi, M H Abd Wahab, A Mustapha, Z A Sayadi, and M Z Saringat. 2018. Conversational Bot for Pharmacy: A Natural Language Approach. In 2018 IEEE Conference on Open Systems (ICOS).

[7] Nahdatul Akma, Mohamad Hafiz, Azaliza Zainal, Muhammad Fairuz Abd Rauf, and Zuraidy Adnan. 2018. Review of Chatbots Design Techniques. International Journal of Computer Applications 181 (08 2018).

[8] Anna A Allen, Howard C Shane, and Ralf W Schlosser. 2018. The Echo™ as a Speaker-Independent Speech Recognition Device to Support Children with Autism: an Exploratory Study. Advances in Neurodevelopmental Disorders 2, 1 (2018).

[9] DBpedia Association. [n.d.]. DBpedia. https://www.dbpedia.org/ Accessed: 2021-05-25.

[10] Rodrigo Bavaresco, Diórgenes Silveira, Eduardo Reis, Jorge Barbosa, Rodrigo Righi, Cristiano Costa, Rodolfo Antunes, Marcio Gomes, Clauter Gatti, Mariangela Vanzin, Saint Clair Junior, Elton Silva, and Carlos Moreira. 2020. Conversational agents in business: A systematic literature review and future research directions. Computer Science Review 36 (May 2020).

[11] R. V. Belfín, A. J. Shobana, Megha Manilal, Ashly Ann Mathew, and Blessy Babu. 2019. A Graph Based Chatbot for Cancer Patients. In 2019 5th International Conference on Advanced Computing Communication Systems (ICACCS).

[12] Nivedita Bhirud, Subhash Tataale, Sayali Randive, and Shubham Nahar. 2019. A Literature Review On Chatbots In Healthcare Domain. International Journal of Scientific & Technology Research 8, 7 (2019).

[13] Jean-Emmanuel Bibault, Benjamin Chaix, Arthur Guillemaës, Sophie Cousin, Alexandre Escande, Morgane Perrin, Arthur Pienkowski, Guillaume Delamon, Pierre Nectoux, and Benoît Brouard. 2019. A Chatbot Versus Physicians to Provide Information for Patients With Breast Cancer: Blind, Randomized Controlled Noninferiority Trial. J. Med. Internet Res. 21, 11 (Nov. 2019).

[14] Timothy Bickmore, Dina Utami, Rohin Matsuyama, and Michael Paasche-Orlowsky. 2016. Improving Access to Online Health Information With Conversational Agents: A Randomized Controlled Experiment. Journal of Medical Internet Research 18 (01 2016).

[15] Eva Bittner, Sarah Oeste-Reiß, and Jan Marco Leimeister. 2019. Where is the bot in our team? Toward a taxonomy of design option combinations for conversational agents in collaborative work. In Proceedings of the 52nd Hawaii international conference on system sciences.

[16] Markus Blut, Cheng Wang, Nancy V Wünderlich, and Christian Brock. 2021. Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI. Journal of the Academy of Marketing Science (Jan. 2021).

[17] Héctor Cadavid, Vasilios Andrikopoulos, and Paris Avgeriou. 2020. Architecting systems of systems: A tertiary study. Information and Software Technology 118 (Feb. 2020).

[18] Rollo Carpenter. [n.d.]. Cleverbot. https://www.cleverbot.com/ Accessed: 2021-05-25.

[19] Jacky Casas, Marc-Olivier Tricot, Omar Abou Khaled, Elena Mugellini, and Philippe Cudré-Mauroux. 2020. Trends & Methods in Chatbot Evaluation. In Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI ’20 Companion). ACM.
Conversational Agents in Software Engineering

[20] Naganna Chetty, Kunwar Singh Vaisla, and Nagamma Patil. 2015. An Improved Method for Disease Prediction Using Fuzzy Approach. In 2015 Second International Conference on Advances in Computing and Communication Engineering.

[21] Kenneth Mark Colby. [n.d.]. PARRY. https://www.chatbots.org/parry/ Accessed: 2021-05-31.

[22] Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. Superagent: A customer service chatbot for e-commerce websites. In Proceedings of ACL 2017, System Demonstrations.

[23] Robert Dale. 2016. The return of the chatbots. Natural Language Engineering 22 (09 2016).

[24] Allan de Barcelos Silva, Marcio Miguel Gomes, Cristiano André da Costa, Rodrigo da Rosa Righi, Jorge Luis Victoria Barbosa, Gustavo Pessin, Geert De Doncker, and Gustavo Federizzi. 2020. Intelligent personal assistants: A systematic literature review. Expert Syst. Appl. 147 (June 2020).

[25] Douglas Derrick, Jeffery Jenkins, and J. Jr. 2011. Design Principles for Special Purpose, Embodied, Conversational Intelligence with Environmental Sensors (SPECIES) Agents. AIS Transactions on Human-Computer Interaction 3 (06 2011).

