Quality Assessment Methods for Textual Conversational Interfaces: A Multivocal Literature Review

information

Review

Quality Assessment Methods for Textual Conversational Interfaces: A Multivocal Literature Review

Riccardo Coppola and Luca Ardito *

Dipartimento di Automatica ed Informatica (DAUIN), Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy; riccardo.coppola@polito.it
* Correspondence: luca.ardito@polito.it

Abstract: The evaluation and assessment of conversational interfaces is a complex task since such software products are challenging to validate through traditional testing approaches. We conducted a systematic Multivocal Literature Review (MLR), on five different literature sources, to provide a view on quality attributes, evaluation frameworks, and evaluation datasets proposed to provide aid to the researchers and practitioners of the field. We came up with a final pool of 118 contributions, including grey (35) and white literature (83). We categorized 123 different quality attributes and metrics under ten different categories and four macro-categories: Relational, Conversational, User-Centered and Quantitative attributes. While Relational and Conversational attributes are most commonly explored by the scientific literature, we testified a predominance of User-Centered Attributes in industrial literature. We also identified five different academic frameworks/tools to automatically compute sets of metrics, and 28 datasets (subdivided into seven different categories based on the type of data contained) that can produce conversations for the evaluation of conversational interfaces. Our analysis of literature highlights that a high number of qualitative and quantitative attributes are available in the literature to evaluate the performance of conversational interfaces. Our categorization can serve as a valid entry point for researchers and practitioners to select the proper functional and non-functional aspects to be evaluated for their products.

Keywords: conversational interfaces; software quality attributes; software quality

Citation: Coppola, R.; Ardito, L. Quality Assessment Methods for Textual Conversational Interfaces: A Multivocal Literature Review. *Information* **2021**, *12*, 437. https://doi.org/10.3390/info12110437

Academic Editor: Ralf Krestel

Received: 28 September 2021
Accepted: 19 October 2021
Published: 21 October 2021

1. Introduction

As defined by Radziwill et al., conversational interfaces are one class of intelligent, conversational software agent activated by natural language input (which can be in the form of text, voice, or both). Conversational interfaces provide conversational output in response, and if commanded, they can sometimes also execute tasks [1]. Conversational interfaces are commonly referred to as chatbots when the interface is only textual.

The research into the evaluation of chatbots dates back to the early 1970s when a team of psychiatrists subjected the two earliest ones (Eliza [2], and Parry [3]) to the Turing tests. Despite the initial interest from the scientific community, the chatbot topic in the broad sense has been explored in more depth by academia only in the last two decades, mainly because of the previous lack of sufficiently sophisticated hardware and theoretical models [4,5]. In the latest years, chatbots have also gained an important commercial interest [6], which has also resulted in constant technological advancement.

Global forecasts have shown that the chatbot market is projected to grow from USD 2.6 billion in 2019 to USD 9.4 billion by 2024 at a compound annual growth rate (C.A.G.R.) of 29.7% [7], with healthcare, educational, customer services and marketing as the most affected application domains. The main driver of such commercial interest is the ability of chatbots to provide rapid responses and well enough support to customer requests [8]. The main shortcomings are instead found in the possibility of unhelpful responses and the lack of a human personality. These issues are still slowing down a widespread acceptance of chatbots [9,10]. Due to the increasing economic impact and the mentioned limitations,

the need to set comprehensive and replicable approaches to test and evaluate chatbots thoroughly has been brought to light [11].

The software behind chatbots is however challenging to verify and validate with traditional testing approaches. Their evaluation is in fact strictly related to their ability to replicate human behavior, and to the user's appreciation of their output [12,13]. The non-determinism of user input also makes the coverage of all possible inputs impractical. The semantic component of the responses to the users must also be taken into account when verifying conversational interfaces. Therefore, academic research has yet failed to converge towards an established set of metrics and actionable approaches to validate conversational interfaces.

The purpose of this paper is to present a comprehensive review of quality properties and attributes related to the quantitative and qualitative verification and validation of chatbots.

To this end, we performed a Multivocal Literature Review (MLR) study that covers not only peer-reviewed works (i.e., white literature or WL) but also grey literature accessible through traditional search engines. By involving the latter, we aimed at capturing valuable information produced by practitioners from the industry, and to compare the practitioners' focus to that of academia. By analyzing the most widespread attributes and properties analyzed in both categories of literature, we discuss potential gaps in current research and practice and the related implications for industry and academia.

The remainder of this paper is organized as follows:

- Section 2 presents the background about the evaluation processes of chatbots and compares this work to existing secondary studies in the field;
- Section 3 describes the adopted research methods by specifying its goals, research questions, search criteria, and analysis methods;
- Section 4 presents the results of the MLR;
- Section 5 discusses the implications of the results and the threats to the validity of our study;
- Finally, Section 6 concludes the study and presents possible future avenues for this work.

2. Background

In this section, we summarize the concepts about the chatbot evaluation processes defined in the academic literature. We also present background concepts regarding applying the MLR research method in the field of Software Engineering, and we discuss the findings of existing secondary studies in this field.

2.1. Overview of Quality Assessment Methods for Conversational Interfaces

The literature on chatbots has highlighted a lack of precise guidelines for designing and evaluating the quality of this type of software. Amershi et al. propose a set of guidelines tailored to the peculiar human–AI nature of the interaction with chatbots [14].

The *quality* attribute for a chatbot can relate to many different aspects of its usage, e.g., the capability of providing the right answers or to infer the right emotions from the human users, or the end user's satisfaction [1]. However, the quality properties to evaluate depend on the purpose and application domain of the specific chatbot to be evaluated, making it difficult to find universal attributes.

Concerning the latest generation of chatbots, based on deep learning and structured information [15], most of the research in the field has emphasized the use of annotated datasets, defined as ground truth. The generation of annotated datasets may imply the presence of human labellers (i.e., a supervised approach) or can be performed automatically based on the characteristics of data (i.e., an unsupervised) approach. Once a ground truth is obtained, the chatbot model infers its behavior learning from this dataset, and the performance is evaluated over a subset of the dataset, the so-called test set.

Several studies in the literature address the application of automated testing methodologies to verify the quality of chatbot software. The many aspects to be considered in chatbot evaluation however make fully automated testing practices harder to adopt than in traditional software domains.

The hardest features to verify are those related to the perception of the chatbot by a human user, and the perceived value of obtained information. For these reasons, manual testing of chatbots is rarely used alone practitioners, and is typically conducted with the aid of questionnaires and interviews [16–18]. Manual testing is, however, inherently error prone and labor intensive, hence it is typically paired with evaluations by domain experts, and aided with platforms for crowdsourced testing [19,20]. Quantitative measurements of the performance of the chatbots (e.g., inspection of abandoned dialogues) are instead most of the time completely automated.

2.2. Multivocal Literature Reviews

Ogawa et al. define the concept of the Multivocal Literature Review (MLR) [21] in the field of Education as a research methodology that applies the approach of a Systematic Literature Review on multiple literature sources, i.e., involving evidence available on regular search engines. In that sense, MLRs differ from regular SLRs (Systematic Literature Reviews) because they include information obtained from non-academic sources, such as blog posts, web-pages, and industry white papers. According to a definition provided by Lefebvre et al., these sources can all be classified as *Grey Literature* (GL), i.e., *literature that is not formally published in sources such as books or journal articles* [22]. Several classifications of the forms of GL have been provided in the literature. Adams et al. provide a three-tier classification of GL: 1st tier GL (high credibility), including books, magazines, government reports, and white papers; 2nd tier GL (moderate credibility), such as annual reports, news articles, presentations, videos, Question and Answer websites (such as StackOverflow) and Wiki articles; 3rd tier GL (low credibility), such as blogs, e-mails, tweets [23].

Albeit that many review studies in the field of Software Engineering (SE) have implicitly incorporated GL to derive their findings, a formalization of the MLR methodology for SE has been provided only recently by Garousi et al. [24]. The authors identified the principal benefit of including GL in literature reviews as the capability of providing useful industry viewpoints and evidence of the quality that cannot always be gathered from peer-reviewed literature [25]. Rigorous MLRs have recently been conducted in the field of SE to investigate, for instance, the need for automation for software testing [26], software test maturity assessment and test process improvement [27], security in DevOps [28], the benefits of the adoption of the Scaled Agile framework [29], requirements engineering in software startups [30], technical debt [31], and the financial aspects of its management [32].

2.3. Related Work

Many studies have already had as their primary purpose an examination of quality attributes for chatbots. Several systematic reviews of chatbot quality assessments have been performed and are available in the literature. In Table 1, we report the secondary studies available at the time this review was conducted. For each of the secondary studies considered, we report the research methodology employed, the number of primary studies referenced, and the main contribution.

The Grey Literature work by Radzwill and Benton [1] analyze the data from 46 primary studies, both from academic and grey literature. In the manuscript, the authors define three different categories of quality attributes for conversational agents: effectiveness, efficiency, and satisfaction. These three categories have been chosen following the definition of Software Usability, paying close attention not only to functionality but also to human-like aspects. In this work, the quality assessment approaches are reviewed, and, ultimately, a synthesized approach is proposed. To date, this study is the one that proposes the most comprehensive list of quality attributes for chatbots. However, the study has some limitations especially in terms of replicability, since there is no explicit adherence to a

review protocol (e.g., Kitchenham's). Moreover, no explicit Inclusion and or Exclusion criteria are provided for the selection of manuscripts. In addition to this lack of formality of the review process, the manuscript is also focused only on quality attributes and not on frameworks or datasets that can be utilized for chatbot evaluations. Finally, the manuscript also lacks a mapping section to provide a categorization of the available academic research in the field.

Maroengsit et al. [12] performed a survey to assess the various architectures and application areas of chatbots, providing a categorization and analysis based on 30 different conversational interfaces. The authors provide a review of natural language processing techniques and evaluation methods for chatbots. The evaluation methods part is divided into content evaluation, user satisfaction, and functional evaluation. This work has a primary focus on natural language processing and low-level white box metrics. However, the work provides a limited analysis of black-box metrics focused on the user perspective. It only considers a high-level subdivision for them (e.g., automatic evaluation and expert evaluation), which we deem not sufficient to cover the complexity of all quality attributes and metrics provided by the available research.

Finally, Ren et al. [33] provided a systematic mapping study about the usability attributes for chatbots. Several specific quality assessment metrics and methods are discussed in this work. However, an exhaustive discussion of quality attributes or frameworks is not provided. The authors also provided a classification of the conversational interfaces in different categories: AIML (Artificial Intelligent Markup Language), NLP (Natural Language Processing), ORM (Object Relational Mapping), and ECA (Embodied Conversational Agents), but no explicit classification of the evaluation techniques used for each of the categories is provided in the manuscript. In addition to the limitation of the analysis to the evaluation of usability, the study does not include an analysis of grey literature sources.

The present work aims to review a broader range of recent sources and to integrate the contribution of works from grey literature and practitioners' reports, which are generally included only to a limited extent in the mentioned works. We also aim at providing an analysis of existing frameworks and datasets used for chatbot evaluations, and a mapping of the literature about the mentioned research facets, which has not been provided yet by related studies.

Table 1. Secondary studies.

Reference	Year	Title	Research Methodology	# of Primary Studies	Description
[1]	2017	Evaluating Quality of Chatbots and Intelligent Conversational Agents	MLR	46	This paper presents a literature review of quality issues and attributes as they relate to the contemporary issue of chatbot development and implementation. The quality assessment approaches are reviewed and based on these attributes a quality assessment method is proposed and examined.
[12]	2019	A Survey on Evaluation Methods for Chatbots	SLR	30	This work presents a survey starting from a literature review, evaluation methods/criteria and comparison of evaluation methods. It is conducted with classication of chatbot evaluation methods and their analysis according to chatbot types and the three main evaluation schemes: content evaluation, user satisfaction and chat function.
[33]	2019	Usability of Chatbots: A Systematic Mapping Study	SLR	19	This paper is focused on identifying the state of the art in chatbot usability and applied human–computer interaction techniques and to analyze how to evaluate chatbots usability. The works were categorized according to four criteria: usability techniques, usability characteristics, research methods and type of chatbots.

3. Research Method

This section provides an overview of the research method that we adopted when conducting this study.

We conducted an MLR by following the guidelines provided by Garousi et al. [24]. These guidelines are built upon Kitchenham's guidelines for conducting SLRs [34], with the addition of specific phases that tackle the procedure of selection and filtering of the grey literature.

According to these guidelines, the procedure for conducting an MLR is composed of three main phases:

1. Planning: in this phase, the need for conducting an MLR on a given topic is established, and the goals and research questions of the MLR are specified;
2. Conducting: in this phase, the MLR is conducted entailing five different sub-steps: definition of the search process, source selection, assessment of the quality of the selected studies, data extraction, and data synthesis;
3. Reporting: in this phase, the review results are reported and tailored to the selected destination audience (e.g., researchers or practitioners from the industry).

In the following subsections, we report all the decisions taken during the Planning and Conducting phases of our study. The Results and Discussion sections of the paper will serve as the output of the *Reporting* phase.

3.1. Planning

This section describes the components of the planning phase according to the guidelines by Garousi et al.: motivation, goals, and RQs. This information is reported in the following.

3.1.1. Motivation Behind Conducting an MLR

To motivate the inclusion of Grey Literature in our literature review, and thus to conduct an MLR, we adopted the approach based on the decision table reported in Table 2, defined by Garousi et al. [24] and based on the guidelines by Benzies et al. [23,35]. One or more positive responses to the questions in the table suggest the inclusion of GL in the review process.

As is evident from our decision table shown in Table 2, we could provide a positive answer to all questions about the addressed subject. Hereby, we provide a brief motivation about each point in the decision table:

1. The subject is not addressable only with evidence from formal literature, since typically real-world limitations of the conversational interfaces are addressed by white literature only to a certain extent;
2. Many studies in the literature do provide methods for the evaluation of conversational interfaces with small controlled experiments, which may not reflect the dynamics of the usage of such technologies by practitioners;
3. The context where they are applied is of crucial importance for conversational interfaces, and grey literature is supposed to provide more information of this kind since they it is more strictly tied to actual practice;
4. Practical experiences reported in grey literature can indicate whether the metrics or approaches proposed in the formal literature are feasible or beneficial in real-world scenarios;
5. Grey literature can reveal the existence of more evaluation methodologies and metrics than those that could be deduced from white literature only;
6. Observing the outcomes of measurements on commercial products can provide researchers with relevant insights regarding where to focus research efforts; conversely, practitioners can deduce new areas in which to invest from the white literature;
7. Conversational interfaces and their evaluation are prevalent in the software engineering area, which accounts for many sources of reliable grey literature.

Table 2. Questions asked to decide whether to include the GL in software engineering reviews.

#	Question	Answer
1	Is the subject "complex" and not solvable by considering only the formal literature?	Yes
2	Is there a lack of volume or quality of evidence, or a lack of consensus of outcome measurement in the formal literature?	Yes
3	Is the contextual information important in the subject under study?	Yes
4	Is it the goal to validate or corroborate scientific outcomes with practical experiences?	Yes
5	Is it the goal to challenge assumptions or falsify results from practice using academic research or vice versa?	Yes
6	Would a synthesis of insights and evidence from the industrial and academic community be useful to one or even both communities?	Yes
7	Is there a large volume of practitioner sources indicating high practitioner interest in a topic?	Yes

3.1.2. Goals

This MLR aims to identify the best practices concerning specific procedures, technologies, methods, or tools by aggregating information from the literature. Specifically, the research is based on the following goals:

- Goal 1: Providing a mapping of the studies regarding the evaluation of the quality of conversational interfaces.
- Goal 2: Describing the methods, frameworks, and datasets that have been developed in the last few years for the evaluation of the quality of conversational interfaces.
- Goal 3: Quantifying the contribution of grey and practitioners' literature to the subject.

3.1.3. Review Questions

Based on the goals defined above, we formulate three sets of research questions.

Regarding the first goal, we identify two mapping research questions that can be considered common to all MLR studies:

- **RQ1.1**—What are the different categories of contributions of the considered sources?
- **RQ1.2**—How many sources present metrics, guidelines, frameworks, and datasets to evaluate textual conversational interfaces?
- **RQ1.3**—Which research methods have been used in the considered sources?

Regarding the second goal, we identify three domain-specific RQs, defined in the following:

- **RQ2.1**—Which are the metrics used for the quality evaluation of textual conversational interfaces?
- **RQ2.2**—Which are the proposed frameworks for evaluating the quality of textual conversational interfaces?
- **RQ2.3**—Which are the datasets used to evaluate textual conversational interfaces?

