Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-023-04626-5

ORIGINAL RESEARCH

Exploring agent-based chatbots: a systematic literature review

Davide Calvaresi[1] · Stefan Eggenschwiler[1] · Yazan Mualla[2] · Michael Schumacher[1] · Jean-Paul Calbimonte[1,3]

Received: 23 March 2022 / Accepted: 2 May 2023

Abstract
In the last decade, conversational agents have been developed and adopted in several application domains, including education, healthcare, finance, and tourism. Nevertheless, chatbots still need to address several limitations and challenges, especially regarding personalization, limited knowledge-sharing capabilities, multi-domain campaign support, real-time monitoring, or integration of chatbot communities. To cope with these limitations, many approaches based on multi-agent systems models and technologies have been proposed in the literature, opening new research directions in this context. To better understand the current panorama of the different chatbot technology solutions employing agent-based methods, this Systematic Literature Review investigates the different application domains, end-users, requirements, objectives, technology readiness levels, designs, strengths, limitations, and future challenges of the solutions found in this scope. The results of this review are intended to provide researchers, software engineers, and innovators with a complete overview of the current state of the art and a discussion of the open challenges.

Keywords Agent-based chatbot · Bots · Agents chatbots · Conversational agents

1 Introduction

Conversational agents have been proposed and designed to enable seamless interactions with people, through computer-based means for communication, language processing, interpretation, and dialogue exchange (Adamopoulou and Moussiades 2020). These agents have substantially evolved from its first incarnation, the seminal project ELIZA developed by Joseph Weizenbaum (Weizenbaum 1966). Ever since, conversational agents have leveraged on Natural Language Processing (NLP), state machine engines, and pattern matching with the intent of engaging in purposeful conversations with human users. Several milestones marked the technological evolution of conversational bots. Towards the end of the 1980 s, Rollo Carpenter developed Jabberwacky (Rollo 1997), a self-learning agent mainly employing contextual pattern matching to identify the best answer (accessible over

✉ Davide Calvaresi
davide.calvaresi@hevs.ch

Stefan Eggenschwiler
stefan.eggenschwiler@hevs.ch

Yazan Mualla
yazan.mualla@utbm.fr

Michael Schumacher
michael.schumacher@hevs.ch

Jean-Paul Calbimonte
jean-paul.calbimonte@hevs.ch

[1] University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland

[2] Université de Technologie de Belfort Montbéliard, Sevenan, France

[3] The Sense Innovation and Research Center, Sion, Switzerland

Published online: 01 June 2023

the internet only later in 1997). Later, in 1994, the term *ChatterBot* made its first appearance, used by Michael Mauldin to describe conversational programs (Mauldin 1994). Nowadays, this term has been shortened to *chatbot*, and it is used on a daily basis to describe these technologies. In the 1990 s, considerable progress was made on conversational agent technologies, based on advances in Artificial Intelligence. For example, Richard Wallace developed ALICE (Artificial Linguistic Internet Computer Entity), which leveraged on heuristical pattern matching.[1]

In the 2010 s, chatbot technologies started to gain adoption, outside the academic sphere, in industrial and mainstream applications. Apple was among the first to commercialize a personal assistant with conversational capabilities in 2011 with the release of Siri.[2] Initially based on the Active platform (Guzzoni 2008), it assisted iPhone users recognizing both written and spoken language. Other major technology companies released their virtual assistants shortly. Google Now for Android and iOS devices appeared in 2012,[3] evolving from a simple recommendation engine to a personal assistant able to dialogue with the user (similar to Siri).[4] Microsoft followed with Cortana,[5] which was released in 2014. The same year, Alexa was launched by Amazon, primarily targeting home automation and online shopping. Although it is not linked to any OS, it quickly gained adoption in the market (Etherington 2014). The widespread acceptance of these major companies' virtual assistants and their usage of asynchronous text-based interactions stimulated instant messenger applications to release APIs for third-party development of chatbots (i.e., Telegram, Facebook Messenger, and WhatsApp), in addition to those mainly dedicated to customer services through their web pages.

The increasing adoption of chatbots has been boosted by anywhere/anytime availability, immediate response, confidentiality, social acceptance, and massive scalability. Leveraging on these aspects, chatbots have proven to be effective in a wide range of domains such as eCommerce (Cui et al. 2017), education (Winkler and Söllner 2018), and in particular for motivational (e.g., social network campaigns (Calvaresi et al. 2019)) and support (e.g., customer management (Xu et al. 2017), eHealth (Calbimonte et al. 2019), and assisted-living scenarios (Fadhil and Gabrielli 2017)).

Recent remarkable technological advancements are pushing the evolution of chatbots using keyword-based text recognition or static finite state machines (FSM) to interpret and orchestrate user interactions (today still representing a significant share of the market), to hybrid solutions merging NLP (for text recognition) and FSM (for the management of intentions and user stories) (DeepLink 2022). However, the solely FSM-based solutions still expose significant limitations, such as inadequate personalization, lack of real-time monitoring, reporting and customization, lack of mechanisms to integrate communities of chatbots, limited knowledge sharing capabilities, and the impossibility of deploying multi-domain campaigns within the same framework. These limitations are linked to the predominantly rigid architectures proposed in most existing approaches. These rely on very specific scenarios translated into chatbot logic, which have to be reprogrammed every time a new scenario arrives. This raises the costs of modifying the behavior of a chatbot and prevents administrators from adapting it to specific situations. Moreover, most chatbot solutions rely on monolithic and centralized data management strategies, making it hard to comply with privacy regulations (e.g., European Union's General Data Protection Regulation – GDPR (Voigt and Von dem Bussche 2017)). The sensitive nature of data collected through chatbot interactions makes it necessary to shift the control of personal data towards the users themselves, empowering them in the process. Many chatbot systems have used AI to boost the accuracy and user-experience of its interactions. Examples include the use of NLP to generate asynchronous follow-up questions (Rao et al. 2021), or the application of neural networks to perform emotion detection in chatbot conversations (Huddar et al. 2021). However, these AI techniques focus more on the generation of responses and monitoring conversational context, without considering the autonomous, decentralized and collaborative nature of chatbots.

Nevertheless, in the last decade, the trend of combining chatbots with multi-agent systems (MAS) models and technologies tried to mitigate the limitations mentioned above. Particular emphasis is given to application domains where the social and collaborative dimensions (e.g., crowdsourcing, user profiling and personalization) is essential in the interaction with users. These features are particularly relevant for domains such as healthcare fostering behavioral change (Pereira and Díaz 2019), where the majority of the studies/contributions bridging chatbots and MAS can be found (Calbimonte et al. 2019; Calvaresi et al. 2019).