[26] Luis Fernando D’Haro, Seokhwan Kim, and Rafael E. Banchs. 2015. A robust spoken Q&A system with scarce in-domain resources. In 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA).

[27] R Dsouza, S Sahu, R Patil, and D R Kalbande. 2019. Chat with Bots Intelligently: A Critical Review Analysis. In 2019 International Conference on Advances in Computing, Communication and Control (ICAC3).

[28] Nikolai Durov and Pável Dúrov. [n.d.]. Telegram. https://telegram.org/ Accessed: 2021-05-25.

[29] Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring Language Style in Chatbots to Increase Perceived Product Value and User Engagement.

[30] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A Taxonomy of Social Cues for Conversational Agents. Int. J. Hum. Comput. Stud. 132 (Dec. 2019).

[31] Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. An intelligent discussion-bot for answering student queries in threaded discussions. International Conference on Intelligent User Interfaces, Proceedings IUI 2006.

[32] Nirosshine Fernando, Felix Ter Chian Tan, Rajesh Vasa, Kon Mouzaki, and Ian Aitken. 2016. Examining Digital Assisted Living: Towards a Case Study of Smart Homes for the Elderly. (2016).

[33] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Amal Abdulrahman, and Willem-Paul Brinkman. 2019. What are We Measuring Anyway?

[34] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Ment Health 4, 2 (06 Jun 2017).

[35] Centre for Reviews and Dissemination Systematic Reviews. 2009. Database of Abstracts of Reviews of Effects. https://www.crd.york.ac.uk/CRDWeb/AboutPage.asp Accessed 22/04/2021.

[36] A.L.I.C.E. Artificial Intelligence Foundation. [n.d.]. AIML: Artificial Intelligence Markup Language. https://web.archive.org/web/20070927192442/http://www.alicebot.org/aiml.html Accessed: 2021-05-25.

[37] Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, and Michiel Rauws. 2018. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. JMIR Ment Health 5, 4 (13 Dec 2018).

[38] Asbjørn Følstad and Petter Brandtzaeg. 2017. Chatbots and the new world of HCI. interactions 24 (06 2017).

[39] Elia Gabarron, Dillys Larbi, Kerstin Denecke, and Eirik Årsand. 2020. What Do We Know About the Use of Chatbots for Public Health? Stud. Health Technol. Inform. 270 (June 2020).

[40] Tobias Gentner, Timon Neitzel, Jacob Schulze, and Ricardo Buettner. 2020. A Systematic Literature Review of Medical Chatbot Research from a Behavior Change Perspective.

[41] Shameek Ghosh, Sammi Bhatia, and Abhi Bhatia. 2018. Quro: Facilitating User Symptom Check Using a Personalised Chatbot-Oriented Dialogue System. Studies in health technology and informatics 252 (01 2018).

[42] Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2017. Towards Designing Cooperative and Social Conversational Agents for Customer Service. In ICIS.

[43] David Griol, Araceli Sanchis, José Manuel Molina, and Zoraida Callejas. 2019. Developing enhanced conversational agents for social virtual worlds. Neurocomputing 354 (2019).

[44] Stanford NLP Group. [n.d.]. Stanford CoreNLP. https://stanfordnlp.github.io/CoreNLP/ Accessed: 2021-05-25.

[45] S. Gupta, Deep Borkar, C. D. Mello, and S. Patil. 2015. An E-Commerce Website based Chatbot.

[46] Eko Handoyo, M. Arfan, Yosua Alvin Adi Soetrinoso, Maman Somantri, Aghus Sofwan, and Enda Wista Sinuraya. 2018. Ticketing Chatbot Service using Serverless NLP Technology. In 2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICTACEE).

[47] Roy Harper, Peter Nicholl, Jonathan Wallace, Lesley-Ann Black, and Patricia Kearney. 2008. Automated Phone Capture of Diabetes Patients Readings with Consultant Monitoring via the Web. In 15th Annual IEEE International Conference and Workshop on the Engineering of Computer Based Systems (echb 2008).
A I Niculescu, K H Yeo, L F D’Haro, S Kim, R Jiang, and R E Banchs. 2014. Design and evaluation of a conversational
agent for the touristic domain. In Signal and Information Processing Association Annual Summit and Conference
(APSIAP), 2014 Asia-Pacific.

Existor Ltd. [n.d.]. Cleverscript. https://www.cleverscript.com/ Accessed: 2021-05-25.

Rosales, and Elizabeth Valencia. 2019. A Survey on Evaluation Methods for Chatbots. In Proceedings of the 2019 7th
International Conference on Information and Education Technology (ICIET’19). ACM.

Rosales, and Elizabeth Valencia. 2019. A Survey on Evaluation Methods for Chatbots. In Proceedings of the 2019 7th
International Conference on Information and Education Technology (ICIET’19). ACM.

A I Niculescu, K H Yeo, L F D’Haro, S Kim, R Jiang, and R E Banchs. 2014. Design and evaluation of a conversational
agent for the touristic domain. In Signal and Information Processing Association Annual Summit and Conference
(APSIAP), 2014 Asia-Pacific.