Regarding the final goal, we defined two research questions to assess attention towards the topic by the research vs. practitioner communities:

- **RQ3.1**—Which technological contributions from the industry and practitioner community are leveraged by white literature?
- **RQ3.2**—How much attention has this topic received from the practitioner and industrial community compared to the research community?

3.2. Conducting the MLR

According to the MLR conduction guidelines formulated by Garousi et al., in this section, we report the source selection, search strings, the paper selection process, and how the data of interest were extracted from the selected sources.

The process that we conducted for this study is described in the following sections and outlined in Figure 1. In the diagram, we report the number of sources in our pool after executing each step of the review.

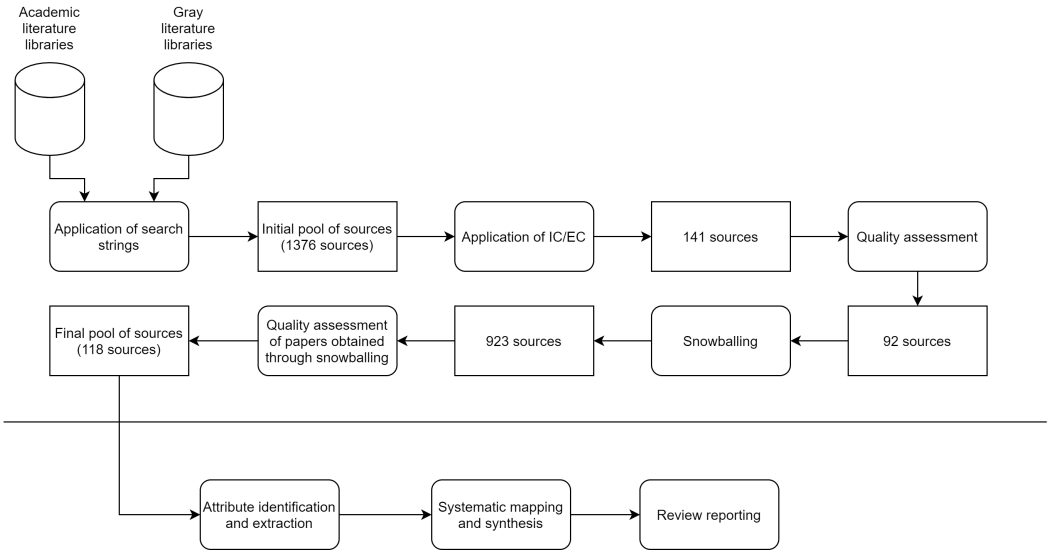

Figure 1. Phases of the literature review.

3.2.1. Search Approach

To conduct the review, we followed the following steps:

- *Application of the search strings*: the specific strings were applied to the selected online libraries (for the white literature search) and on the Google search engine (for the grey literature search);
- *Search bounding*: To stop the searching of grey literature and to limit the number of sources to a reasonable number, we applied the *Effort Bounded* strategy, i.e., we limited our effort to the first 100 Google search hits;
- *Removal of duplicates*: in our pool of sources, we consider a single instance for each source that is present in multiple repositories;
- *Application of inclusion and exclusion criteria*: we defined and applied the inclusion and exclusion criteria directly to the sources extracted from the online repositories, based on an examination of titles, keywords, and abstracts of the papers;
- *Quality assessment*: every source from the pool was entirely read and evaluated in terms of the quality of the contribution.
- *Backward Snowballing* [36]: all the articles in the reference lists of all sources were added to the preliminary pool and evaluated through the application of the previous steps. We also added to the pool of grey literature the grey literature sources cited by white literature;
- *Documentation and analysis*: the information about the final pool of paper was collected in a form including fields for all the information needed to answers the formulated research questions.

3.2.2. Selected Digital Libraries

To find white literature sources regarding our research goal, we searched the following academic online repositories:

1. ACM Digital Library: https://dl.acm.org/ (accessed on 18 October 2021);
2. IEEE Xplore: http://www.ieeexplore.ieee.org (accessed on 18 October 2021);
3. Springer Link: https://link.springer.com/ (accessed on 18 October 2021);
4. Science Direct Elsevier: https://www.sciencedirect.com/ (accessed on 18 October 2021);
5. Google Scholar: http://www.scholar.google.com (accessed on 18 October 2021);

The repository held by the Association for Computational Linguistic (ACL Anthology) was excluded from this list, given that the results showed a complete overlap with those obtained from the Google Scholar engine.

To these sources, we added Google's regular search engine to find grey literature sources related to our research goal.

3.2.3. Search Strings

A pool of terms was defined through brainstorming to determine the most appropriate terms for the search strings:

Initial pool of terms

- conversational [agent, assistant, system, interface, AI, artificial intelligence*, bot*, ...], dialog [system*, ...], chatbot [system*, ...], virtual [agent*, assistant*, ...], personal digital assistant, ...
- evaluat*, measur*, check*, metric*, quality, quality assessment, quality attribute*, criteria, analys*, performance, rank*, assess*, benchmark, diagnostic, test*, compar*, scor*, framework, dataset...
- user [interaction, experience, satisfaction, ...], customer [interaction, experience, satisfaction, ...], engagement, intent, psychometric, usability, perception, QoE, naturalness, personal*, QoS, ...

In Table 3, we report the search strings based on the pool of terms, and formulated for each digital library. The search strings include all the elicited synonyms of the terms *chatbot*, *quality assessment*, *framework* and *datasets*.

Table 3. Search string for research questions RQ2.

#	Search String
IEEE Xplore	(((chatbot* OR conversational) AND (interface* OR agent*)) AND (metric* OR evaluat* OR "quality assessment" OR analysis OR measur*))
Elsevier Science Direct	(((chatbot OR conversational) AND (interface OR agent)) AND (metric OR evaluation OR "quality assessment" OR analysis OR measurement))
ACM Digital Library	(((chatbot* OR conversational) AND (interface* OR agent*)) AND (metric* OR evaluat* OR "quality assessment" OR analysis OR measur*))
Springer Link	((chatbot* OR conversational) AND (interface* OR agent*)) AND (metric* OR evaluat* OR "quality assessment" OR analysis OR measur*)
Google Scholar	metric OR evaluation OR "quality assessment" OR analysis OR measurement "chatbot interface"
Google	metric OR evaluation OR "quality assessment" OR analysis OR measurement "chatbot interface"

In the search on digital libraries, we filtered the results for publication dates between 2010 and September 2021. For the search on Google Scholar, we used the Publish or Perish (PoP) (https://harzing.com/resources/publish-or-perish, accesssed on 18 October 2021) tool; for the other sources, we used the official utilities and APIs exposed. Since the final objective was to extract and inspect all the related sources published in the 2010–2021 time frame, the search ordering was not taken into account. We excluded patents from Google Scholar results.

We used a Python script to remove exact duplicates, by retrieving pairs of articles with more than 80% overlapping words in their titles. We developed a stand-alone script that analyzed the results provided by the PoP tool and by the APIs (in the form of .csv files), and that cycled over all manuscript titles to signal potential overlaps.

The correctness of the signalled overlaps were verified by a manual check on the resulting list. A single entry was maintained for each pair of identical articles published in more than one source. A total of 1376 unique white literature papers were gathered in this step.

Regarding grey literature, we collected 100 contributions using the Google search engine. Before performing the search, we cleaned the browser of cookies and history before performing the search to avoid influencing the search's replicability. We narrowed down the search results to web pages published before the end of September 2021, by applying the before: 30 September 2021 modifier at the end of the search string. We excluded academic

sources that resulted from searches on the regular Google Search engine. The search hits were ordered by relevance, by keeping the default Google Search behavior.

3.2.4. Inclusion/Exclusion Criteria

Inclusion Criteria (from now on, IC) and Exclusion Criteria (from now on, EC) were defined to ensure gathering only the sources relevant to our research goal.

- **IC1** The source is directly related to the topic of chatbot evaluation. We include papers that explicitly propose, discuss or improve an approach regarding evaluation of conversational interfaces.
- **IC2** The source addresses the topics covered by the research questions. This means including papers using or proposing metrics, datasets, and instruments for the evaluation of conversational interfaces.
- **IC3** The literature item is written in English;
- **IC4** The source is an item of white literature available for download and is published in a peer-reviewed journal or conference proceedings; or, the source is an item of 1st tier Grey Literature;
- **IC5** The source is related (not exclusively) to text-based conversational interfaces.

Conversely, the exclusion criteria we applied were:

- **EC1** The source does not perform any investigation nor reports any result related to chatbots, corresponding evaluation metrics, datasets used to evaluate chatbots.
- **EC2** The source is not in a language directly comprehensible by the authors.
- **EC3** The source is not peer-reviewed; or, the paper is Grey Literature of 2nd or 3rd tier.
- **EC4** The source is related exclusively to a different typology of conversational interface.

Sources that did not meet the above Inclusion Criteria, or that met any of the Exclusion Criteria, were excluded from our analysis.

The first round of IC/EC and the theoretical saturation was applied considering the title and the abstract: 115 papers passed the round. From the grey literature, other 28 documents were added to the pool, 24 from google search engine and 4 from white literature snowballing that led to artifacts of grey literature.

The order in which the source is considered influences the final pool due to exhaustion criteria.

3.2.5. Quality Assessment of Sources

Each author evaluated the quality of the sources based on some aspects advised by Garousi's guidelines to perform MLRs: authority of the source, methodology, objectivity, position, novelty, impact. Each source was hence voted on by using a Likert scale. We adopted a threshold of an average score of 2.5 to keep the sources in the final pool.

3.2.6. Data Extraction and Synthesis

Once we gathered our final pool of sources, we executed the step of data extraction and synthesis on all white and grey literature works. All the studies were inserted into an online repository that was shared among the authors to facilitate concurrent analysis of the sources.

The contributions were initially described in the Google Docs spreadsheet by comments, summary texts, and inferences drawn from the documents in a descriptive way.

We did not use pre-determined categories to categorize and map the papers and extract quality attributes from them (and so, respectively, address RQ1 and RQ2). However, hence we applied the Grounded Theory approach. We adopted the Straussian definition of Grounded Theory [37], which allows up-front definition the Research Questions instead of letting them emerge from the data analysis.

The categories responding to research questions were generated through *Open Coding* [38]. We did not consider the inferred categories as mutually exclusive for the types of contribution and research for the quality attributes to extract. For quality attributes and contribution types, we also applied the *Axial Coding* procedure [39], to remove redundancies from the categories and potentially merge the less populated ones.

The open and coding procedures were performed on each paper by all the authors of this literature review independently, and divergences were discussed to find a single (set of) categories for each manuscript and attribute.

3.2.7. Final Pool of Sources

After the application of all the stages described above, as shown in Table 4, our final pool included 118 sources. The information about all contributions is included in a publicly-available spreadsheet, hosted on Google Docs (https://docs.google.com/spreadsheets/d/18hEL36Qx7VGzansemcUIpqgTcLQIhibsx103lmaTTZU/edit?usp=sharing, accessed on 18 October 2021).

Table 4. Number of papers after quality assessment of sources defining the final pool.

Repository	Number of Sources
IEEE Xplore	18
Elsevier Science Direct	17
ACM Digital Library	4
Springer Link	14
Google Scholar	12
Google	23
Snowballing WL	18
Snowballing GL	12
Final Pool	118

In Figure 2, we report the distribution of the contributions per year, discriminating between the sources gathered with direct search and those obtained through snowballing.

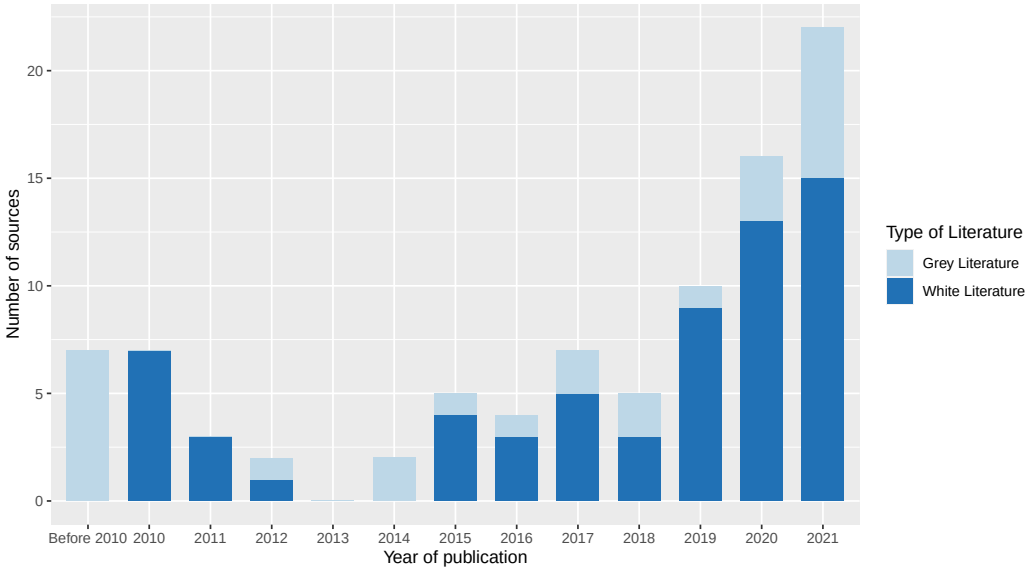

Figure 2. Amount of white and grey literature sources per year.

Since we did not apply the EC regarding the publication year on the sources obtained through snowballing, we obtained papers published before 2010. We grouped them in the plot's first column. On the other hand, we deem it not meaningful to report the publication

year of grey literature sources. By adopting the *Effort bounded* strategy (i.e., taking into account only the first 100 hits on the Google search engine), in fact, the results are naturally biased towards the most recent sources. Older grey literature (mainly blog posts or similar sources) may no longer be available due to missing systematic archiving. By analyzing the publication years of WL sources, we can see that the sources have experienced an increasing trend in the recent years. At the same time, there was little interest in the decade's central portion. This trend can be justified by taking into account the current higher availability of machine learning algorithms and repositories, which can be used to perform more agile assessments and evaluations on conversational agents.

Figure 3 reports the distribution of the source type for each of the two categories (i.e., white vs. grey literature). Among grey literature sources, we filed as *White literature pointers* the sources that could be found through searches on the regular Google search engine and linked to open-access publications in academic libraries. White literature establishes the main contributor to the final pool of sources, with 83 contributions out of 118. Despite this, the number of grey literature sources, and their variety, can be deemed a confirmation of the necessity of including such typology of literature in a comprehensive literature review. In the final pool of white literature sources there were 28 journal papers, 51 works in conference proceedings, and 4 works in companion proceedings of conferences (i.e., a workshop paper). In the final pool of grey literature sources, there were 15 blog posts, 7 documentation pages of commercial tools, 4 industry reports, 3 master's theses, one webinar source, one white paper, and 4 pointers to works in white literature. Blog posts represent the primary type of contribution in grey literature. This can be explained by the nature of blogs that can be considered the quickest means to communicate novel, high-level, practical, and actionable ideas about the design and evaluation of conversational agents.

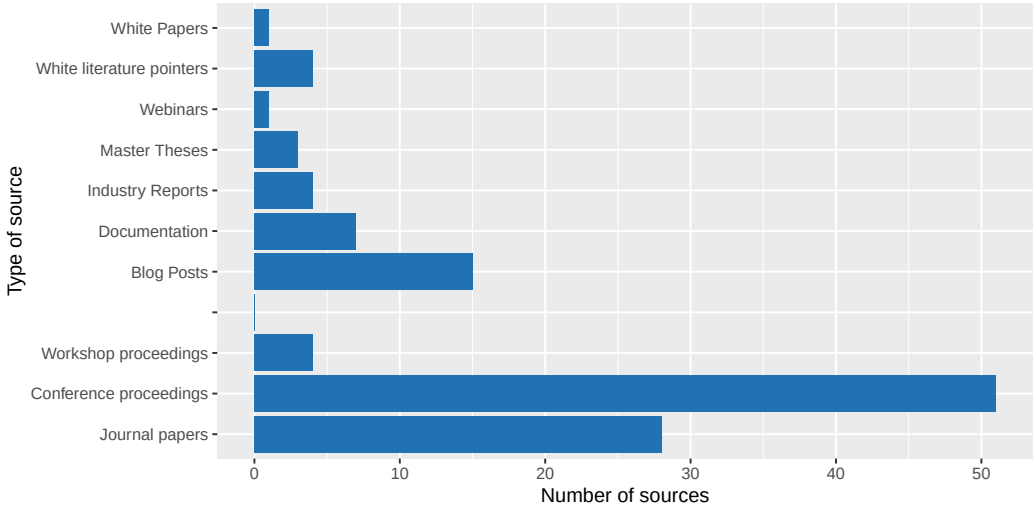

Figure 3. Types of grey and white literature sources.