To better understand the current panorama of the different chatbot technology solutions employing agent-based approaches, this work presents a Systematic Literature Review (SLR) investigating application domains, end-users, requirements, objectives, technology readiness level (TRL) (European Commission 2017), designs, strengths, limitations, and future challenges of the solutions found in the literature. The goal is to provide a tool for researchers, software engineers, innovation managers, and other

[1] Program D: 2012 version of ALICE: https://github.com/noelbush-xx/programd.

[2] https://www.apple.com/siri/.

[3] https://developers.google.com/events/io/2012.

[4] Now called Google Assistant: https://assistant.google.com/.

[5] https://www.microsoft.com/en-us/cortana.

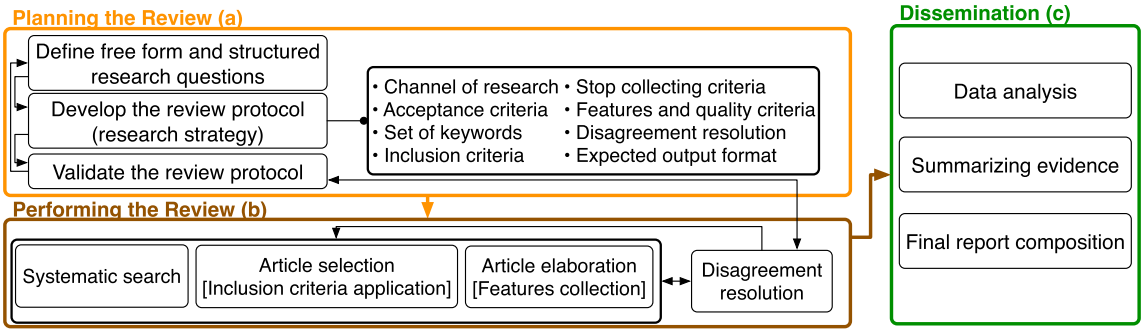


Fig. 1 Systematic literature review phases (Kitchenham et al. 2009)

practitioners to investigate the current state of the art and discuss the open challenges.

The rest of the paper is structured as follows: Sect. 2 presents the methodology applied for performing the SLR. Section 3 presents the review planning phase, including the definition of the protocol and the research questions. Section 4 describes how the review was performed. Section 5 analyses the outcomes of the applied methodology structured according to the research questions. Section 6 discusses the obtained results, projecting them into the stated (by the primary studies) and envisioned (by the authors of this paper) future directions. Finally, Sect. 7 concludes the paper.

2 Systematic literature review methodology

The approach employed in this paper aims at being both rigorous and reproducible. It relies on the methodology outlined by Kitchenham (Kitchenham et al. 2009), which has also been employed in a similar contexts (Palmarini et al. 2018; Calvaresi et al. 2021b; Anjomshoae et al. 2019; Mualla et al. 2019; Calvaresi et al. 2018). Figure 1 proposes a schematic representation of the adopted procedure. In particular, it comprises three stages:

P1: Planning the review. This phase consists of defining the main generic question(s) and deriving *Structured Research Questions*, characterizing the entire search protocol, matching the requirements (*rigorousness and reproducibility*), and validating the protocol.

P2: Performing the review. Entails the execution of the following planned activities: collection and selection of literature, literature elaboration, and disagreement resolution.

P3: Dissemination. Includes analysis, documentation, reporting, and summary of the learned lessons.

3 Review planning

This section describes the definition of the *structured research questions* and the *development of the review protocol* describing the *search strategy*, the *inclusion and exclusion criteria*, the *biases and disagreement resolution*, and the *quality criteria*.

3.1 Research questions

As introduced in Sect. 1, the research community has proposed the usage of multi-agent-based chatbots in recent years, for different domains, stakeholders, and purposes. Therefore, the main research question can be contextualized in these terms as follows: *How are agent-based chatbots characterized, envisioned, and employed?* To better investigate such a question, we comply with the Goal-Question-Metric (GQM) approach introduced by (Galster et al. 2014; Kitchenham et al. 2010). Such an approach has been employed in several other studies in the computer science-related domain (e.g., augmented reality for maintenance (Palmarini et al. 2018), virtual reality for education (Radianti et al. 2020), explainable agents and robots (Anjomshoae et al. 2019), agents and blockchains (Calvaresi et al. 2018)) and other domains (e.g., tourism (Yang et al. 2017; Calvaresi et al. 2021b). The dimensions targeted in this study apply to "intelligent" technologies and research. In particular, they are *scientific interest over the years, application domains, stakeholders, requirements, goals, technologies, advantages, limitations, countermeasures, and future research*. By formulating questions addressing such aspects, provide investigations and analysis in support of practitioners (providing an aggregated understanding of the current works), new tech pioneers (understanding what has been tried and

what might be future targets), and industrial researchers (to bring research ideas onto the real-world market). Thus, we devised a set of ten structured research questions.

SRQ1 To establish an understanding of the demographic evolution of agent-based chatbots, we inquire: *How are the research efforts temporally and geographically distributed?*

SRQ2 To elicit the domains on which the agent-based chatbots research focuses, we inquire: *Which application domains have employed agent-based chatbots?*

SRQ3 To clarify who are the stakeholders of agent-based chatbots, we inquire: *Who are the users of the chatbot systems relying on the agent paradigm?*

SRQ4 To formalize the requirements arranged w.r.t. the given stakeholders, we inquire: *What are the requirements standing behind the employment of agent-based chatbots?*

SRQ5 To explore what research tried to achieve with agent-based chatbots, we inquire: *What are the objectives set for agent-based chatbots?*

SRQ6 To better understand the technological characterization, we have structured SRQ6 in four sub-questions:

a) *Which chatbot* ***design*** *(e.g., paradigms) and* ***implementations*** *have been proposed?*

b) *Which technologies have been employed in the proposed solutions?*

c) *Which technologies have been previously employed?*

d) *What is the* ***Technology Readiness Level*** (European Commission 2017) *of the solutions proposed in the primary studies?*

SRQ7 To explore the benefit of existing solutions, we inquire: *What are the* **strengths** *of employing agent-based chatbots?*