D Niculescu and Badri Nath. 2003. Ad hoc positioning system (APS) using AOA. In IEEE INFOCOM 2003. Twenty-second
Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428), Vol. 3.

Panitan Muangkammuen, Narong Intiruk, and Kanda Runapongsa Saikaew. 2018. Automated Thai-FAQ Chatbot
using RNN-LSTM. In 2018 22nd International Computer Science and Engineering Conference (ICSEC).

Takuma Okuda and Sanae Shoda. 2018. AI-based chatbot service for financial industry. Fujitsu Sci. Tech. J. (2018).

Kulothunkan Palasundram, Nurfadhilla Mohd Sharef, Khairul Azhar Kasmiran, and Azreen Azman. 2020. Enhancements
to the Sequence-to-Sequence-Based Natural Answer Generation Models. IEEE Access 8 (2020).

Dr Lal PathLabs. [n.d.]. Pathology Lab Chatbot. https://www.chatbots.org/chatbot/lpl/ Accessed: 2021-05-25.
[123] Hiroki Tanaka, Hideki Negoro, Hitomi Iwasaka, and Satoshi Nakamura. 2017. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. In 2016 IEEE 8th International Conference on Intelligent Systems (IS).

[124] J Todorov, S Stoyanov, V Valkanov, B Daskalov, and I Popchev. 2016. Learning Intelligent System for Student Assistance - LISSA. In 2016 IEEE 8th International Conference on Intelligent Systems (IS).

[125] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding Chatbot-mediated Task Management.

[126] Van Pinxteren Michelle M, Mark Pluymaekers, and Lemmink Jos G A. 2020. Human-like communication in conversational agents: a literature review and research agenda. Journal of Service Management 31, 2 (Jan. 2020).

[127] Anusha Vegesna, Pranjul Jain, and Dhruv Porwal. 2018. Ontology based chatbot (for e-commerce website). Int. J. Comput. Appl. Technol. 179, 14 (2018).

[128] Viswanath Venkatesh, John A. Aloysius, Hartmut Hoehle, and Scot Burton. 2017. Design and Evaluation of Auto-ID Enabled Shopping Assistance Artifacts in Customers’ Mobile Phones: Two Retail Store Laboratory Experiments. MIS Q. 41, 1 (March 2017).

[129] Akshay Vijayan, S Janmasree, C Keerthana, and L Baby Syla. 2018. A Framework for Intelligent Learning Assistant Platform Based on Cognitive Computing for Children with Autism Spectrum Disorder. In 2018 International CET Conference on Control, Communication, and Computing (IC4).

[130] Richard S. Wallace. [n.d.]. A.L.I.C.E. bot. https://web.archive.org/web/20070926031437/http://www.alicebot.org/ Accessed: 2021-05-31.

[131] Haolin Wang, Qingpeng Zhang, Mary Ip, and Joseph Tak Fai Lau. 2018. Social Media–Based Conversational Agents for Health Management and Interventions. Computer 51, 8 (2018).

[132] Xiaojie Wang and Caixia Yuan. 2016. Recent Advances on Human-Computer Dialogue. CAAI Transactions on Intelligence Technology 1, 4 (2016).

[133] Joseph Weizenbaum. 1983. ELIZA — a Computer Program for the Study of Natural Language Communication between Man and Machine. Commun. ACM 26, 1 (Jan. 1983).

[134] Bruce Wilcox. [n.d.]. ChatScript. https://github.com/ChatScript/ChatScript Accessed: 2021-05-25.

[135] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE ’14, Article 38). ACM.

[136] Claes Wohlin, Per Runeson, Höst Martin, Magnus C. Ohlsson, Regnell Björn, and Wesslén Anders. 2012. Experimentation in software engineering: an introduction. Springer-Verlag New York.

[137] Wilson Wong, John Thangarajah, and Lin Padgham. 2011. Health Conversational System Based on Contextual Matching of Community-Driven Question-Answer Pairs. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM ’11). ACM.

[138] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI ’17). ACM.

[139] Q Xu, X Chen, S Li, H Zhang, M A Babar, and N K Tran. 2020. Blockchain-based Solutions for IoT: A Tertiary Study. In 2020 IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C).

[140] Zijun Xue, Ting-Yu Ko, Neo Yuchen, Ming-Kuang Daniel Wu, and Chu-Cheng Hsieh. 2018. Isa: Intui Smart Agent, A Neural-Based Agent-Assist Chatbot. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW).

[141] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, JianShe Zhou, and Zhoujun Li. 2017. Building Task-Oriented Dialogue Systems for Online Shopping. AAAI 31, 1 (Feb. 2017).

[142] Ran Zhao, Oscar J Romero, and Alex Rudnicky. 2018. SOGO: A Social Intelligent Negotiation Dialogue System. In Proceedings of the 18th International Conference on Intelligent Virtual Agents (IVA ’18). ACM.

[143] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. arXiv:1704.01074