Figure 4 reports the distribution of the affiliations of the authors of the white literature sources, including those gathered through Snowballing. The United States is the country with the highest number of author affiliations (84), followed by China (56) and South Korea (26).

Figure 5 shows the number of sources by type of contributors. We divided the sources into three different categories: (i) sources of which all authors were academic; (ii) sources of which all authors were working in industry; (iii) sources that were the output of a collaboration between authors from industry and from academia. All-academic sources outnumbered all-industrial sources (58 vs. 42). Of white literature sources, 52 were the output of academic studies, 18 were the output of collaborations, and 13 were industrial studies. On the other hand, most grey literature papers were industrial (29 sources vs. 6 academic studies).

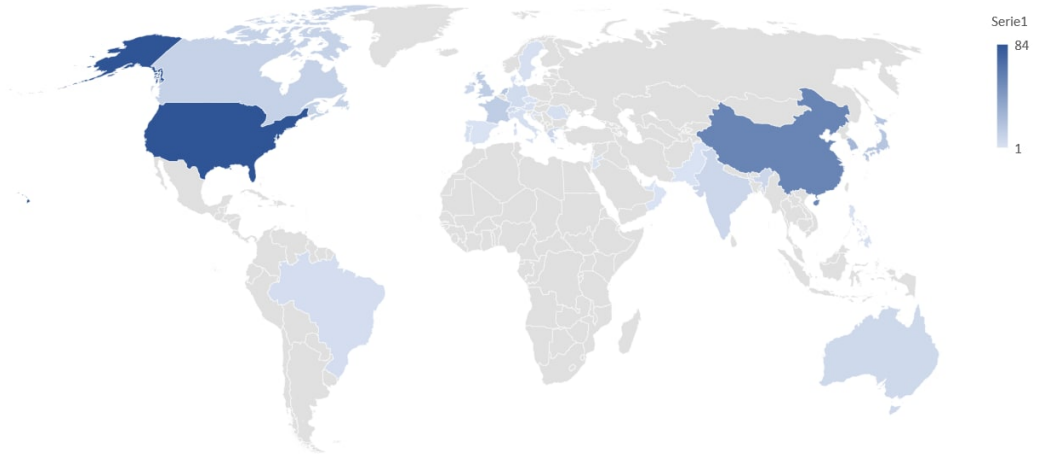

Figure 4. Affiliations of authors of white literature sources.

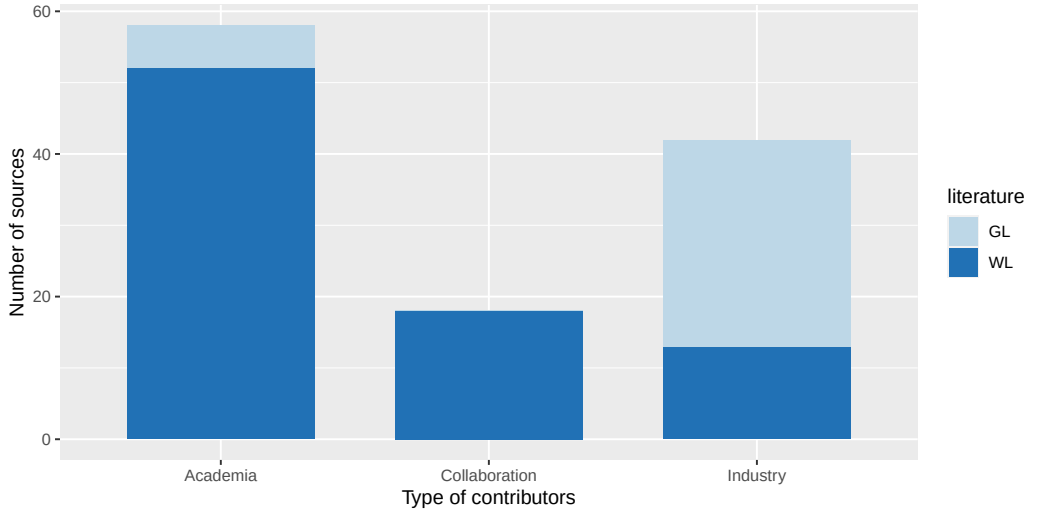

Figure 5. Number of sources by types of contributors.

4. Results

This section presents the results of our analysis of gathered sources, and the answers to the Research Questions that guided the data extraction from the selected pool of sources.

4.1. RQ1—Mapping

4.1.1. Types of Contributions (RQ1.1)

While thoroughly analyzing all papers in the final pool, we applied the Grounded Theory approach to categorize the paper's type of contribution. The categorization was based only on the main contributions of the papers (i.e., accessory content that is not deeply discussed or that is not the primary finding of a source is not considered for its categorization). The categorization was not considered mutually exclusive.

After the examination of all the sources in the final pool, we came up with the following seven categories of contributions:

- Chatbot description: sources whose primary focus is the description of the implementation of a novel conversational interface.
- Guidelines: sources (typically more descriptive and high-level) that list sets of best practices that should be adopted by chatbot developers and/or researchers to enhance their quality. Sources discussing guidelines do not need to explicitly adopt or measure quality attributes or metrics for chatbot evaluation.
- Quality attributes: sources that discuss explicitly or implicitly one or more qualitative attributes for evaluating textual conversational interfaces.

- Metrics: sources that explicitly describe metrics—with mathematical formulas—for the quantitative evaluation of the textual conversational interfaces.
- Model: sources whose main contribution is a presentation or discussion of machine learning modules finalized to enhance one or more quality attributes of textual conversational interfaces. Regarding models, many different models were mentioned in the analyzed studies. Some examples are: Cuayáhuitl et al., who adopt a 2-layer Gated Recurrent Unit (GRU) neural network in their experiments [40]; Nestorovic also adopts a two-layered model, with the intentions of separating the two components contained in each task-oriented dialogue, i.e., intentions of the users and passive data of the dialogue [41]; Campano et al. use a binary decision tree, which allows for a representation of a conditional rule-based decision process [42].
- Framework: sources that explicitly describe an evaluation framework for textual conversational interfaces, or that select a set of parameters to be used for the evaluation of chatbots. In all cases, this typology of sources clearly defines the selected attributes deemed essential to evaluate chatbots. The difference between this category and Quality attributes and Metrics lies in the combination of multiple quality attributes or metrics into a single comprehensive formula for evaluating chatbots.
- Dataset: sources that describe, make available, or explicitly utilize publicly available datasets that can be used to evaluate textual conversational interfaces.

4.1.2. Distribution of Sources per Type of Contribution (RQ1.2)

Figure 6 reports the distribution of all the studies of the final pool according to the type of contribution they provide. In the bar plot, we differentiated the number of white literature and grey literature sources providing each type of contribution. It is worth underlining that the total sum of the contributing sources is higher than the number of papers in the final pool we used for our review, since contribution type category was not an attribute that classified the sources exclusively.

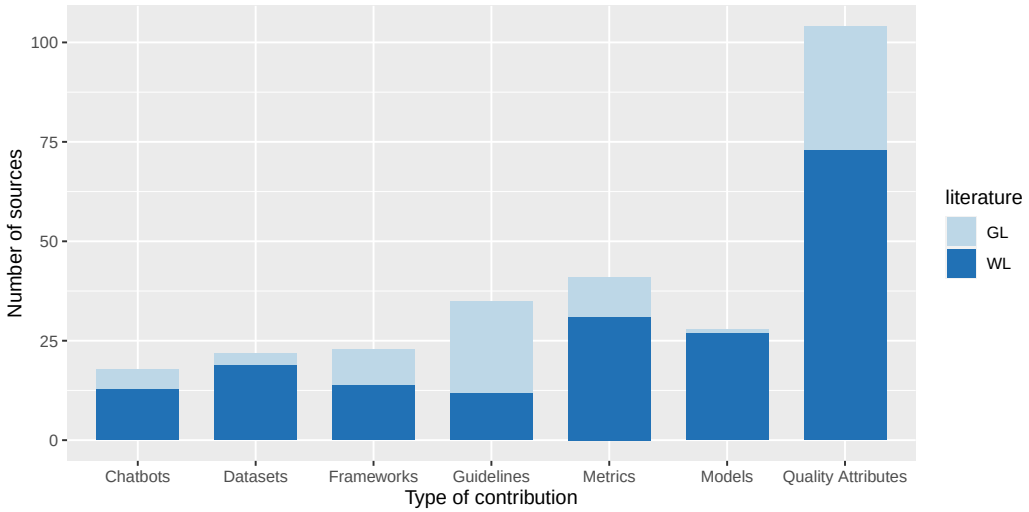

Figure 6. Number of sources for type of contribution.

The most present contribution facet was *Quality Attributes*, with 104 different sources (around 88% of the total); this result was expected since keywords related to quality evaluation and assessments were central in the search strings fed to the engines. A total of 41 sources (35%) presented metrics for the evaluation of conversational agents, and 29 sources (25%) presented guidelines for the evaluation of conversational agents. The least present category of contribution was a chatbot presentation, with only 12 sources (10%).

Figure 7 shows the total number of white literature studies for each year, grouped by the type of contribution provided. The graph shows that some contributions (especially models and published datasets) represented a significantly higher portion of the

publications in recent years. On the other hand, a high percentage of sources providing quantitative metrics date back to 2010 or before.

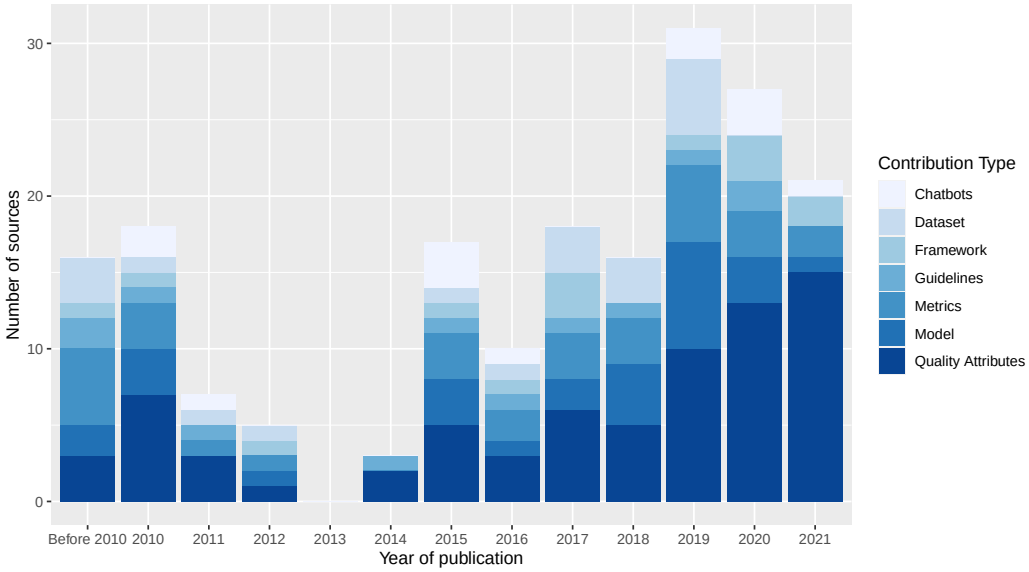

Figure 7. Number of white literature sources per year by type of contribution.

4.1.3. Mapping of Sources by Research Methodology (RQ1.3)

Several guidelines are available in the Software Engineering literature to map existing studies according to the type of research they describe. Said guidelines can be very low-level, e.g., Glass et al. propose 22 different research methods and 13 research approaches [43]. We mutuated the categorization provided by Petersen et al. [44]: to avoid having too sparse a distribution of our sources among the categories; we resorted to adopting four high-level categories to describe the research type of the analyzed manuscripts.

The four research typologies that we considered are the following:

- Descriptive and opinion studies: Studies that discuss issues about conversational interface validation and measurement and that propose metrics and frameworks for their evaluation from a theoretical perspective. This category's studies do not propose technical solutions to improve conversational interfaces or compute quality attributes upon them. Neither do they set up and describe experiments to measure and/or compare them.
- Solution proposals: Studies proposing technical solutions (e.g., new chatbot technologies, machine learning models, and metric frameworks) to solve issues in the field of conversational interfaces, and that explicitly mention quality attributes for chatbots. However, the studies only propose solutions without performing measurements, case studies, or empirical studies about them.
- Experience reports and case studies: Studies in which quality attributes and metrics about conversational interfaces are explicitly measured and quantified, in small-scale experiments that do *not* involve formal empirical methods (e.g., the definition of controlled experiments, formulation of research questions, and hypothesis testing).
- Empirical studies: Studies in which quality attributes and metrics about conversational interfaces are explicitly measured and quantified through the description and reporting of the results of formal empirical studies.

Figure 8 reports the distribution of sources from the final pool according to the type of adopted research methodology. The largest subset was that of descriptive and opinion studies (40 out of 119 sources, 34%). This number is mostly impacted by the inclusion of grey literature in the review: 27 grey literature sources (out of the total number of 35, 77%) featured documentation of existing technologies or opinion-based studies without setting up experiments, case studies, or providing quantitative means of evaluating conversational

agents. The only sources among grey literature that provided empirical evaluations were four studies that we categorize as pointers to white literature documents or Master's theses. The lowest number of occurrences was for solution proposals (15 sources, 13%). We interpret this low number to be due to the keywords used in the search strings, which led us to exclude papers proposing technological advancements, without featuring explicit evaluations of the technologies in terms of quality attributes and metrics.

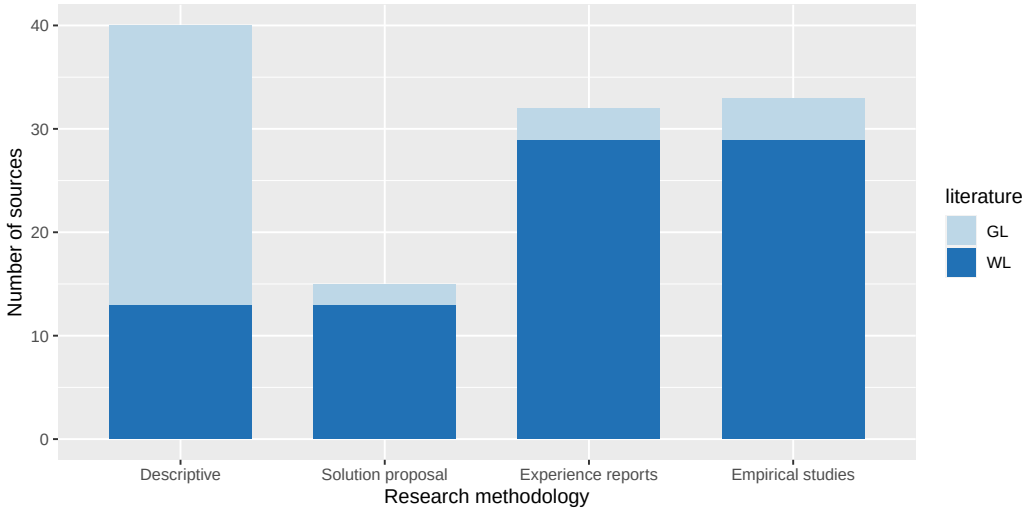

Figure 8. Number of papers by research methodology adopted.

Figure 9 shows the total number of white literature studies for each year, grouped by the type of research methodology employed. It emerged that solution proposal studies were absent in the considered pool from 2010 until 2016, and a predominance of experience reports and empirical studies in newer literature.

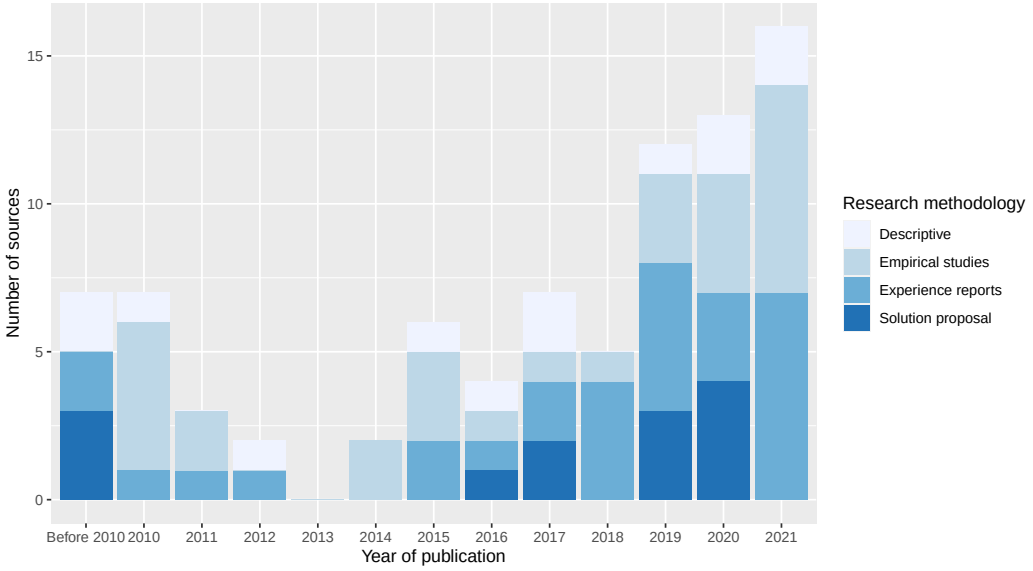

Figure 9. Number of white literature sources per year by research methodology.