SRQ8 To identify the shortcomings of the existing solutions, we inquire: *What are the* **limitations** *of employing agent-based chatbots?*

SRQ9 To understand the measures employed by the authors to achieve their objectives and overcome the limitations, we inquire: *What are the proposed* **solutions** *for the limitations identified in SQR8?*

SRQ10 Finally, to foster the establishment of future objectives, we inquire: *What are the future challenges for chatbot-based solutions envisioned by the primary studies?*

3.2 Review protocol

The *search strategy* included the selection of the following information sources: IEEE Xplore,[6] ScienceDirect,[7] ACM Digital Library,[8] Citeseerx,[9] and Pubmed.[10] The selection of the keywords relied on the reviewers' background and knowledge in the context of agent-based chatbots, and they include the following: *Multi-agent system, MAS, agent-based, chatbot, conversational agent, virtual assistant, personal assistant.* To increase the results' accuracy, some keywords have been aggregated. For example, *MAS* was expanded to three different queries: *(i)* MAS + chatbot + virtual assistant, *(ii)* MAS + chatbot + personal assistant, and *(iii)* MAS + chatbot + conversational agent.

Each search query produced a set of articles added to the list of papers to be considered. The result of each query has been screened by the reviewers to evaluate the articles' coherence with the study. In particular, title and abstract have been pre-processed according to the criteria presented in the next section.

3.2.1 Inclusion and exclusion criteria

The initial search collected **108** papers, hereafter referred to as *primary studies*. Additional filtering criteria have been applied (see Table 1). In particular, such criteria have been selected aiming at (i) avoiding multiple papers (usually incremental) describing the same work, (ii) bounding the time window for the investigation (e.g., excluding too old and less-relevant works, given the technological advancements), (iii) selecting works contributing to the actual investigated topic, and (iv) ensuring the presence of a tangible theoretical/practical contribution – avoiding purely visionary

[6] http://ieeexplore.ieee.org.

[7] http://www.sciencedirect.com/.

[8] http://dl.acm.org/.

[9] http://citeseerx.ist.psu.edu/index.

[10] http://www.ncbi.nlm.nih.gov/pubmed.

Table 1 Inclusion and exclusion criteria

Criteria	Description
Temporal	Most major chat-based companies released their APIs between 2002 and 2020. To provide a fair selection and include recent publications, the mean was set to 2015 with a deviation of up to 6 years, i.e., a publication window from 2009 to 2021
Originality	Papers presenting minor variations or duplicated papers should be discarded
Purpose	The purpose of the primary study has to refer to asynchronous human-bot communications utilizing an agent-based approach
Relevance	The primary study should define its contributions in the context of agent-based chatbots
Primary Study	Reviews focusing on applying agent-based chatbots are excluded from the analysis but have to be gathered and considered separately
Theoretical foundation	The primary study should provide at least one of the following elements: innovative formulation, theoretical definition, system design
Practical contribution	The primary study should provide at least one of the following elements: practical implementation, tests, critical analysis, evaluations or discussion

Table 2 Summary of the inclusion/exclusion phase of the collected papers

Rev	Conflict solver	Y-Y	N-N	Conflicts	Accepted out of conflicts	Total	Accepted	Acceptance %
$\mathcal{A}\Leftrightarrow\mathcal{B}$	$\mathcal{C}$	14	15	11	2	40	16	40.0%
$\mathcal{A}\Leftrightarrow\mathcal{C}$	$\mathcal{B}$	11	19	5	2	35	13	37.1%
$\mathcal{B}\Leftrightarrow\mathcal{C}$	$\mathcal{A}$	7	16	10	2	33	9	27.3%
Sum		32	50	26	6	108	38	21.1%

and blue sky studies). The criteria definition is usually quite specific per topic/review. Nevertheless, several studies recall similar criteria selections (Yang et al. 2017; Anjomshoae et al. 2019). Applying the criteria defined in Table 1, we purged unrelated papers and narrowed them down to a set of **38** contributions. Three reviewers were instructed to verify the compliance of the papers with the aforementioned inclusion criteria. Each reviewer operated independently while filtering out the list of papers. After the filter process ended, a review process was established so that a paper was included if at least two reviewers agreed on it.

3.2.2 Biases and disagreement resolution policy

The policy for biases and disagreement resolution allows the reviewers to cross-examine each task to limit biases and resolve disagreements among themselves. In particular, during the articles selection task, three reviewers cross-validated the inclusion/exclusion. During the elaboration of the articles, uncertainties have been discussed during periodic meetings.

3.2.3 Features and quality criteria

Assessing the quality of the extracted information is crucial. The following set of features has been chosen to answer the structured research questions: Publication year, geographical localization, main purpose, context, kind of users involved, scenarios, level of abstraction†, architectures and designs, development methodologies, techniques, technologies and devices, user needs coverage‡, need - offered support relation, kind of disease or difficulties supported‡, awareness provided, architectural evidence‡, technological evidence‡, technical evidence‡, architectural limitations‡, technological limitations‡, technical limitations‡, identified future directions, identified future challenges. The features annotated with (†) are classified with C, P, or T as possible values, that respectively stand for C = Conceptual; P = Prototype Architectures and Frameworks, no results are provided; T = Tested Architectures and Frameworks, results are provided. The features annotated with (‡) are associated to Y, P, or N values, that stand for: Y = information are explicitly defined / evaluated; P = information are implicit / stated; N = information are not inferable. Such a categorization of the collected features has been performed according to the DARE criteria, elaborated and proposed by (Kitchenham et al. 2009).

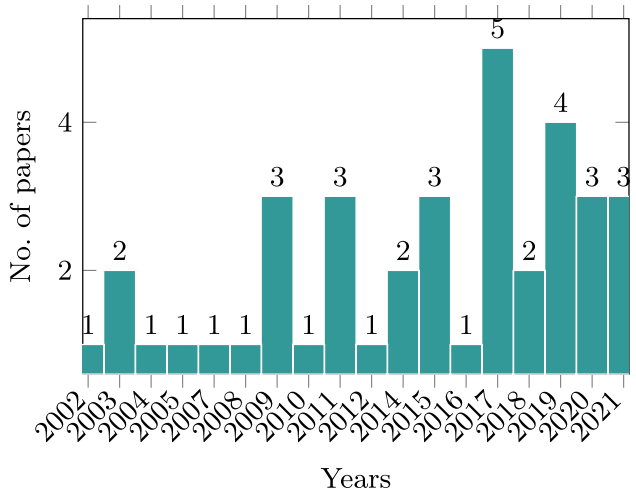


Fig. 2 Total papers per year

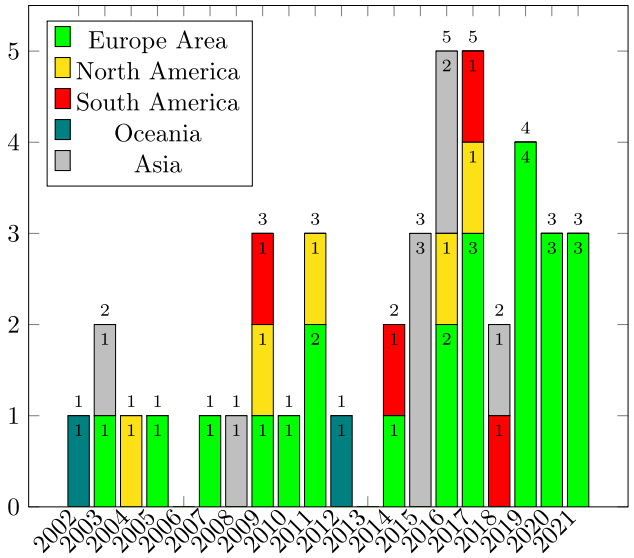


Fig. 3 Number of papers per continent per year

4 Review execution

This section details the *Perform Review* task in Fig. 1. In particular, it elaborates on the review's execution, including details on the article collection, selection, and elaboration. The semi-automatic search presented in Sect. 2 resulted in a total of 108 selected articles. The assessment of the primary studies to be finally included in the elaboration phase has been conducted by a total of three reviewers. In particular, the articles have been organized into three equally distributed groups, each of them elaborated by two reviewers (in rotation) with the third one involved in the case of conflict.[11] Table 2 details the selection assessments, referring to the reviewers with the letters $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$.

The papers have been listed following the collection order and respecting the relevance-based sorting obtained when querying the scientific web collectors. It is possible to notice that the third set of papers recorded a drastic reduction in the acceptance rate. Such a piece of information offers two possible reading keys: *(i)* the stop criteria has acted too loosely and/or *(ii)* title & abstract do not mirror the papers' content properly.

The filtering phase concludes by selecting 38 papers to be elaborated out of the 108 initially collected (21.1% total acceptance rate). In turn, the features presented in section 3.2 have been extracted and collected in a tabular format to facilitate their elaboration and the understanding of possible correlations to be discussed. Nevertheless, in some cases, the extraction of relevant information has been challenging

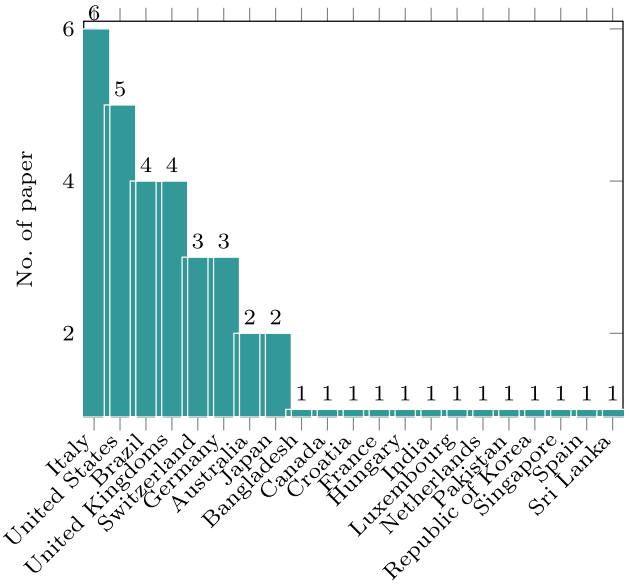


Fig. 4 Number of papers per country

due to the lack of explicit statements (e.g., very few studies have clearly mentioned the limitations of their approaches). To cope with this situation, the reviewers have leveraged their knowledge of the topic to produce a more comprehensive understanding and propose to the reader additional information (rigorously decoupled with the presentation of the results and solely addressed in the discussion).

[11] Possible evaluations: Y = Yes, N = No, X = Not sure. If both reviewers agreed on the assessment, no further review was required. However, if a conflict occurred (e.g., Y-X, X-X, X-Y, N-X, X-N), the third reviewer was consulted.

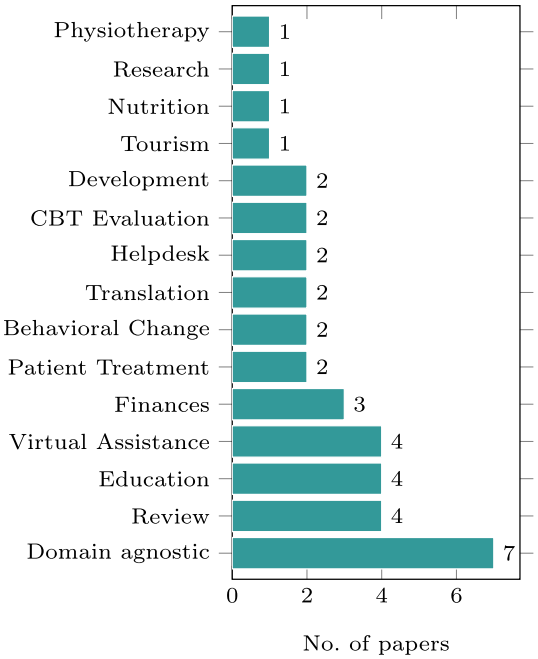


Fig. 5 Contributions per application domain

5 Review results and analysis

In the following, we structure the results of the SLR according to the research questions defined in Sect. 3.1.

5.1 Demographics

Referring to question **SRQ1**, Figs. 2 and 4 show the temporal and geographical distribution of papers targeting agent-based chatbots. Figure 2 reports the primary studies' distribution over the time-window selected for this study. A slight upward trend can be observed in recent years. Nevertheless, the research field of multi-agent-based chatbots still seems to be a niche area. Indeed, looking at Fig. 4, the geographical localization of the first authors' institutions (organized per country) relates to the distribution of research groups in the field of multi-agent systems (i.e., centered in the US and Europe). Finally, Fig. 3 provides a further view on the selected primary studies by grouping the papers per continent.