4.2. RQ2—Quality Evaluation of Textual Conversational Interfaces

4.2.1. Proposed Quality Metrics (RQ2.1)

We applied the Grounded Theory methodology to define a taxonomy of the quality attributes used to evaluate textual conversational interfaces. We refer to the guidelines by P. Ralph for the definition of taxonomies through Grounded Theory in Empirical

Software Engineering [45]. We applied the Axial Coding technique to derive macro- and sub-categories in our derived taxonomy of quality attributes.

Our investigation about quality attributes used in evaluating conversational agents demonstrated that the researchers' practice had taken quite some distance from the traditional distinction between functional and non-functional quality evaluation. More recent work in the field has started considering the tight connection between conversational agents' responses and their users' emotional sphere [46,47]. Thus, the separation between the concepts of usability and functionality is not so evident for chatbots as it is for traditional categories of software.

We found four main macro-categories: Relationship, Conversation, Application Usability, and Application Metrics. Each category was divided into sub-clusters. We performed an analysis of the leaves of the taxonomy in each cluster to group together synonyms and different definitions of equivalent non-functional attributes in different sources.

Figure 10, reports the taxonomy of categories of quality attributes obtained from our literature review.

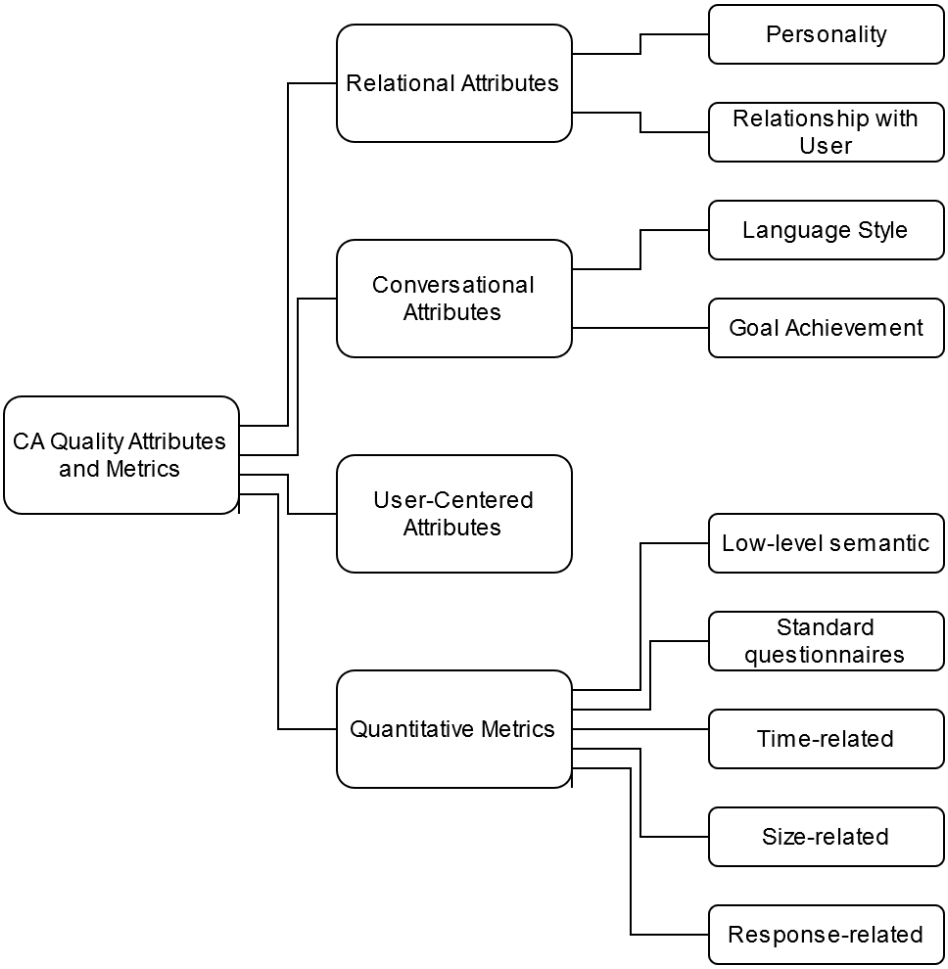

Figure 10. Taxonomy of categories of quality attributes inferred from the final pool of sources.

In Table 5, we report the full list of quality attributes found in the considered sources, along with the list and number of sources mentioning each of them.

Below we describe the taxonomy categories and report some examples of how the quality attributes are described in the primary sources:

- Relational attributes. Quality attributes that measure the relationship with the user on human-related aspects or describe the human characteristics of a chatbot. Relational aspects do not directly affect the communication correctness but rather enrich it by creating emotional bonds between the user and the chatbot.

These attributes cannot always be clearly separated from functionality, since in various applications establishing a human connection with the user is the main functional goal for which the conversational agent is used (e.g., in the medical field). As an example, Looije et al. report that "Research on persuasive technology and affective computing is providing technological (partial) solutions for the development of this type of assistance, e.g., for the realization of social behavior, such as social talk and turn-taking, and empathic behavior, such as attentiveness and compliments" [48].
Among Relational Attributes, we identify two sub-categories:

- Personality: Attributes that are related to the perceived humanness of the chatbot, which are generally reserved to describe essential and distinctive human traits.
  In this category, the most prominent attribute considered is the Social Capacity of the Conversational Agent. Chen et al., for instance, [46] identify several attributes that can all be considered as expressions of the social capacity of an agent working in a smart home, and that they translate into several guidelines for the behavior of a chatbot (e.g., "Be Friendly", "Be Humorous", "Have an adorable persona").
  Several manuscripts also refer to the Empathy of chatbots, which implies the ability of the chatbot to correctly understand the emotional tone of the user and avoid being perceived as rude.
  Another frequently mentioned attribute is the Common Sense of the Conversational Agent, also defined as the context sensitiveness of the agent [49], or the match between the system and the real world [50].
  Many studies in the pool have also mentioned the ethics of the Conversational Agents: in a grey literature source about the faults of a commercial chatbot, the concept is defined as as "the need to teach a system about what is not appropriate like we do with children" [51].
- Relationship with the user: quality attributes that directly affect the relationship between the chatbot and the user. Trust [52] and Self-Disclosure [53], for instance, are essential to triggering rich conversations. Memory (also defined as User History) is a basic principle to keep alive the relationship with the user over time. Customization (also defined as Personalization, User-Tailored content [47], Personalized experience [54]) is an important parameter to improve the uniqueness of the conversation for each specific user. Finally, Engagement [46] and Stimulating Companionship [55] are relevant attributes to measure the positive impact on the user.

- Conversational attributes. Quality attributes are related to the content of the conversation happening between the chatbot and the user. We can identify two sub-categories of Conversational attributes:

  - Language Style: attributes related to the linguistic qualities of the Conversational Agents' language.
    The most mentioned language style attribute is Naturalness, defined by Cuaydhuitl et al., as the fact that the dialogue is "naturally articulated as written by a human", [40], also referred to as Human-like Tone and Fluency.
    Relevance refers to the capability of the system to convey the information in a way that is relevant to the specific context of application [56], to keep the answers simple and strictly close to the subjects [57], and to avoid providing information overload to the users [58].
    Diversity refers to the capability of the chatbot to use a varied lexicon to provide information to the users and the capability to correctly manage homonymy and polysemy [59].
    Conciseness is a quality attribute that takes into account the elimination of redundancy without removing any important information (a dual metric is Repetitiveness).

  - Goal Achievement: attributes related to the way the chatbot provides the right responses to the users' goal. They measure the correctness of the output given by the chatbot in response to specific inputs.
    The most cited quality attribute for this category is the Informativeness of the chatbot, i.e., the chatbot capability to provide the desired information to the user in a given task. Informativeness is also defined as Usefulness, measured in terms of the quantity of the content of the answers given to the users [60], or Helpfulness [61].
    Correctness instead evaluates the quality of the output provided by the chatbot measured in terms of the correct answers provided to the users [62], and the accuracy of the provided content.
    Proactiveness (in some sources referred to as Control of topic transfer [63], Initiate a new topic appropriately [46], Topic Switching [64] and Intent to Interact [65]), is the capability of the chatbot to switch or initiate new topics autonomously.
    Richness is defined as the capability of the chatbot to convey rich conversations with a high diversity of topics [66].
    Goal achievement attributes include the capability of the chatbot to understand and tailor the conversation to the Context (i.e., Context Understanding, also defined as Context-Awareness [57], Context Sensitiveness [49] and Topic Assessment [67,68]) and to the user (i.e., User Understanding).

- User-centered Attributes: attributes related to the user's perception of a chatbot. These attributes are mostly compatible with traditional Usability software non-functional requirements.
  The most-frequently cited user-centered attributes are Aesthetic Appearance, User Intention to Use, and User Satisfaction.
  Aesthetic Appearance refers to the interface that is offered to the user for the conversation. Bosse and Provost mentions the benefits of having photorealistic animations [69]; Pontier et al. performed a study in which the aesthetic perception of the participants was evaluated in a scale going from attractive to bad-looking [70].
  User Satisfaction is defined as the capability of a conversational agent to convey competent and trustworthy information [71], or the capability of the chatbot to answer questions and solve customer issues [72].
  User Intention to Use is defined as the intention of a user to interact with a specific conversational agent [65]. Jain et al. measured the user's intention to use the chatbot again in the future as an indicator of the quality of the interaction [73].
  Ease of Use is defined as the capability of the chatbot to offer easy interaction to the user [57], i.e., the capability of the chatbot to allow the users to write easy questions that are correctly understood [58], and to keep the conversation going with low effort [74]. The ease of use of a chatbot can be enhanced, for instance, by providing routine suggestions during the conversation [75].
  Other important parameters for the usability of a conversational agent are the Mental Demand and Physical Demand required by an interaction with it [50,56,73].
- Quantitative Metrics. Quantitative metrics can be objectively computed with mathematical formulas. Metrics are generally combined to provide measurements of the quality attributes described in the other categories.
  We can divide this category of quality attributes into the following sub-categories:

  - Low-Level Semantic: grey box metrics that evaluate how the conversational agent's models correctly interpret and classify the input provided by the user.
    Several papers, especially those based on the machine learning approaches, report metrics related to these fields, e.g., the use of word embeddings metrics [76] or confusion matrices [77].
    The most cited metrics of the category are common word-overlap-based metrics (i.e., they rely on frequencies and position of words with respect to ground truth, e.g., a human-annotated dataset), BLEU, ROUGE, METEOR, CIDEr, and word-

embedding-based metrics like Skip-Thought, Embedding Average, Vector Extrema and Greedy Matching [60].
Low-level Semantic metrics are typically employed to evaluate the learning models adopted by the chatbots and to avoid common issues like underfitting (i.e., the inability to model either training data or new data) which translates to a very low BLEU metric, or overfitting (i.e., poor performance on new data and hence small generalizability of the algorithm) which translates to a very high BLEU metric.
In this category of metric we also include traditional metrics used to evaluate the prediction of Semantic Textual Similarity based on Regression Models. Examples are the Mean Squared Error (MSE), defined as the average of sum of squared difference between actual value and the predicted or estimated value; Root Mean Squared Error (RMSE) that considers whether the values of the response variable are scaled or not; and R-Squared metric (or R2), defined as the ratio of Sum of Squares Regression (SSR) and Sum of Squares Total (SST). Values closer to 0 for (R)MSE, or closer to 100% for R2, indicate optimum correlation.

- Standard Questionnaires: sets of standard questions to be answered using a Likert scale, and that can be used to quantify abstract quality attributes defined in the previous categories.
  Edwards et al. list three instruments belonging to this category: the Measure of Source Credibility, an 18-item instrument designed to assess perceptions of an individual's credibility across three dimensions of competence, character and caring; the Measure of interpersonal attraction, to assess attraction to another along two dimensions, task and social; and the Measure of computer-mediated Communication Competence, to examine how competent the target was perceived as in the communication by the participants [65].
  Valtolina et al., in the context of an evaluation of the usability of chatbots for healthcare and smart home domains, leveraged three standard questionnaires for general-purpose software: SUS (System Usability Scale), CSUQ (Computer System Usability Questionnaire), and UEQ (User Experience Questionnaire) [78].
- Time-related metrics: time and frequency of various aspects of the interaction between the user and the chatbot. The most mentioned metric is the Response Time, i.e., the time employed by the chatbot to respond to a single request by the user [79]. Other manuscripts report the time to complete a task (i.e., a full conversation leading to a result) or the frequency of requests and conversations initiated by the users.
- Size-related metrics: quantitative details about the length of the conversation between the human user and the chatbot. The most common is the number of messages, also referred to as the number of interactions or utterances, with several averaged variations (e.g., the number of messages per user [41] or per customer [80]).
- Response-related metrics: measures of the number of successful answers provided by the chatbot to meet the users' requests. Examples of this category of metrics are the frequency of responses (among the possible responses given by the chatbot) [81], the task success rate (i.e., entire conversations leading to a successful according to the user's point of view) [82], the number of correct (individual) responses, the number of successful (or, vice-versa, of incomplete) sessions [83].

Figure 11 reports the number of quality attributes and metrics for each of the categories that were inferred through Grounded Theory. The highest number of quality attributes belongs to the category of Quantitative Metrics. This result can be justified by the higher specificity of quantitative metrics, instead of higher-level quality attributes that are filed under the other categories. A total of 123 metrics were found in the selected pool of primary

sources, divided in these proportions: 30 Relational attributes, 22 Conversational attributes, 20 User-Centered attributes, and 51 Quantitative metrics.

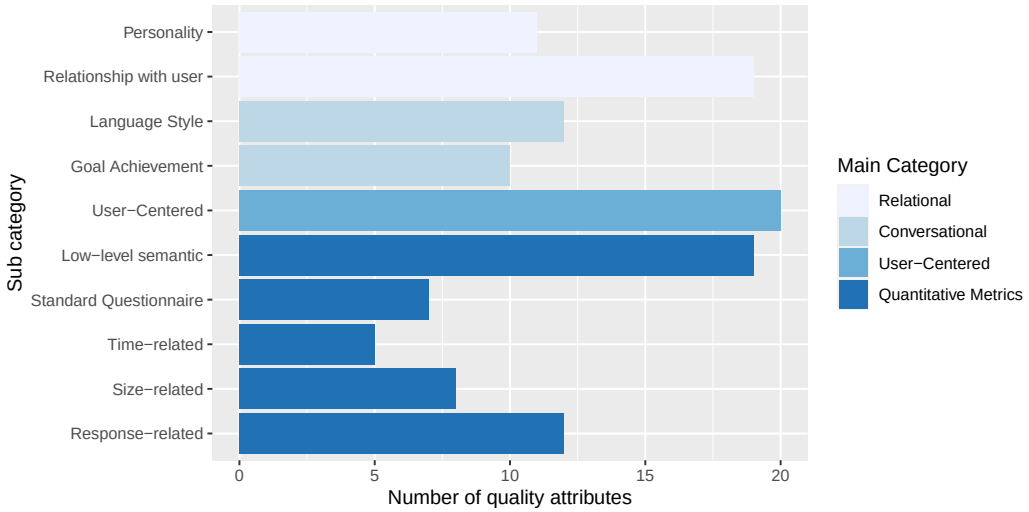

Figure 11. Number of quality attributes per category.

Figure 12 summarizes the number of different papers mentioning attributes in a given category, divided by the typology of sources (WL or GL). The most common metrics in sources from both typologies of literature are those belonging to the categories Relationship with the User, Goal Achievement, and User-Centred attributes (respective totals of 43, 42 and 41 different papers mentioning them). Few different sources defined or used Quantitative Metrics, ranging from 2 sources mentioning Standard Questionnaires to 12 sources mentioning Low-level Semantic metrics.