5.2 Application domains

Regarding **SRQ2**, we graphically represented in Fig. 5 the selected application domains of the primary studies. It is noticeable that the panorama of the application domains is remarkably broad and diversified. For example, it ranges from education (Alencar and Netto 2014) to healthcare (Kökciyan et al. 2021) and financing (de Bayser et al. 2018). Nevertheless, it appears that personalized assistive purposes have attracted most efforts across domains.

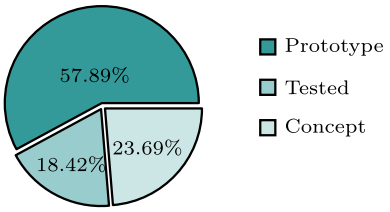


Fig. 6 Type of studies

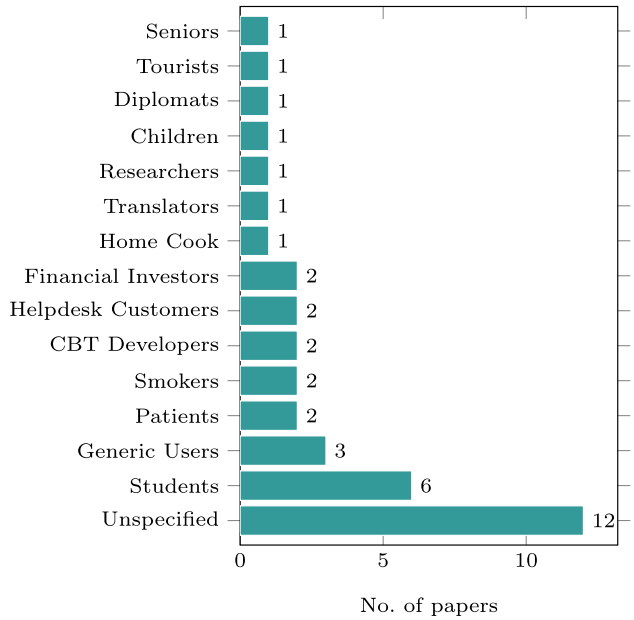


Fig. 7 Number of papers per type of users

5.3 Intended user classes

Concerning **SRQ3**, Fig. 7 shows the distribution of the diverse intended user classes identified by the selected primary studies, which is a direct consequence of the application domains. On the one hand, it is evident that most of the literature operates in the context of education, having either students, tutors, or professors as the main users. On the other hand, although being a minority, a considerable amount of studies is solely conceptual or general (see Fig. 6) and does not tackle a specific use case. Overall, the majority (57.89%) of the primary studies presented some form of prototypes, 23.69% deal with technical or scientific concepts, and 18.42% of the selected papers contains extensively tested artifacts.

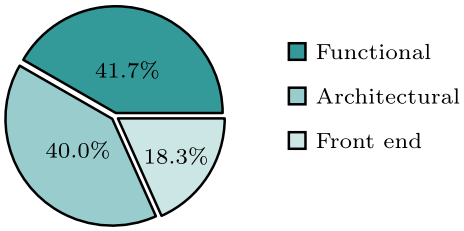


Fig. 8 Type of requirements

5.4 Requirements

Concerning question **SRQ4**, we elicited the requirements expressed by the primary studies. We can see the evolution of the main features captured by these requirements in Fig. 9. We categorized the requirements as follows:

- **Functional Requirements:** requirements affecting the behavior of the platforms (see Table 3);

Table 3 Functional requirements

Study	Functional Requirements
(Hettige and K. 2015)	Multi-language chatbot. Semantic-context inference. Knowledge ontology and agent rules updated based on user input
(Hung et al. 2009)	Evaluate qualitative and quantitative aspects. Quantify aspects like "naturalness" and "friendliness"
(de Bayser et al. 2017)	Allow multiple users/bots to communicate with each other. Support multiple roles with different behaviors & interaction norms (trigger, pre-conditions, behaviors)
(Vasconcelos et al. 2017)	Automate user interaction simulation. Analyze metrics, e.g., answer accuracy, frequency, response time. Provide a standardized unit test format (input, expected output)
(de M. Batista et al. 2009)	Answer FAQ-like questions regarding Java development
(Kalia et al. 2017)	Identify human roles that can be automated, their goals and their commitments. Generate an interaction set translated into an IBM Watson Model
(Angara et al. 2017)	Specify diet requirements, food preferences, goals, medical condition. Check available ingredients and provide recipe recommendations
(Z. et al. 2016)	Record user-chatbot interactions for data analysis purposes
(Memon et al. 2018)	Allow chatbots to communicate with each other
(de Bayser et al. 2018)	Allow multiple users and bots to communicate with each other
(Alencar and Netto 2014)	Enable distance learning tutoring for student activities, communicate deadlines, answer questions regarding course material. Dynamically update chatbot knowledge through tutor input
(Pilato et al. 2007)	Expand knowledge base through user inquiries, mapping, and linking with concepts in the associative area. Combine Latent Semantic Analysis (LSA), common sense & traditional knowledge representation. Provide greater expressiveness over traditional rule-based systems
(Augello et al. 2011)	Adapt to specific dialogue requirements through specific knowledge modules
(Tarau and Figa 2004)	Cover interactive story telling, online teaching, user support
(Wong et al. 2012)	Act as a virtual child companion providing structured activities and free-flow dialogue with unpredictability for children
(Noori et al. 2014)	Provide training for junior diplomats on consular activities. Catch expert knowledge and update information if necessary
(Huang et al. 2008)	Generic Embodied Conversational Agent framework (GECA) should be usable for rapid system prototyping and ease of Embodied Conversational Agent (ECA) development
(Bentivoglio et al. 2010)	Implement a modular Learning Management System with reusable components. Implement a pedagogic-didactic paradigm for self-regulated learning
(Calvaresi et al. 2019)	Support behavioral change campaigns (such as smoking cessation)
(Thosani et al. 2020)	Store user input (profiling) to reduce necessary future user input
(Calbimonte et al. 2019)	Support profile learning, persuasive argumentation, and symmetric- & asymmetric communication
(Kökciyan et al. 2021)	Help stroke patients in self-managing their condition and to adhere to treatment plans, in collaboration with healthcare professionals. Collect and process data from external devices such as heart rate, etc
(Chapman et al. 2019)	Integrate data from multiple sources, including commercial wellness sensors, EHR, to produce an adaptive care plan. Apply computational argumentation & provenance to track data from disparate sources, and to identify reinforcing and conflicting data
(Calvaresi et al. 2021a)	Possess reactive and proactive behaviors, persuasive, learning, and profiling capabilities
(Tatai et al. 2003)	Provide high quality NLP. Manipulate user emotion based on Plutchik's emotional model