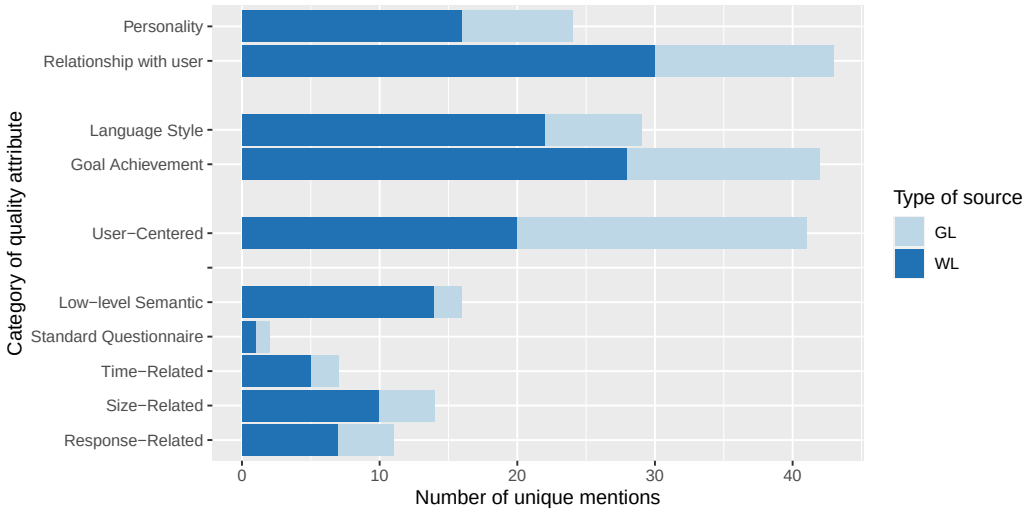

Figure 12. Number of unique mentions per category of quality attributes.

4.2.2. Proposed Frameworks for the Evaluation of Conversational Agents (RQ2.2)

In this section, we describe a set of explicitly mentioned frameworks, proposed, and/or implemented in the final pool sources that we examined. A summary of the frameworks is provided in Table 6.

- ADEM: an automated dialogue evaluation model that learns to predict human-like scores to responses provided as input, based on a dataset of human responses collected using the crowdsourcing platform Amazon Mechanical Turk (AMT).

The ADEM framework computes versions of the Response Satisfaction Score and Task Satisfaction Score metrics. According to empirical evidence provided by Lowe et al. [84] the framework outperforms available word-overlap-based metrics like BLEU.

- Botest: a framework to test the quality of conversational agents using divergent input examples. These inputs are based on known utterances for which the right outputs are known. The Botest framework computes as principal metrics the size of the utterances and the conversations and the quality of the responses.
  The quality of responses is evaluated at syntactical level by identifying several possible errors, e.g., word order errors, incorrect verb tenses, and wrong synonym usage.
- Bottester: a framework to evaluate conversational agents through their GUIs. The tool computes time and size metrics (mean answer size, answer frequency, word frequency, response time per question, mean response time) and response-based metrics, i.e., the number and percentage of correct answers given by the conversational interface. The tool receives the files with all questions submitted, the expected answers, and configuration parameters for the specific agent to test to give an evaluation of the proportion of correct answers. The evaluation provided by the tool is focused on the user perspective.
- ParlAI: a unified framework for testing dialog models. It is based on many popular datasets and allows seamless integration of Amazon Mechanical Turk for data collection and human evaluation, and is integrated with chat services like Facebook Messenger. The framework computes accuracy and efficiency metrics for the conversation with the chatbot.
- LEGOEval: an open-source framework to enable researchers to evaluate dialogue systems by means of the online crowdsourced platform Amazon Mechanical Turk. The toolkit provides a Python API that allows the personalization of the chatbot evaluation procedures.

4.2.3. Proposed Datasets for the Evaluation of Conversational Agents (RQ2.3)

Table 7 reports the list of datasets mentioned in the selected pool of sources. In our analysis of the pool, we do not find any dataset explicitly defined to evaluate chatbots. Exceptions are made for conversational agents based on machine and deep learning approaches, where datasets are divided into training, validation, and test sets to validate the machine learning approaches.

Table 5. Complete list of quality attributes in primary sources.

Macro-Category	Sub-Category	Attribute	WL Refs.	GL Refs.	Refs.
Relational	Personality	Social Capacities	[46,48,69,85–88]	[49,58,59,64,89]	12
		Common Sense	[90–92]	[49,51,56,59]	7
		Ethics	[70,91,93]	[49,51]	5
		Empathy	[47,48,94]	[59,95]	5
		Freewill	[92,96]	-	2
		Extraversion	[91,97]	-	2
		Warmth	[85]	-	1
		Judgement	-	[51]	1
		Perceived Intelligence	[87]	-	1
		Neuroticism	[91]	-	1
		Openness	[91]	-	1
	Relationship with user	Customization	[47,52,71,98]	[49,54,72,99–101]	10
		Trust	[48,52,53,62,87,102,103]	[59,104]	9
		Believability	[42,65,69,71,96,97,105]	[59]	8
		Engagingness	[40,42,70,87,106,107]	[57,89]	7
		Memory	[47,60,108]	[54,58,64]	6
		Companionship	[46,55,105,109]	[54,110]	6
		Adaptability	[47,55,87,111]	-	4
		Playfulness	[62,71,77,87]	-	4
		Utilitarian Value	[62],	[72,99,104]	4
		Affect Understanding	[102,105,112]	[57]	4
		Reliability	[55,92,109]	-	3
		Intimacy	[53,109]	-	2

Table 5. *Cont.*

Macro-Category	Sub-Category	Attribute	WL Refs.	GL Refs.	Refs.
		Modeling Capability	[55]	[54]	2
		Persuasiveness	[48]	-	1
		Reciprocity	[53]	-	1
		Self-Disclosure	[53]	-	1
		Perceived Sacrifice	[103]	-	1
		Transparency	[105]	-	1
		Clarity	[105]	-	1
Conversational	Language Style	Naturalness	[40,63,68,87,91,102] [92,94,98,103,107,113]	[58,59,95,101],	16
		Diversity	[46,114,115]	[59]	4
		Interaction Style	[67,69]	[56,59]	4
		Language Relevance	[73]	[56–58]	4
		Conciseness	[116]	[57]	2
		Repetitiveness	[42,47]	-	2
		Colloquiality	[46]	-	1
		Initiative to User	[41]	-	1
		Lexical Performance	[117]	-	1
		Politeness	-	[54]	1
		Restatement	[118]	-	1
		Shallowness	[47]	-	1
	Goal Achievement	Informativeness	[46,60,63,108,116], [62,73,87,109,111]	[54,56,57,64,110]	16
		Correctness	[46,62,65,69,71,85,103]	[100,119]	12
		Context Understanding	[46,60,67,68,98]	[49,57]	7
		Proactiveness	[47,63,65,93]	[58,64]	6
		Richness	[46,60,108,114]	[49,120]	6
		Consistency	[40,46,111]	[56,57]	5
		Clarity	[47,55,98]	[59,110]	5
		Relevance	[46,60,70]	[59]	4
		Robustness	[117]	[51]	2
		User Understanding	[117,121]	-	2
User-Centered Attributes		Aesthetic Appearance	[47,69,70,86,87,92,98]	[49,50,56–58,89]	13
		User Satisfaction	[53,62,71,103,106,122]	[54,57,59,82,95,99,104]	13
		Ease of Use	[62,74,87,116,123]	[57,58,99,104,124]	10
		User Intention to Use	[53,62,65,70,73]	[57,82,124]	8
		Mental Demand	[62,73,87]	[50,54,56,58,99]	8
		Availability	[47]	[54,64,72,95,99]	6
		Acceptability	[48,93]	[59,124]	4
		Presence of Notifications	[47,74]	[75,100]	4
		Protection of Sensitive Data	[116]	[59,101]	3
		Number of Channels	-	[100,101]	2
		Physical Demand	[73]	[99]	2
		Presence of Documentation	-	[56,78]	2
		Hedonic Value	[62,103]	-	2
		Information Presentation	-	[49]	1
		Integration in External Channels	-	[99]	1
		Presence of Ratings	-	[49]	1
		Responsiveness (Graphical)	-	[80]	1
		Integration with multiple systems	-	[119]	1
		Human Escalation	-	[100]	1
		Number of Languages	-	[100]	1
Quantitative	Low-level semantic	ML Accuracy	[63,90,125–129]	[82]	8
		BLEU	[60,63,76,81,108,130,131]	-	7
		ML Precision	[63,90,125,126,128]	[82]	6
		ML Recall	[63,90,125,126]	[82]	5
		Distinct-1	[63,81,130,131]	-	4
		METEOR	[60,76,108,132]	-	4
		ROUGE	[60,76,108,131]	-	4
		CIDEr	[60,108]	-	2
		Skip-Thought	[76,108]	-	2
		Embedding Average	[76,108]	-	2
		Vector Extrema	[76,81]	-	2
		Perplexity (PPL)	[130,131]	-	2
		Greedy Matching	[76,108]	-	1
		Distinct-2	[81]	-	1

Table 5. *Cont.*

Macro-Category	Sub-Category	Attribute	WL Refs.	GL Refs.	Refs.
		ASR Confidence Score	-	[120]	1
		NLU Confidence Score	-	[120]	1
		Semantic Similarity Metrics	[108]	-	1
		MRR	[67]	-	1
		P1	[67]	-	1
	Standard Questionnaires	Measure of Source Credibility	[65]	-	1
		Measure of interpersonal attraction	[65]	-	1
		Measure of computer-mediated Communication Competence	[65]	-	1
		Communicability Evaluation Method (CEM)	-	[78]	1
		System Usability Scale (SUS)	-	[78]	1
		Computer System Usability Questionnaire (CSUQ)	-	[78]	1
		User Experience Questionnaire (UEQ)	-	[78]	1
	Time-Related Metrics	Response Time	[79,116,133]	[57]	4
		Task Completion Time	[73,97,116,133]	-	4
		Peak Usage Time	-	[80]	1
		Frequency of Requests	-	[120]	1
		Frequency of Conversations	[116]	-	1
	Size-Related Metrics	Number of messages/utterances	[73,97,115,127,134–136]	[54,64]	9
		Number of conversations	[133–135]	[54,64]	5
		Number of Keywords	[68,77,135]	-	3
		Number of users	[134]	[54,57,119]	4
		Mean Answer Size	[79,97,135]	-	3
		Number of user ended sessions	-	[54]	1
		Number of system ended sessions	-	[54]	1
		Number of emotions per dialogue	[135]	-	1
	Response-Related Metrics	Response Frequency	[79,81]	-	2
		Number of successful sessions	-	[83,137]	2
		Number of correct responses	-	[64,83]	2
		Topic Diversity	[138]	[120]	2
		Word Frequency	[81]	-	1
		Number of queries related to the topic	[116]	-	1
		Number of utterances with poor understanding	-	[82]	1
		Response Satisfaction Score	[139]	-	1
		Task Success Rate	[79]	-	1
		Reliability Coefficient	[65]	-	1
		User Satisfaction Estimate (USE)	-	[120]	1
		Response Specificity	[81]	-	1

Table 6. Frameworks for evaluating conversational interfaces in the selected pool of primary sources.

Framework	References	Language	License
ADEM	[12,84]	Python	-
BoTest	[140]	Python	Open Source
Bottester	[79]	Mocha/chai	MIT License
ParlAI	[141]	Python	MIT License
LEGOEval	[107]	Python	Open Source

Table 7. Categories and examples of datasets mentioned in the selected pool of sources.

Type of Source	Source Examples	WL References	GL References
Forums	IMDB Movie Review Data	[90]	-
	Ubuntu Dialog corpora	[108,142]	-
	Yahoo! Answer	[67,114]	-
	2channel Internet bulletin board	[68]	-
	Slashdot	[68]	
Social Media	Twitter	[46,47,60,68,71,76,90,96,108,114,117,121,130,142,143]	[64,144]
Customer Service	Twitter Customer Service Corpus	[60]	
	Didi Customer Service Corpus	[63]	-
Movies and Subtitles	OpenSubtitles	[60]	-
	Cornell Movie Dialog Corpus 3	[63,67]	-
	Movie-DiC	[67,145]	-
	Hello-E	[131]	-
	CMU_DoG	[131]	-
	DuConv	[131]	-
Challenges	Persona-Chat dataset	[40]	-
	Dialog State Tracking Challenge datasets	[81,130,143]	-
	Dialog System Technology Challenges (DSTC6) Track 2	[60,108,143]	-
	NTCIR-13 Short Text Conversation	[81]	
External Knowledge	WordNet, WordNet Affect	[76,112,114,115,117,118,139,146,147]	-
	Wikipedia	[68,90,114,141]	-
	Wizard of Wikipedia	[131]	-
	OpenDialKG	[131]	-
	Unified Medical Language System (UMLS)	-	[59]
	SNOMED CT® (Clinical Terms)	-	[59]
	Bioportal	-	[59]
	CISMEF	-	[59]
	Schema.org	-	[50]
Others	Debate Chat Contexts	[67]	-

In this section, we list all the specific data sources used to evaluate chatbots or mentioned in the papers that we evaluated, with a categorization based on the type of data they contain.

- Forums and FAQs. Since forums and FAQ pages are based on a question and answer, conversational nature, they can be leveraged as sources of dialogue to train and test chatbots.
  The datasets of this category used in the examined sources are the following:
  - IMDB Movie Review Data: a dataset of 50K movie reviews, containing 25,000 reviews for training and 25,000 for testing, with a label (polarity) for sentiment classification. The dataset is used, for instance, by Kim et al. to address common sense end ethics through a neural network for text generation [90].
  - Ubuntu Dialog Corpora: a dataset of 1 million non-labeled multi-turn dialogues (more than 7 million utterances and 100 million words). The dataset is oriented to machine learning and deep learning algorithms and models.
  - Yahoo! Answers: datasets hosted on GitHub repositories, containing scrapings of the questions and answers on the popular community-driven website.
  - 2channel: a popular Japanese bulletin board that can be used as a source for natural, colloquial utterances, containing many boards related to diverse topics.
  - Slashdot: a social news website featuring news stories on science, technology, and politics, that are submitted and evaluated by site users and editors.
- Social Media. Social media are sources of continuous, updated information and dialogue. It is often possible to model the users and access their history and metadata (e.g., preferences, habits, geographical position, personality, and language style). It is also possible to reconstruct the social graphs, i.e., the social relations between users and how they are connected. Several social media, to maximize the exploitation of

data and render a higher amount of analyses possible, make their data available through APIs, allowing developers to extract datasets.

  - Twitter: it provides simple and accessible tools and API to download tweets and related metadata. Moreover, it allows performance of data scraping.

- Customer Service: collections of technical dialogues for customer service coming from multiple sources, such as private and company channels, social platforms, or e-commerce platforms. Some examples are the following:
  - Twitter Customer Service Corpus: a large corpus of tweets and replies to and from customer service accounts on Twitter.
  - Didi Customer Service Corpus: a dataset of 11,102 Facebook wall posts, 6507 wall comments, and 22,218 private messages from 136 South Tyrolean users who participated in the "DiDi" project in the year 2013.
- Movies and Subtitles. Subtitles can be used to collect dialogues by modeling the characters and, virtually, the users, if metadata are present.
  - Open Subtitles Corpus: A dataset composed of movie and television subtitle data from OpenSubtitles, in 62 languages [148]. Consecutive lines in the subtitle data could be used as conversational dialogues.
  - Cornell Movie Dialog Corpus 3: an extensive collection of fictional conversations extracted from raw movie scripts. There are 220,579 conversations between 10,292 pairs of movie characters, 9035 characters from 617 movies.
  - Movie-DiC: a dataset comprising 132,229 dialogues containing a total of 764,146 turns that have been extracted from 753 movies.
  - HelloE: a dialogue dataset containing around 9K conversations from about 921 movies.
  - CMU_DoG: a conversation dataset where each conversation is followed by documents about popular movies from Wikipedia articles. The dataset contains around 4K conversations.
  - DuConv: a conversation dataset where each conversation is grounded with a factoid knowledge graph, containing 270K sentences within 30K conversations.
- Challenges. Chatbot competitions are regularly held by both research and industry. Generally, datasets are given and left available for these competitions even after their conclusion; these datasets are often annotated by humans. The transcripts of the competitions are usually made available by the authors or organizers and can be utilized for further comparisons.
  - Persona-Chat dataset: a dataset consisting of conversations between crowd-workers who were paired and asked to act as a given provided persona. The dataset is oriented to machine learning and deep learning algorithms and models.
  - Dialog State Tracking Challenge datasets: a corpus collecting multi-turn dialogues from Amazon Mechanical Turk.
  - Dialog System Technology Challenges (DSTC6) Track 2: a dataset composed of customer service conversations scraped from Twitter.
  - NTCIR-13 Short Text Conversation: a dataset used in the NTCIR conference context to clarify the effectiveness and limitations of retrieval-based and generation-based methods and to advance the research on automatic evaluation of natural language conversation.
- External Knowledge. Sources of high interest for those chatbots that require specific knowledge in a specialized domain. They are also often represented as knowledge graphs and are useful for semantic and conceptual abstraction. Prominent examples of this category are medical knowledge bases. In our pool of sources, the following examples are used:
  - WordNet: a lexical database of English, where nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct

concept interlinked in a network utilizing conceptual–semantic and lexical relations.
  - Wikipedia and related projects: the content of the free crowd-based online encyclopedia and its related projects, as Wikidata, the open knowledge base that acts as central storage for the structured data of other Wikimedia projects.
  - Wizard of Wikipedia: an open-domain dialogue benchmark containing 22K conversations linked to knowledge retrieved from Wikipedia.
  - OpenDialKG: a human-to-human dialogue dataset containing 91K utterances across 15K dialog sessions about varied topics.
  - Unified Medical Language System (UMLS): created by MedLine (the U.S. National Library of Medicine bibliographic database), the UMLS, or Unified Medical Language System, it brings together many health and biomedical vocabularies and standards.
  - SNOMED CT: a dataset of clinical terminology for Electronic Health records to be used in electronic clinical decision support, disease screening and enhanced patient safety.
  - Bioportal: a dataset released by NCBO (National Center for Biomedical Ontology) and accessible through a web portal that provides access to a library of biomedical ontologies and terminologies.
  - CISMEF: Catalogue et Index des Sites Médicaux de langue Française, a French database dedicated to teaching and research.
  - Schema.org: founded by Google, Microsoft, Yahoo, and Yandex, and maintained by an open community process, Schema.org has the purpose of creating, maintaining, and promoting schemas for structured data on the Internet, on web pages, and other applications.
- Others:. Some heterogeneous datasets cannot be filed under the previous categories. We report the following example mentioned in one source of our set:
  - Debate Chat Contexts Wang et al. reported the creation of chat contexts spanning a range of topics in politics, science, and technology [67]. These chats can be considered examples of typical conversations between humans and can be used to train the chatbots to improve their evaluation regarding Conversational Quality Attributes.

4.3. RQ3—Comparison between Formal and Grey Literature

4.3.1. Industrial Contributions Leveraged by White Literature (RQ3.1)

To answer RQ3.1, we analyzed white literature to find mentions of industrial products, to conduct experiments or to compute quality attributes. Figure 13 reports the number of mentions in white literature for these technologies. The highest number of mentions (14) for a product from the industry was obtained by Alexa, the platform from Amazon that provides developer sets for the creation of skills, i.e., sets of voice-driven capabilities [80]. Amazon has defined several metrics to evaluate the performance of Alexa skills (e.g., in terms of time and size of the conversations) [83]. Alexa is closely followed in mentions by the Facebook platform (13), which offers developer tools allowing the creation of bots on the Messenger messaging platform. Other frequently mentioned products are Siri, the conversational agent developed by Apple, Cortana by Microsoft (that comes with the definition of many measurable quality attributes [120,149]) and Watson by IBM (for which a series of best practices have been defined [82]).

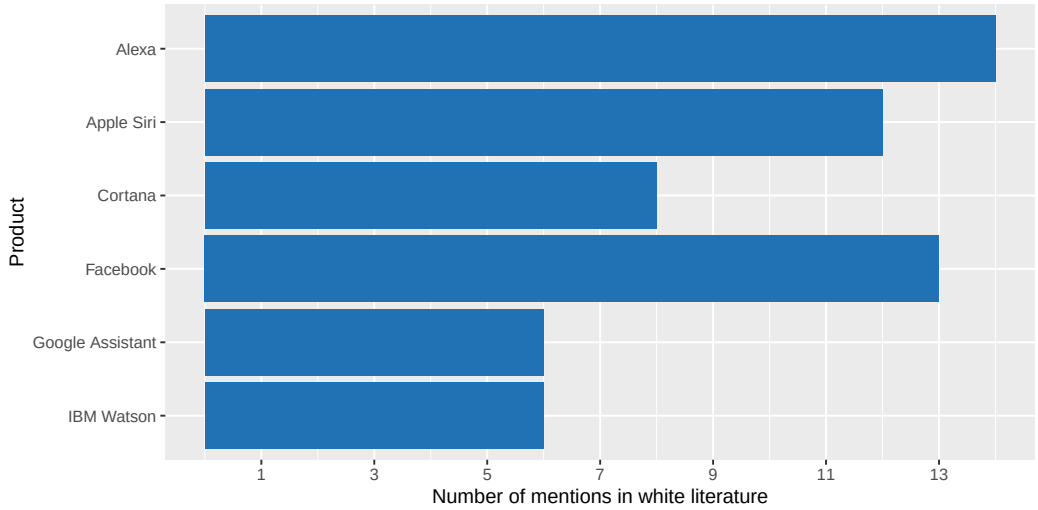

Figure 13. Number of mentions for industrial products in selected white literature sources.

4.3.2. Attention from Formal and Grey Literature (RQ3.2)

To answer RQ3.2, we analyze how much the selected white literature and grey literature sources contributed to the answers to the previous research questions of the paper.

In Figure 14, we report how many metrics for each category were exclusively present in white literature and grey literature and how many are common to both types of sources. A total of 27 quality attributes (23% of the total) were only presented or discussed in grey literature sources. The categories where the contribution of grey literature was more significant were those closest to the user point of view, i.e., the User-Centred, Standard Questionnaires and Response-related quality attributes. This result can be justified by the fact that the grey literature on the field of conversational agents is typically more related to real-world measurements on commercial chatbots. Hence, it is based on the measurements of responses provided to real users of such systems. Conversely, white literature on the topic is more related to the definition of models to drive the conversation between the human and the conversational agent. Hence, it favours the discussion of Low-level semantic metrics. We also observe a predominancy of white literature in the exploration of relational attributes of the conversational agents, which involve human science-related analyses.

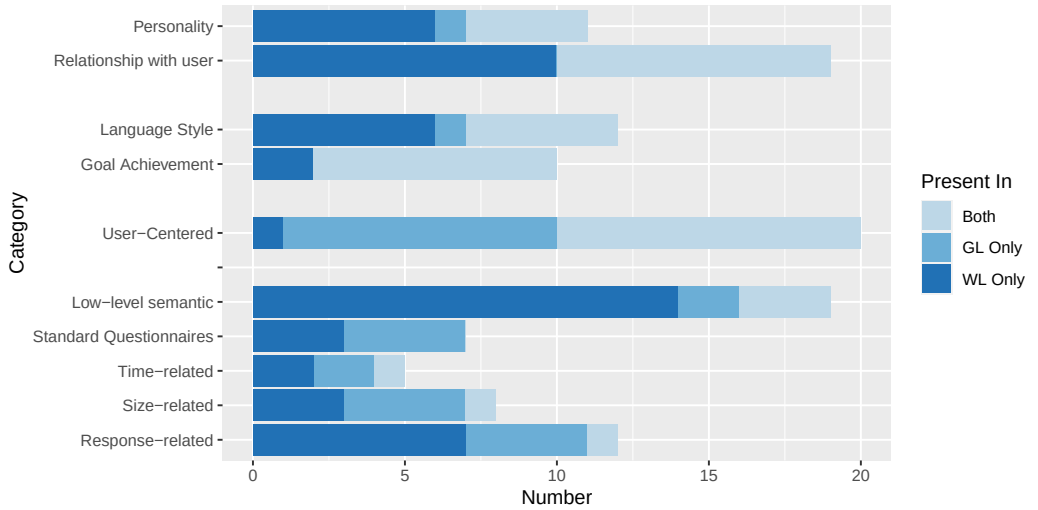

Figure 14. Exclusive and common quality attributes for typology of source, grouped by attribute category.

In Figure 15, we report the number of datasets exclusively mentioned, used, or presented in white literature and grey literature, and the number of ones common to both types of sources. The contribution from grey literature to the knowledge gathered to answer RQ2.3 is not negligible, since 5 of the 28 datasets are mentioned only in grey literature sources (while 22 are mentioned only in white literature, and one, Twitter, is mentioned in both the types of literature). More specifically, the medical databases and Schema.org (see Section 4.2.3) are only retrievable in grey literature sources.

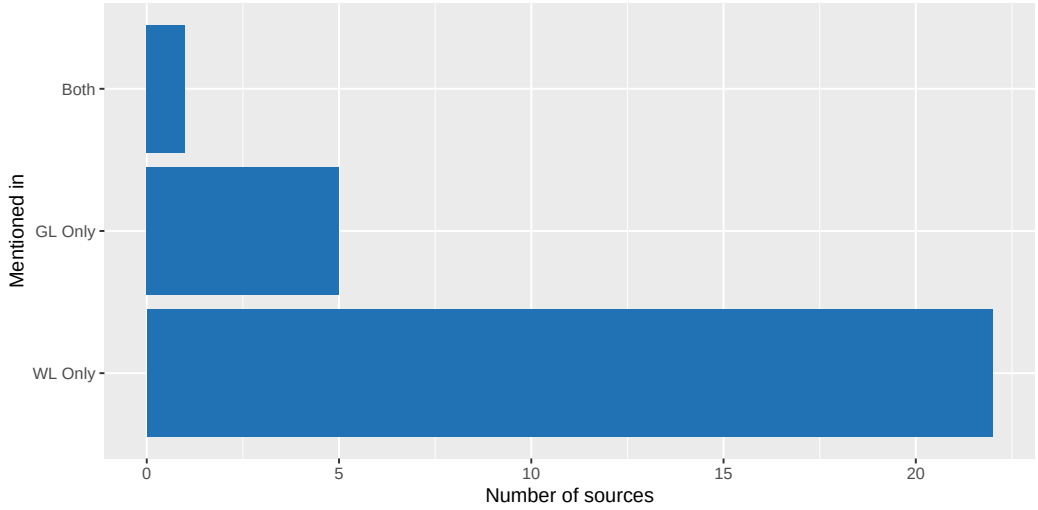

Figure 15. Exclusive and common datasets for typology of source.

5. Discussion

5.1. Summary of Findings

The main objective of our work was to classify quality attributes for conversational agents and describe the different categories under which they can be filed. To that extent, we considered both white and grey literature to analyze quality attributes mentioned and utilized in peer-reviewed manuscripts, procedures, and evaluations used by practitioner and industrial literature.

The first goal was to produce a mapping of all the studies, subdividing them according to the type of contribution provided and the research methodology employed. We found that the white literature sources discussing conversational interface evaluation and assessment were homogeneously distributed between the different categories. In contrast, grey literature leaned towards presenting quality attributes, guidelines, and frameworks, with less formalized models, metrics, and presentation of chatbots and datasets. This result can be justified by the nature of most of the grey literature sources considered, e.g., blog posts that are less likely to feature quantitative studies than formally published academic works.

By analyzing the distribution per year of the typologies of contributions in white literature from 2010 to 2021, we could deduce two principal trends: in general, more peer-reviewed papers about conversational interface evaluations have been published in recent years than at the beginning of the decade; furthermore, we can observe an increasing trend in the number of papers defining and/or using systematically quality attributes and metrics for the evaluation of conversational agents. Identifying these trends can encourage researchers in the field of software metrics to adopt established frameworks and software metrics instead of defining new ones, given that a relevant corpus of quality attributes has been defined in the latest years.

To answer RQ2.1, we built a taxonomy of quality attributes for conversational agents, obtaining ten different typologies of attributes grouped in four macro-categories. From a strictly numerical perspective, the most populated macro-category was the one including quantitative metrics (51 out of a total of 123 different metrics). We found that the most

commonly adopted and mentioned quality attributes belonged to the qualitative categories related to the relationship with the user, the conversation with the relational agent, and the user's perception of this conversation. This result suggests that the research in conversational agent evaluation currently lacks a wide adoption of quantitative methods.

Our findings answering RQ2.2 suggest a lack of structured, comparable, and standardized evaluation procedures since we could find only four different frameworks in the selected pool of sources. It is worth underlining that—in the studies that we analyzed—we did not find mentions of fully structured approaches to aid decision-making approaches in the design of conversational interfaces, based on either qualitative or quantitative attributes that have been measured. A recent and promising example in the literature is presented in a work by Radziwill et al., where the computation of quality attributes for chatbots is integrated into an Analytic Hierarchy Process (AHP). This process can be used to compare two or more versions of the same conversational system. These versions can either be a current available one (*as-is*) and one or more future ones in development (*to-be*) [1].

Finally, to answer RQ3, we compared the contribution of white and grey literature to the facets analyzed in the previous research questions. By considering the contribution to RQ2.1 and RQ2.3, we found that 27 out of 123 metrics (the 23%) and 5 out of 28 datasets (the 18%) for the evaluation of conversational agents are only mentioned in grey literature sources from our pool. These results underline how taking into account manuscripts that are not peer-reviewed provides an important added value for researchers when the research objective is to assess and evaluate conversational agents.

5.2. Threats to Validity

Threats to Construct validity for a Literature Review concern possible failures in the coverage of all the studies related to the review topic. In this study, we mitigated the threat by selecting five essential sources of white literature studies and including grey literature to consider chatbots, datasets, and evaluation metrics that are not presented in peer-reviewed papers.

For both typologies of literature, we applied a reproducible methodology based on established guidelines. To broaden the research as much as possible, we included the most commonly used terms in the search strings, as well as various synonyms. However, it is still possible that some terms describing other relevant works in the literature may have been overlooked. The definition and identification of the right keywords for the search string could be influenced. For this particular study, the missing unanimity about some concepts is defined (e.g., the same concepts are defined as quality attributes or metrics in different studies).

Regarding grey literature, there is a possibility that some relevant chatbots, frameworks, and evaluation procedures were not included in the analysis due to the inability to access the documents where they are presented.

Threats to Internal validity are related to the data extraction and synthesis phases of the Literature Review. All the primary sources resulting from the search strings application were read and evaluated by all authors and collaborators of this study to assess their quality, apply inclusion and exclusion criteria, and extract the information to answer the Research Questions of the study. Hence, the validity of the study is threatened by possible errors in the author's judgment when examining the sources. It may suffer from misinterpretations of the original content of the papers. This threat was mitigated by multiple readings of the same sources by the different researchers and discussions among the same people about the disagreements during the review phase.

Threats to External validity concern the generalizability of the findings of the Literature Review. We limited our investigation to textual chatbots or voice chatbots whose inputs can be reconducted to text for this study. It is not ensured that the found quality attributes can be generalized to any category of chatbots available in the literature.

Due to the differing accessibility of grey literature sources, it is also possible that this review provides only partial geographical coverage of commercial chatbots.

6. Conclusions and Future Work

In this study, we defined, conducted, and documented the results of a Multivocal Literature Review, i.e., an SLR conducted by taking into account different typologies of sources—not only peer-reviewed white literature. We applied this research methodology to the field of assessment and evaluation of conversational interfaces.

The principal goal of our review was to identify quality attributes that are used by either researchers or practitioners to evaluate and test conversational interfaces. We came up with 123 different quality attributes, four metric tools and frameworks for systematic and automated evaluations of conversational interfaces, and 28 datasets commonly using for performing evaluations. The quantity of information coming from grey literature only can be deemed a confirmation of the necessity of including grey literature in this very specific topic, since evaluation methods can be useful to researchers are often disseminated in non peer-reviewed sources.

The primary objective of this manuscript is to serve as a comprehensive reference for researchers and practitioners in the field of conversational agents development and testing, providing a categorization and a set of references to available quality attributes already defined in the literature. As future extensions of the present study, we plan to explore the possibility of developing automated or semi-automated tools to collect the measurable quality attributes for existing chatbots. We also plan to find strategies to prioritize the different quality attributes and implement a subset of them into a framework that can serve as a global means of evaluating any chatbot type. Finally, we aim to analyze multiple empirical measurements on a diverse set of chatbots, both commercial and academic, to seek correlations and dependencies between different metrics.

Funding: This work was partially funded by the H2020 EU-funded SIFIS-Home (GA #952652) project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable, the study does not report any data.