Table 4 Architectural requirements

Studies	Architectural Requirements
(de Bayser et al. 2017)	Communication structured around 4 dimensions: What is the message about? Who should reply to the message? How should the reply look like? When should the reply be sent?
(de M. Batista et al. 2009)	Generic knowledge base architecture based on pattern-matching (i.e., AIML)
(Jiang et al. 2015)	Support multi-domain contexts, provide integration to different knowledge bases (e.g., Frame-based, AIML, SQL, RDF, Rule-based). Provide integration with different NLP possibilities (e.g., speech-recognition/speech synthesis, NL understanding/generation)
(Z. et al. 2016)	Integrate diverse chatbot back ends into one unified system. Provide access to external API resources (e.g. weather data, yelp)
(Memon et al. 2018)	Provide rule-based NLP techniques for communication
(de Bayser et al. 2018)	Provide mediator and expert bots. Implement a deontic approach and follow the so-called norms (e.g., obligations, permissions and prohibitions)
(Alencar and Netto 2014)	Integration in the Moodle platform
(Agostaro et al. 2005)	Provide foundation chatbots based on the LSA-framework
(Pilato et al. 2007)	Split the knowledge base (KB) into a rational area (structured, rule-based KB) and an associative area (data-driven semantics space)
(Pilato et al. 2011)	Provide modularity to adapt to different KB domains. Implement a component (i.e., corpus callosum) that chooses and triggers the most adequate KB section
(Augello et al. 2009)	Use associative space to allow the better exploitation of semi-structured data and increase the dialogue capabilities
(Augello et al. 2011)	Development as web-platform for chatbots with modular knowledge bases
(Tarau and Figa 2004)	Support both semantic and lexical knowledge bases
(Noori et al. 2014)	Incorporate a pattern matching the algorithm to create a goal orientated Arabic bot
(Huang et al. 2008)	Act as low-level layer between different system components (NLG, Language generation, 3D module, question analyzer)
(Bentivoglio et al. 2010)	Split different domains by domain-expert agents
(Calvaresi et al. 2019)	Provide user profiling capabilities and the possibility to share knowledge between user agents
(Zolitschka 2020)	Orchestration framework should be MAS-based, reusable, scalable, and provide topic-specific agents
(Calbimonte et al. 2019)	System scalability
(Shashaj et al. 2019)	Robust, powerful, and flexible enough to easily adjust to any business context
(Calvaresi et al. 2021a)	Store personal data in a GDPR-compliant and consent-based database where the user has control
(Maher and Gu 2002)	User-centered virtual architecture should change the current state of designing virtual architectures
(Mori et al. 2003)	Provide proactive and reactive behaviors

Table 5 Front end requirements

Studies	Front end Requirements
(Vasconcelos et al. 2017)	Provide an intuitive web user interface (UI) to execute test scenarios, display test outcomes and analyze results
(Jiang et al. 2015)	Provide a cross-platform UI for the user (i.e., web, mobile, desktop)
(Z. et al. 2016)	Unified web-based front end for text or speech interface, 2D/3D avatars
(Alencar and Netto 2014)	Interaction through an agent embodied as a 3D animated avatar. 3D animation to convey meaning through body movement
(Tarau and Figa 2004)	Provide a voice-enabled web-based chat interface
(Wong et al. 2012)	UI in a 3D embodied agent with speech input and output
(Calvaresi et al. 2019)	The bot should be integrated in Facebook Messenger
(Shashaj et al. 2019)	Provide a web interface for design, development, test and deployment. Provide a separate web interface for maintenance and monitoring
(Kökciyan et al. 2021)	Interaction through a tablet app as a dashboard or a bot
(Chapman et al. 2019)	Interaction with the system via argumentation-based dialogue

- **Architectural Requirements:** requirements stirring the system or the back end of the platforms (see Table 4);
- **Front end Requirements:** requirements applied to the front end of the platforms (see Table 5);

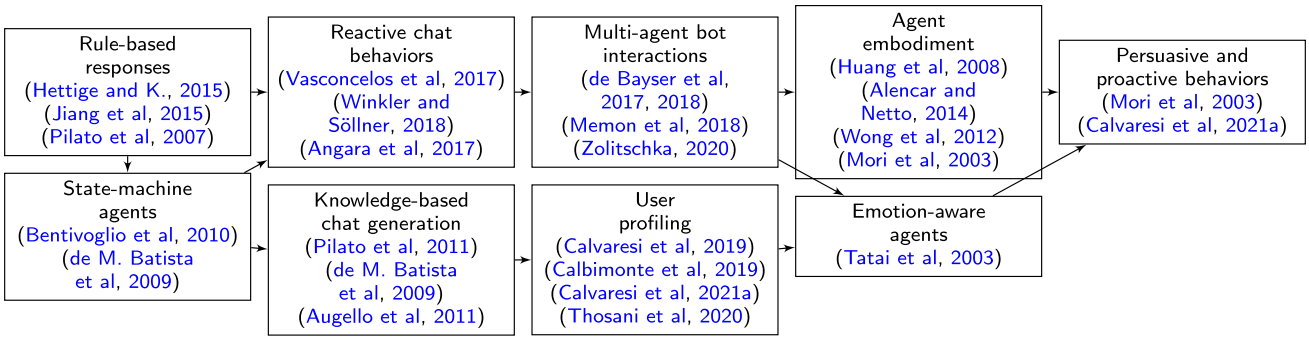


Fig. 9 Evolution of features in agent-based chatbots according to the requirements