Acknowledgments: The authors thank Elona Merepeza and Isabeau Oliveri for their contribution to the paper management and assessment activities.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Radziwill, N.M.; Benton, M.C. Evaluating quality of chatbots and intelligent conversational agents. *arXiv* **2017**, arXiv:1704.04579.
2. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45.
3. Colby, K.M. Modeling a paranoid mind. *Behav. Brain Sci.* **1981**, *4*, 515–534.
4. Klopfenstein, L.C.; Delpriori, S.; Malatini, S.; Bogliolo, A. *The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms*; Association for Computing Machinery: New York, NY, USA, 2017.
5. Dale, R. The return of the chatbots. *Nat. Lang. Eng.* **2016**, *22*, 811–817, doi:10.1017/S1351324916000243.
6. Følstad, A.; Brandtzæg, P.B. Chatbots and the New World of HCI. *Interactions* **2017**, *24*, 38–42, doi:10.1145/3085558.
7. Shanhong, L. Chatbot Market Revenue Worldwide 2017 and 2024, 2019. Available online: https://www.statista.com/statistics/966893/worldwide-chatbot-market-value (accessed on 18 October 2021).
8. Brandtzaeg, P.B.; Følstad, A. Why People Use Chatbots. In *Internet Science*; Kompatsiaris, I., Cave, J., Satsiou, A., Carle, G., Passani, A., Kontopoulos, E., Diplaris, S., McMillan, D., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 377–392.
9. Müller, L.; Mattke, J.; Maier, C.; Weitzel, T.; Graser, H. Chatbot Acceptance: A Latent Profile Analysis on Individuals' Trust in Conversational Agents. In Proceedings of the SIGMIS-CPR'19: 2019 on Computers and People Research Conference, Nashville, TN, USA, 20–22 June 2019; pp. 35–42, doi:10.1145/3322385.3322392.
10. A Taxonomy of Social Cues for Conversational Agents. *Int. J. Hum. Comput. Stud.* **2019**, *132*, 138–161, doi:10.1016/j.ijhcs.2019.07.009.
11. Yu, Z.; Xu, Z.; Black, A.W.; Rudnicky, A. Chatbot evaluation and database expansion via crowdsourcing. In Proceedings of the chatbot workshop of LREC; International Conference on Language Resources and Evaluation, Portorož, Slovenia, 23–28 May 2016; Volume 63, p. 102.

12. Maroengsit, W.; Piyakulpinyo, T.; Phonyiam, K.; Pongnumkul, S.; Chaovalit, P.; Theeramunkong, T. A Survey on Evaluation Methods for Chatbots. In Proceedings of the 2019 7th International Conference on Information and Education Technology, Aizu-Wakamatsu, Japan, 29–31 March 2019; pp. 111–119.
13. Jokinen, K. Natural Language and Dialogue Interfaces. Journal of Human Factors and Ergonomics, 2009. Available online: http://www.ling.helsinki.fi/~kjokinen/Publ/200906UAIHandbookCh41_NaturalLanguage_Jokinen_Final.pdf (accessed on 18 October 2021).
14. Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P.N.; Inkpen, K.; et al. Guidelines for Human-AI Interaction. In Proceedings of the CHI'19: 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–13, doi:10.1145/3290605.3300233.
15. Nuruzzaman, M.; Hussain, O.K. A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks. In Proceedings of the 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), Xi'an, China, 12–14 October 2018; pp. 54–61, doi:10.1109/ICEBE.2018.00019.
16. Kocaballi, A.B.; Laranjo, L.; Coiera, E. Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires. In Proceedings of the HCI '18: 32nd International BCS Human Computer Interaction Conference, Belfast, UK, 4–6 July 2018; BCS Learning & Development Ltd.: Swindon, UK, 2018, doi:10.14236/ewic/HCI2018.21.
17. Jain, M.; Kumar, P.; Kota, R.; Patel, S.N. *Evaluating and Informing the Design of Chatbots*; Association for Computing Machinery: New York, NY, USA, 2018.
18. Hingston, P. A turing test for computer game bots. *IEEE Trans. Comput. Intell. AI Games* **2009**, *1*, 169–186.
19. Liu, D.; Bias, R.G.; Lease, M.; Kuipers, R. Crowdsourcing for usability testing. *Proc. Am. Soc. Inf. Sci. Technol.* **2012**, *49*, 1–10.
20. Tung, Y.H.; Tseng, S.S. A novel approach to collaborative testing in a crowdsourcing environment. *J. Syst. Softw.* **2013**, *86*, 2143–2153.
21. Ogawa, R.T.; Malen, B. Towards rigor in reviews of multivocal literatures: Applying the exploratory case study method. *Rev. Educ. Res.* **1991**, *61*, 265–286.
22. Higgins, J.P.; Thomas, J.; Chandler, J.; Cumpston, M.; Li, T.; Page, M.J.; Welch, V.A. *Cochrane Handbook for Systematic Reviews of Interventions*; John Wiley & Sons: Hoboken, NJ, USA, 2019.
23. Adams, R.J.; Smart, P.; Huff, A.S. Shades of grey: Guidelines for working with the grey literature in systematic reviews for management and organizational studies. *Int. J. Manag. Rev.* **2017**, *19*, 432–454.
24. Garousi, V.; Felderer, M.; Mäntylä, M.V. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Inf. Softw. Technol.* **2019**, *106*, 101–121.
25. Garousi, V.; Felderer, M.; Mäntylä, M.V. The Need for Multivocal Literature Reviews in Software Engineering: Complementing Systematic Literature Reviews with Grey Literature; In Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering, New York, NY, USA, 1–3 June 2016.
26. Garousi, V.; Mäntylä, M.V. When and what to automate in software testing? A multi-vocal literature review. *Inf. Softw. Technol.* **2016**, *76*, 92–117.
27. Garousi, V.; Felderer, M.; Hacaloğlu, T. Software test maturity assessment and test process improvement: A multivocal literature review. *Inf. Softw. Technol.* **2017**, *85*, 16–42.
28. Myrbakken, H.; Colomo-Palacios, R. DevSecOps: A multivocal literature review. In *International Conference on Software Process Improvement and Capability Determination*; Springer: New York, NY, USA, 2017; pp. 17–29.
29. Putta, A.; Paasivaara, M.; Lassenius, C. Benefits and Challenges of Adopting the Scaled Agile Framework (SAFe): Preliminary Results from a Multivocal Literature Review. In *Product-Focused Software Process Improvement*; Kuhrmann, M., Schneider, K., Pfahl, D., Amasaki, S., Ciolkowski, M., Hebig, R., Tell, P., Klünder, J., Küpper, S., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 334–351.
30. Tripathi, N.; Klotins, E.; Prikladnicki, R.; Oivo, M.; Pompermaier, L.B.; Kudakacheril, A.S.; Unterkalmsteiner, M.; Liukkunen, K.; Gorschek, T. An anatomy of requirements engineering in software startups using multi-vocal literature and case survey. *J. Syst. Softw.* **2018**, *146*, 130–151.
31. Tom, E.; Aurum, A.; Vidgen, R. An exploration of technical debt. *J. Syst. Softw.* **2013**, *86*, 1498–1516.
32. Ampatzoglou, A.; Ampatzoglou, A.; Chatzigeorgiou, A.; Avgeriou, P. The financial aspect of managing technical debt: A systematic literature review. *Inf. Softw. Technol.* **2015**, *64*, 52–73.
33. Ren, R.; Castro, J.W.; Acuña, S.T.; de Lara, J. Usability of Chatbots: A Systematic Mapping Study. In Proceedings of the 31st International Conference on Software Engineering and Knowledge Engineering, SEKE 2019, Hotel Tivoli, Lisbon, Portugal, 10–12 July 2019; Perkusich, A., Ed.; KSI Research Inc. and Knowledge Systems Institute Graduate School: Singapore, 2019; pp. 479–617.
34. Kitchenham, B.A.; Budgen, D.; Brereton, P. *Evidence-Based Software Engineering and Systematic Reviews*; CRC Press: Boca Raton, FL, USA, 2015; Volume 4.
35. Benzies, K.M.; Premji, S.; Hayden, K.A.; Serrett, K. State-of-the-evidence reviews: Advantages and challenges of including grey literature. *Worldviews Evid.-Based Nurs.* **2006**, *3*, 55–61.
36. Jalali, S.; Wohlin, C. Systematic literature studies: Database searches vs. backward snowballing. In Proceedings of the 2012 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, Lund, Sweden, 20–21 September 2012; IEEE Computer Society: Los Alamitos, CA, USA, 2012; pp. 29–38.
37. Corbin, J.M.; Strauss, A. Grounded theory research: Procedures, canons, and evaluative criteria. *Qual. Sociol.* **1990**, *13*, 3–21.

38. Khandkar, S.H. Open coding. *Univ. Calg.* **2009**, *23*, 2009.
39. Scott, C.; Medaugh, M. Axial Coding. *Int. Encycl. Commun. Res. Methods* **2017**, *1*, 1–2.
40. Cuayáhuitl, H.; Lee, D.; Ryu, S.; Cho, Y.; Choi, S.; Indurthi, S.; Yu, S.; Choi, H.; Hwang, I.; Kim, J. Ensemble-based deep reinforcement learning for chatbots. *Neurocomputing* **2019**, *366*, 118–130.
41. Nestorovič, T. Creating a general collaborative dialogue agent with lounge strategy feature. *Expert Syst. Appl.* **2012**, *39*, 1607–1625.
42. Campano, S.; Langlet, C.; Glas, N.; Clavel, C.; Pelachaud, C. *An ECA Expressing Appreciations*; IEEE Computer Society: Washington, DC, USA, 2015.
43. Glass, R.L.; Vessey, I.; Ramesh, V. Research in software engineering: An analysis of the literature. *Inf. Softw. Technol.* **2002**, *44*, 491–506.
44. Petersen, K.; Feldt, R.; Mujtaba, S.; Mattsson, M. *Systematic Mapping Studies in Software Engineering*; EASE'08; BCS Learning & Development Ltd.: Swindon, UK, 2008; pp. 68–77.
45. Ralph, P. Toward methodological guidelines for process theories and taxonomies in software engineering. *IEEE Trans. Softw. Eng.* **2018**, *45*, 712–735.
46. Chen, X.; Mi, J.; Jia, M.; Han, Y.; Zhou, M.; Wu, T.; Guan, D. *Chat with Smart Conversational Agents: How to Evaluate Chat Experience in Smart Home*; Association for Computing Machinery: New York, NY, USA, 2019.
47. Ly, K.H.; Ly, A.M.; Andersson, G. A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interv.* **2017**, *10*, 39–46.
48. Looije, R.; Neerincx, M.A.; Cnossen, F. Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *Int. J. Hum.-Comput. Stud.* **2010**, *68*, 386–397.
49. Kuligowska, K. Commercial chatbot: Performance evaluation, usability metrics and quality standards of embodied conversational agents. *Prof. Cent. Bus. Res.* **2015**, *2*, 1–16.
50. Earley, S. Chatbot Best Practices—Webinar Overflow Questions Answered. Available online: https://www.earley.com/blog/chatbot-best-practices-webinar-overflow-questions-answered (accessed on 23 September 2021).
51. Reese, H. Why Microsoft's 'Tay' AI Bot Went Wrong. Available online: https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/ (accessed on 23 September 2021).
52. Luthra, V.; Sethia, A.; Ghosh, S. Evolving Framework for Building Companionship Among Human and Assistive Systems. In *Human-Computer Interaction. Novel User Experiences*; Kurosu, M., Ed.; Springer International Publishing: Cham, Switzerland, 2016; pp. 138–147.
53. Lee, S.; Choi, J. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *Int. J. Hum. Comput. Stud.* **2017**, *103*, 95–105.
54. Reply. Chatbot in the Travel Industry | Reply Solutions. Available online: https://www.reply.com/en/travel-with-a-bot (accessed on 23 September 2021).
55. Abdulrahman, A.; Richards, D. Modelling Therapeutic Alliance Using a User-Aware Explainable Embodied Conversational Agent to Promote Treatment Adherence. In Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, Paris, France, 2–5 July 2019.
56. Götzer, J. Engineering and User Experience of Chatbots in the Context of Damage Recording for Insurance Companies. 2018. Available online: https://shorturl.at/yBEQZ (accessed on 18 October 2021).
57. Slesar, M. How to Design a Chatbot: Creating a Conversational Interface. Available online: https://onix-systems.com/blog/how-to-design-a-chatbot-creating-a-conversational-interface (accessed on 23 September 2021).
58. Linh, P.N. Want to Design a World-Class Customer Service Chatbot? Not without UX Testing! Available online: https://in.solvemate.com/blog/want-to-design-a-world-class-customer-service-chatbot-not-without-ux-testing (accessed on 23 September 2021).
59. Sanofi. Healthcare Chatbots. Available online: https://www.sanofi.fr/fr/-/media/Project/One-Sanofi-Web/Websites/Europe/Sanofi-FR/Newsroom/nos-publications/Livre-blanc-BOT-ENG-HD.pdf (accessed on 23 September 2021).
60. Xu, Z.; Sun, C.; Long, Y.; Liu, B.; Wang, B.; Wang, M.; Zhang, M.; Wang, X. Dynamic Working Memory for Context-Aware Response Generation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1419–1431.
61. Okanović, D.; Beck, S.; Merz, L.; Zorn, C.; Merino, L.; van Hoorn, A.; Beck, F. Can a Chatbot Support Software Engineers with Load Testing? Approach and Experiences. In Proceedings of the ACM/SPEC International Conference on Performance Engineering, Edmonton, AB, Canada, 20–24 April 2020.
62. Mimoun, M.S.B.; Poncin, I. A valued agent: How ECAs affect website customers' satisfaction and behaviors. *J. Retail. Consum. Serv.* **2015**, *26*, 70–82.
63. Chang, J.; He, R.; Xu, H.; Han, K.; Wang, L.; Li, X.; Dang, J. NVSRN: A Neural Variational Scaling Reasoning Network for Initiative Response Generation. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019; pp. 51–60.
64. Solutions, A. Chatbots: The Definitive Guide. Available online: https://www.artificial-solutions.com/chatbots (accessed on 23 September 2021).
65. Edwards, C.; Edwards, A.; Spence, P.R.; Shelton, A.K. Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Comput. Hum. Behav.* **2014**, *33*, 372–376.