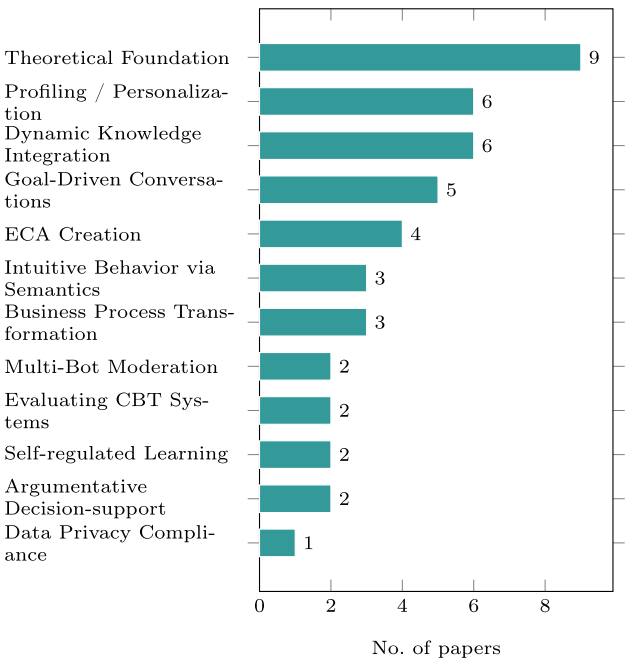


Fig. 10 Primary studies' objectives

Figure 8 depicts the distribution of types of requirements characterizing the primary studies. The authors of the elaborated papers focus primary on functional (41.7%) and architectural (40.0%) requirements. Requirements concerning the front end were only explicitly formalized in 18.3% of the studies.

5.5 Objectives of the studies

Investigating **SRQ5**, we collected and clustered the objectives of the primary studies as depicted in Fig. 10. Most of the papers tackle the theoretical foundations of MAS-based chatbots (i.e., nine studies focus primarily on conceptual aspects of the current state of the art or non-concrete systems). Among them, we can mention (Augello et al. 2017), where a notion of "social intelligence" for chatbots is defined, and linked to current technologies' capability to develop social chatbots. Also, (Hung et al. 2009) defines a method for an evaluation process to assess the "naturalness" of a chatbot system.

Concerning more practical studies, goal-driven behaviors (e.g., intended to tackle user personalization) have been studied for dietary and entertainment proposes. (Angara et al. 2017) describes a chatbot designed to support users in their kitchen by providing recipe recommendations while adhering to their dietary goals, medical conditions, preferences, and available ingredients. Similarly, (Wong et al. 2012) describe a goal-oriented virtual chat companion for children with a focus on structured entertainment (e.g., story-telling, collaborative games) and engaging in "free-flowing" dialogue with unstructured responses. Concerning behavioral change, studies such as (Calvaresi et al. 2019; Calbimonte et al. 2019) target profiling and cravings' analysis to tailor smoking cessation support, (Calvaresi et al. 2021a) target the maintenance/improvement of physical balance capabilities with personalized exercises. (Chapman et al. 2019), and (Kökciyan et al. 2021) demonstrate the development of a chatbot system to help stroke patients manage their care. The system processes data from multiple inputs (e.g., blood pressure monitor, electronic health record) to serve a computational argumentation engine and respond to user queries.

From a different perspective, data-driven behavior has been addressed in contributions including (Agostaro et al. 2005; Pilato et al. 2007; Augello et al. 2009) which deal with the limitations of the conventional, rule-based, data-driven semantics by introducing the paradigm of LSA. Indeed, according to (Landauer et al. 1998), LSA allows overcoming rule-based pattern matching limits and introduces an element of intuitiveness by constructing a *conceptual* space. Another targeted objective is the integration of multiple domain-specific knowledge sources into one chatbot system. For example, (Jiang et al. 2015; Augello et al. 2011) deal with the integration of different static sources (i.e., vector space model-based indices, XML, relational databases,

Table 6 Technology readiness levels according to the definition provided by (European Commission 2017)

Level	Description
1	basic principles observed
2	technology concepts formulated
3	experimental proof of concept
4	technology validated in lab
5	technology validated in relevant environment (industrially relevant environment in the case of key enabling technologies)
6	technology demonstrated in relevant environment (industrially relevant environment in the case of key enabling technologies)
7	system prototype demonstration in operational environment
8	system complete and qualified
9	actual system proven in operational environment (competitive manufacturing in the case of key enabling technologies)

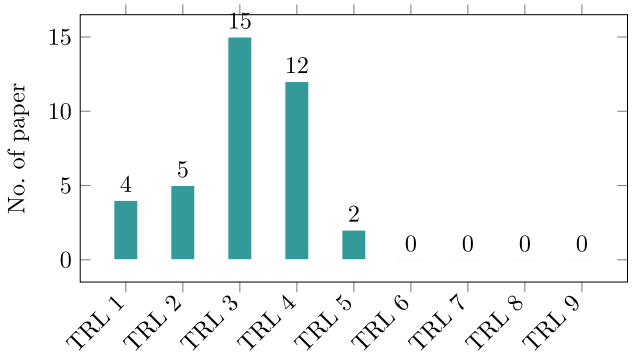


Fig. 11 Technology readiness level distribution of the primary studies

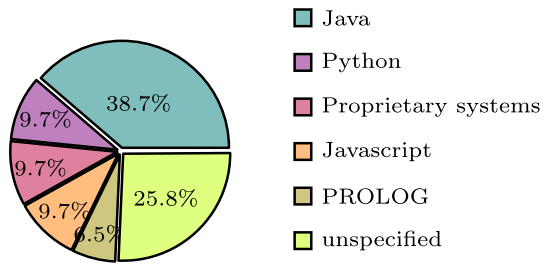


Fig. 12 Overview of utilized back-end technologies

SPARQL queries, and AIML), while (Pilato et al. 2011; Tarau and Figa 2004) are intended to manage knowledge dynamically based on the current dialogue context.

While the studies mentioned above are in a user-to-single agent scope, a few studies are in the user to multi-agents (i.e., chatbots) scope. For example, (de Bayser et al. 2017, 2018) address the coordination of multiple bots providing financial advice within the same chat. Their final goal is the *moderation* of the user-bots' interaction. Finally, (Calvaresi et al. 2021a) focused, among other aspects, on the facets of data protection and data privacy.

5.6 Technology characterization

Studying **SRQ6**, we have classified the primary studies according to the technology readiness level (European Commission 2017) (see Table 6). In turn, we have analyzed the technologies, architecture, and design principles employed in the primary studies.

Assessing the TRL is a valuable way to measure the maturity of a technology/system. The scale was originally devised by NASA ((Sadin et al. 1989)) and is nowadays used in many areas in various forms. In this context, we rely on the definition provided by the European Commission in the context of research and innovation projects ((European Commission 2017)) as shown in Table 6.