66. Chalaguine, L.A.; Hunter, A.; Potts, H.; Hamilton, F. Impact of argument type and concerns in argumentation with a chatbot. In Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, USA, 4–6 November 2019; pp. 1557–1562.
67. Wang, D.; Jojic, N.; Brockett, C.; Nyberg, E. Steering Output Style and Topic in Neural Response Generation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017.
68. Inoue, M.; Matsuda, T.; Yokoyama, S. Web Resource Selection for Dialogue System Generating Natural Responses. In *HCI International 2011—Posters' Extended Abstracts*; Stephanidis, C., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 571–575.
69. Bosse, T.; Provoost, S. Integrating Conversation Trees and Cognitive Models Within an ECA for Aggression De-escalation Training. In *PRIMA 2015: Principles and Practice of Multi-Agent Systems*; Chen, Q., Torroni, P., Villata, S., Hsu, J., Omicini, A., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 650–659.
70. Pontier, M.; Siddiqui, G.; Hoorn, J.F. Speed Dating with an Affective Virtual Agent—Developing a Testbed for Emotion Models. In *Intelligent Virtual Agents*; Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 91–103.
71. Chatbot e-service and customer satisfaction regarding luxury brands. *J. Bus. Res.* **2020**, *117*, 587–595, doi:10.1016/j.jbusres.2018.10.004.
72. Arthur, R. Louis Vuitton Becomes Latest Luxury Brand to Launch a Chatbot. Available online: https://www.forbes.com/sites/rachelarthur/2017/12/08/louis-vuitton-becomes-latest-luxury-brand-to-launch-a-chatbot/#46b9941afe10 (accessed on 23 September 2021).
73. Jain, M.; Kota, R.; Kumar, P.; Patel, S.N. Convey: Exploring the Use of a Context View for Chatbots. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montréal, QC, Canada, 21–26 April, 2018.
74. Ali, M.R.; Crasta, D.; Jin, L.; Baretto, A.; Pachter, J.; Rogge, R.D.; Hoque, M.E. *LISSA—Live Interactive Social Skill Assistance*; IEEE Computer Society: Washington, DC, USA, 2015.
75. Google. User Engagement. Available online: https://developers.google.com/assistant/engagement (accessed on 23 September 2021).
76. Liu, C.W.; Lowe, R.; Serban, I.V.; Noseworthy, M.; Charlin, L.; Pineau, J. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; Volume 1, pp. 2122–2132.
77. Hwang, S.; Kim, B.; Lee, K. A Data-Driven Design Framework for Customer Service Chatbot. In *Design, User Experience, and Usability. Design Philosophy and Theory*; Marcus, A., Wang, W., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 222–236.
78. Valtolina, S.; Barricelli, B.R.; Di Gaetano, S. Communicability of traditional interfaces VS chatbots in healthcare and smart home domains. *Behav. Inf. Technol.* **2020**, *39*, 108–132.
79. Vasconcelos, M.; Candello, H.; Pinhanez, C.; dos Santos, T. Bottester: Testing Conversational Systems with Simulated Users. In Proceedings of the IHC 2017: XVI Brazilian Symposium on Human Factors in Computing Systems, Joinville, Brazil, 23–27 October 2017. doi:10.1145/3160504.3160584.
80. Amazon. Alexa Skills Kit, Alexa Skills. Available online: https://developer.amazon.com/it-IT/blogs/alexa/alexa-skills-kit (accessed on 23 September 2021).
81. Zhang, R.; Guo, J.; Fan, Y.; Lan, Y.; Xu, J.; Cheng, X. Learning to Control the Specificity in Neural Response Generation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018.
82. Benvie, A.; Eric Wayne, M.A. Watson Assistant Continuous Improvement Best Practices. 2019. Available online: https://www.ibm.com/downloads/cas/V0XQ0ZRE (accessed on 10/20/2021).
83. Amazon. Alexa Skills Kit Metrics API. Available online: https://developer.amazon.com/it-IT/docs/alexa/smapi/metrics-api.html (accessed on 23 September 2021).
84. Lowe, R.; Noseworthy, M.; Serban, I.V.; Angelard-Gontier, N.; Bengio, Y.; Pineau, J. Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. *arXiv* **2017**, arXiv:1708.07149.
85. Niewiadomski, R.; Demeure, V.; Pelachaud, C. Warmth, competence, believability and virtual agents. In *International Conference on Intelligent Virtual Agents*; Springer: New York, NY, USA, 2010; pp. 272–285.
86. Thaler, M.; Schlögl, S.; Groth, A. Agent vs. Avatar: Comparing Embodied Conversational Agents Concerning Characteristics of the Uncanny Valley. In Proceedings of the 2020 IEEE International Conference on Human-Machine Systems (ICHMS), Rome, Italy, 7–9 September 2020; pp. 1–6.
87. Herath, D.C.; Binks, N.; Grant, J.B. To Embody or Not: A Cross Human-Robot and Human-Computer Interaction (HRI/HCI) Study on the Efficacy of Physical Embodiment. In Proceedings of the 2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV), Shenzhen, China, 13–15 December 2020; pp. 848–853.
88. Grimes, G.M.; Schuetzler, R.M.; Giboney, J.S. Mental models and expectation violations in conversational AI interactions. *Decis. Support Syst.* **2021**, *144*, 113515.
89. Knidiri, H. How Artificial Intelligence Impacts the Customer Experience. 2021. Available online: https://matheo.uliege.be/bitstream/2268.2/13565/8/ISU_Template_with_Journal_Article_Format__ver_3_01_2021_%20%284%29.pdf (accessed on 18 October 2021).
90. Kim, W.; Lee, K. A Data-Driven Strategic Model of Common Sense in Machine Ethics of Cares. In *Human-Computer Interaction. Perspectives on Design*; Kurosu, M., Ed.; Springer International Publishing: Cham, Switzerland, 2019; pp. 319–329.

91. Iwase, K.; Gushima, K.; Nakajima, T. "Relationship Between Learning by Teaching with Teachable Chatbots and the Big 5. In Proceedings of the 2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech), Nara, Japan, 9–11 March 2021; pp. 191–194.
92. Vukovac, D.P.; Horvat, A.; Čižmešija, A. Usability and User Experience of a Chat Application with Integrated Educational Chatbot Functionalities. In *International Conference on Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 216–229.
93. Komori, M.; Fujimoto, Y.; Xu, J.; Tasaka, K.; Yanagihara, H.; Fujita, K. Experimental Study on Estimation of Opportune Moments for Proactive Voice Information Service Based on Activity Transition for People Living Alone. In *Human-Computer Interaction. Perspectives on Design*; Kurosu, M., Ed.; Springer International Publishing: Cham, Switzerland, 2019; pp. 527–539.
94. Pelau, C.; Dabija, D.C.; Ene, I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Comput. Hum. Behav.* **2021**, *122*, 106855.
95. Verstegen, C. The Pros and Cons of Chatbots. Available online: https://www.chatdesk.com/blog/pros-and-cons-of-chatbots (accessed on 23 September 2021).
96. Ishida, Y.; Chiba, R. Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design. *Procedia Comput. Sci.* **2017**, *112*, 2506–2518.
97. Ruane, E.; Farrell, S.; Ventresque, A. User Perception of Text-Based Chatbot Personality. In *International Workshop on Chatbot Research and Design*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 32–47.
98. Langevin, R.; Lordon, R.J.; Avrahami, T.; Cowan, B.R.; Hirsch, T.; Hsieh, G. Heuristic Evaluation of Conversational Agents. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–15.
99. Morgan, B. How Chatbots Improve Customer Experience in Every Industry: An Infograph. Available online: https://www.forbes.com/sites/blakemorgan/2017/06/08/how-chatbots-improve-customer-experience-in-every-industry-an-infograph/#2162528867df (accessed on 23 September 2021).
100. Max, D. The 13 Best AI Chatbots for Business in 2021 and Beyond [Review and Key Features]. Available online: https://www.netomi.com/best-ai-chatbot (accessed on 23 September 2021).
101. TechLabs, M. Your Go-To Chatbot Guide 101—All You Need to Know About Chatbots. Available online: https://marutitech.com/complete-guide-chatbots/ (accessed on 23 September 2021).
102. Hu, P.; Lu, Y.; others. Dual humanness and trust in conversational AI: A person-centered approach. *Comput. Hum. Behav.* **2021**, *119*, 106727.
103. Ameen, N.; Tarhini, A.; Reppel, A.; Anand, A. Customer experiences in the age of artificial intelligence. *Comput. Hum. Behav.* **2021**, *114*, 106548.
104. Raunio, K. Chatbot Anthropomorphism: Adoption and Acceptance in Customer Service. Master's Thesis, University of Twente, Enschede, The Netherlands, 2021.
105. Shin, D. How do people judge the credibility of algorithmic sources? 2021. Available online: https://philpapers.org/rec/SHIHDP-2 (accessed on 18 October 2021).
106. Ashfaq, M.; Yun, J.; Yu, S.; Loureiro, S.M.C. I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telemat. Inform.* **2020**, *54*, 101473.
107. Li, Y.; Arnold, J.; Yan, F.; Shi, W.; Yu, Z. LEGOEval: An Open-Source Toolkit for Dialogue System Evaluation via Crowdsourcing. *arXiv* **2021**, arXiv:2105.01992.
108. Wang, Z.; Wang, Z.; Long, Y.; Wang, J.; Xu, Z.; Wang, B. Enhancing generative conversational service agents with dialog history and external knowledge. *Comput. Speech Lang.* **2019**, *54*, 71–85.
109. Campos, J.; Paiva, A. A Personal Approach: The Persona Technique in a Companion's Design Lifecycle. In *Human-Computer Interaction—INTERACT 2011*; Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 73–90.
110. Dušan, O.; Samuel, B.; Lasse, M.; Christoph, Z.; Leonel, M.; André, v.H.; Fabian, B. Can a Chatbot Support Software Engineers with Load Testing? Approach and Experiences. 2020. Available online: https://www.vis.wiwi.uni-due.de/uploads/tx_itochairt3/publications/2020_ICPE_IndustryTrack_Chatbots.pdf (accessed on 18 October 2021).
111. Reeves, L.M.; Lai, J.; Larson, J.A.; Oviatt, S.; Balaji, T.; Buisine, S.; Collings, P.; Cohen, P.; Kraal, B.; Martin, J.C.; others. Guidelines for multimodal user interface design. *Commun. ACM* **2004**, *47*, 57–59.
112. Zhang, L. Exploration on Affect Sensing from Improvisational Interaction. In *Intelligent Virtual Agents*; Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 385–391.
113. Bührke, J.; Brendel, A.B.; Lichtenberg, S.; Greve, M.; Mirbabaie, M. Is Making Mistakes Human? On the Perception of Typing Errors in Chatbot Communication. In Proceedings of the 54th Hawaii International Conference on System Sciences, Kauai, HI, USA, 5 January 2021; p. 4456.
114. Krommyda, M.; Kantere, V. Improving the Quality of the Conversational Datasets through Extensive Semantic Analysis. In Proceedings of the 2019 IEEE International Conference on Conversational Data & Knowledge Engineering (CDKE), San Diego, CA, USA, 9–11 December 2019; pp. 1–8.
115. Hijjawi, M.; Bandar, Z.; Crockett, K. A general evaluation framework for text based conversational agent. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 23–33.

116. Crutzen, R.; Peters, G.J.Y.; Portugal, S.D.; Fisser, E.M.; Grolleman, J.J. An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: An exploratory study. *J. Adolesc. Health* **2011**, *48*, 514–519.
117. Guichard, J.; Ruane, E.; Smith, R.; Bean, D.; Ventresque, A. Assessing the robustness of conversational agents using paraphrases. In Proceedings of the 2019 IEEE International Conference On Artificial Intelligence Testing (AITest), Newark, CA, USA, 4–9 April 2019; pp. 55–62.
118. Jordan, P.; Albacete, P.; Katz, S. Exploring the effects of redundancy within a tutorial dialogue system: Restating students' responses. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Prague, Czech Republic, 2–4 September 2015; pp. 51–59.
119. Michelsen, J. Chatbots: Tip of the Intelligent Automation Iceberg. Available online: https://kristasoft.com/chatbots-tip-of-the-intelligent-automation-iceberg/ (accessed on 23 September 2021).
120. Microsoft. Bot Analytics. Available online: https://docs.microsoft.com/it-it/azure/bot-service/bot-service-manage-analytics?view=azure-bot-service-4.0 (accessed on 23 September 2021).
121. Ogara, S.O.; Koh, C.E.; Prybutok, V.R. Investigating factors affecting social presence and user satisfaction with mobile instant messaging. *Comput. Hum. Behav.* **2014**, *36*, 453–459.
122. Casas, J.; Tricot, M.O.; Abou Khaled, O.; Mugellini, E.; Cudré-Mauroux, P. Trends & Methods in Chatbot Evaluation. In Proceedings of the Companion Publication of the 2020 International Conference on Multimodal Interaction, Virtual, 25–29 October 2020; pp. 280–286.
123. Piao, M.; Kim, J.; Ryu, H.; Lee, H. Development and Usability Evaluation of a Healthy Lifestyle Coaching Chatbot Using a Habit Formation Model. *Healthc. Inform. Res.* **2020**, *26*, 255–264.
124. Mavridis, P.; Huang, O.; Qiu, S.; Gadiraju, U.; Bozzon, A. Chatterbox: Conversational interfaces for microtask crowdsourcing. In Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization, Larnaca, Cyprus, 9–12 June 2019; pp. 243–251.
125. Epstein, M.; Ramabhadran, B.; Balchandran, R. Improved language modeling for conversational applications using sentence quality. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; IEEE: New York, NY, USA, 2010; pp. 5378–5381.
126. Walker, M.; Langkilde, I.; Wright, J.; Gorin, A.; Litman, D. Learning to Predict Problematic Situations in a Spoken Dialogue System: Experiments with How May I Help You? In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, Seattle, WA, USA, 29 April–4 May 2000; pp. 210–217.
127. Shalaby, W.; Arantes, A.; GonzalezDiaz, T.; Gupta, C. Building chatbots from large scale domain-specific knowledge bases: challenges and opportunities. In Proceedings of the 2020 IEEE International Conference on Prognostics and Health Management (ICPHM), Detroit, MI, USA, 8–10 June 2020; pp. 1–8.
128. Teixeira, M.S.; da Costa Pereira, C.; Dragoni, M. Information Usefulness as a Strategy for Action Selection in Health Dialogues. In Proceedings of the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Melbourne, Australia, 14–17 December 2020; pp. 323–330.
129. Zhang, Y.; Song, D.; Li, X.; Zhang, P.; Wang, P.; Rong, L.; Yu, G.; Wang, B. A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Inf. Fusion* **2020**, *62*, 14–31.
130. Wang, W.; Huang, M.; Xu, X.S.; Shen, F.; Nie, L. Chat more: Deepening and widening the chatting topic via a deep model. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 255–264.
131. Wang, H.; Guo, B.; Wu, W.; Liu, S.; Yu, Z. Towards information-rich, logical dialogue systems with knowledge-enhanced neural models. *Neurocomputing* **2021**, *465*, 248–264.
132. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
133. Fergencs, T.; Meier, F.M. Engagement and Usability of Conversational Search–A Study of a Medical Resource Center Chatbot. In Proceedings of iConference 2021, Beijing, China, 17–31 March, 2021. Available online: https://vbn.aau.dk/en/publications/engagement-and-usability-of-conversational-search-a-study-of-a-me (accessed on 18 October 2021).
134. Karakostas, A.; Nikolaidis, E.; Demetriadis, S.; Vrochidis, S.; Kompatsiaris, I. colMOOC–an Innovative Conversational Agent Platform to Support MOOCs A Technical Evaluation. In Proceedings of the 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), Tartu, Estonia, 6–9 July 2020; pp. 16–18.
135. Firdaus, M.; Thangavelu, N.; Ekba, A.; Bhattacharyya, P. Persona aware Response Generation with Emotions. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
136. Beredo, J.; Ong, E. Beyond the Scene: A Comparative Analysis of Two Storytelling-based Conversational Agents. In Proceedings of the Asian CHI Symposium 2021, Yokohama, Japan, 8–13 May 2021; pp. 189–195. Available online: https://dl.acm.org/doi/abs/10.1145/3429360.3468208 (accessed on 18 October 2021).
137. Chug, P. 12 Experts Share The Biggest Chatbot Trends For 2020! Available online: https://botcore.ai/blog/12-experts-share-the-biggest-chatbot-trends-for-2020/ (accessed on 23 September 2021).
138. Bailey, D.; Almusharraf, N. Investigating the Effect of Chatbot-to-User Questions and Directives on Student Participation. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 85–90.

139. Schumaker, R.P.; Chen, H. Interaction analysis of the alice chatterbot: A two-study investigation of dialog and domain questioning. *IEEE Trans. Syst. Man Cybern. Part A Syst. Humans* **2009**, *40*, 40–51.
140. Ruane, E.; Faure, T.; Smith, R.; Bean, D.; Carson-Berndsen, J.; Ventresque, A. Botest: A framework to test the quality of conversational agents using divergent input examples. In Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, Tokyo, Japan, 7–11 March 2018; pp. 1–2. Available online: https://researchrepository.ucd.ie/handle/10197/9305?mode=full (accessed on 18 October 2021).
141. Miller, A.; Feng, W.; Batra, D.; Bordes, A.; Fisch, A.; Lu, J.; Parikh, D.; Weston, J. ParlAI: A Dialog Research Software Platform. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Association for Computational Linguistics: Copenhagen, Denmark, 9–11 September 2017.
142. Lowe, R.; Pow, N.; Serban, I.; Pineau, J. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue; Association for Computational Linguistics: Prague, Czech Republic, 2–4 September 2015.
143. Hori, C.; Perez, J.; Higashinaka, R.; Hori, T.; Boureau, Y.L.; Inaba, M.; Tsunomori, Y.; Takahashi, T.; Yoshino, K.; Kim, S. Overview of the sixth dialog system technology challenge: DSTC6. *Comput. Speech Lang.* **2019**, *55*, 1–25.
144. TheBotForge. How Much Does It Cost to Build a Chatbot in 2020? 2020. Available online: https://www.thebotforge.io/how-much-does-it-cost-to-build-a-chatbot-in-2020/ (accessed on 18 October 2021).
145. Banchs, R.E. On the construction of more human-like chatbots: Affect and emotion analysis of movie dialogue data. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC); IEEE: New York, NY, USA, 12–15 December 2017; pp. 1364–1367.
146. Mairesse, F.; Walker, M. PERSONAGE: Personality generation for dialogue. In Proceedings of the 45th annual meeting of the association of computational linguistics; Association for Computational Linguistics, Prague, Czech Republic, 23–30 June 2007; pp. 496–503.
147. Neff, M.; Wang, Y.; Abbott, R.; Walker, M. Evaluating the Effect of Gesture and Language on Personality Perception in Conversational Agents. In *Intelligent Virtual Agents*; Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 222–235.
148. Lison, P.; Tiedemann, J. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16); European Language Resources Association (ELRA), Portorož, Slovenia, 23– 28 May 2016.
149. Microsoft. Analyze Your Bot's Telemetry Data. Available online: https://docs.microsoft.com/en-us/azure/bot-service/bot-builder-telemetry-analytics-queries?view=azure-bot-service-4.0 (accessed on 23 September 2021).