The TRL distribution of the primary studies is depicted in Fig. 11. It is noticeable that most of the studies are in Levels 3 and 4 (68.1%). This entails that the final *outcome* of these studies is either a non-validated prototype (TRL 3) or is at the laboratory test stage (TRL 4). Two studies (i.e., (Calvaresi et al. 2019) and (Calvaresi et al. 2021a)) are classified as TRL 5. Indeed, such studies have been deployed and validated in real-world health and social-related campaigns.

In addition to analyzing the TRL of each study, the front-end and back-end technologies applied in the presented systems were analyzed. All studies with a TRL of 3 and higher were considered. Figure 12 depicts the distribution of the back-end technologies used in the primary studies. The majority (38.7%) of the systems employ Java-based back ends. This prevalence can be related to the wide use of MAS frameworks such as JADE[12] and MaSMT.[13] For example, studies such as (Alencar and Netto 2014), (de M. Batista et al. 2009), and (Bentivoglio et al. 2010) rely on JADE and (Hettige and K. 2015) implemented the system based on MaSMT. Although not relying on a pre-existing MAS framework, (Pilato et al. 2007) and (Tarau and Figa 2004) implemented their own ad-hoc Java-based systems. Moreover,

[12] https://jade.tilab.com/.

[13] https://sourceforge.net/projects/masmt/.

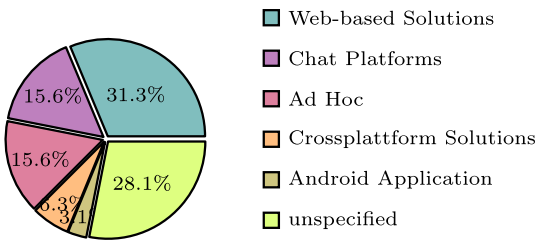


Fig. 13 Overview of utilized frontend technologies

(Estes 2011) exploit features of the Java Enterprise Edition platform (JavaEE) to develop their chatbot system and (Memon et al. 2018) use communication sockets of the Java Standard Edition (Java SE). Several studies use unconventional technologies to develop MAS. For example, (de Bayser et al. 2017) use Akka,[14] an actor-based framework, and (Z. et al. 2016) relied on ActiveMQ,[15] a multi-protocol messaging server.

Python-based back ends are 9.7% of the total. In particular, (Jiang et al. 2015) and (Calvaresi et al. 2019) have developed ad hoc systems, while (Calvaresi et al. 2021a) rely on the SPADE framework.[16]

Several studies (9.7%) relied on existing proprietary systems. For example, (Kalia et al. 2017) and (Angara et al. 2017)) rely on IBM Watson's Conversation Platform[17] and (Zolitschka 2020) rely on Aimpulse Spectrum.[18]

A number of studies (9.7%) developed their systems' back end as ad-hoc solution using JavaScript (i.e., (de Bayser et al. 2018), (Thosani et al. 2020) and (Bosse. 2021)).

6.5% of all studies (i.e., (Tarau and Figa 2004) and (Bosse. 2021)) implemented a PROLOG[19]-based back end. Finally, With a share of 25.8%, a substantial number of studies have developed prototypes but failed to mention details regarding their back end implementation. One such example is (Kökciyan et al. 2021). Although the authors specify the human interface, it does not go into detail about how the actual backend is implemented.

Figure 13 displays the distribution of the front-end technologies used in the developed chatbot systems. Web-based technologies have received the most attention (31.3%), mostly using JavaScript or JavaServer Pages (JSP) in Java.

Using existing web/mobile messaging platform is a choice undertaken by 15.6% of the studies. In particular, (Calvaresi et al. 2019) rely on Facebook Messenger,[20] (Calvaresi et al. 2021a) offer Telegram Messenger[21] among the available interfaces, (Tarau and Figa 2004) use Yahoo Instant Messenger (deprecated since 2012), and (Bentivoglio et al. 2010) adopt Jabber.[22]

The development of ad hoc solutions accounts for 15.6%. the programming languages involved are Java (e.g., (Hettige and K. 2015) or (Tatai et al. 2003)), C#, and C++ (e.g., (Huang et al. 2008)).

6.3% of the elaborated solutions' front ends uses cross-platform frameworks. Such frameworks allow the same code base to be used for web and smartphone app development. For example, The studies used (Thosani et al. 2020) use Ionic,[23] and (Calvaresi et al. 2021a) offer among the possible interfaces HemerApp, which is written in Flutter.[24]

3.1% of systems used an Android application as front end (e.g., (Kökciyan et al. 2021)).

Finally, 28.1% of the studies do not mention what technologies are used in their solution or provide only simplistic and non-classifiable descriptions. For example, (de Bayser et al. 2018) focuses primarily on the conception of the back-end side without mentioning how their human interfacing system was implemented.

5.7 Strengths of the primary studies

Referring to question **SRQ7**, the strengths of the primary studies are listed in Table 7. Among all the strengths, 22% of the strengths are classified as Y, which means that the strengths are explicitly defined and evaluated, 21% are classified as P, indicating that the information is implicitly defined, 57% are classified as N, denoting that the information is not inferrable (see Fig. 14). Figure 15 shows, in particular, the classification per strength.

5.8 Limitations and solutions of the primary studies

Referring to questions **SRQ8** and **SRQ9**, the limitations stated in the studies and their proposed solutions were analyzed. Table 8 lists all limitations acknowledged by the authors and their proposed solutions. Only five of the ten papers that point out limitations proposed solutions to address them. As an unfortunate habit, limitations are often overlooked. However, among those who mentioned limitations, it is possible to identify two main categories: architectural and functional. As architectural limitation, we specify limitations that are of technical nature and can be solved by changing the applied architecture or technologies.

[14] https://akka.io/.

[15] https://activemq.apache.org/.

[16] https://spade-mas.readthedocs.io/en/latest/readme.html.

[17] https://www.ibm.com/cloud/watson-assistant.

[18] https://www.aimpulse.com/.

[19] https://www.iso.org/standard/21413.html.

[20] https://www.messenger.com/.

[21] https://telegram.org/.

[22] https://www.cisco.com/c/en/us/products/unified-communications/jabber/index.html.

[23] https://ionicframework.com/.

[24] https://flutter.dev/.

Table 7 Strengths of the primary studies

Study/strength	S1	S2	S3	S4	S5	S6	S7
(Hettige and K. 2015)	Y						P
(de Bayser et al. 2017)		P					Y
(Estes 2011)	Y	Y			Y	P	P
(de M. Batista et al. 2009)		P					
(Kalia et al. 2017)		P			P		
(Jiang et al. 2015)	P	Y			P	P	
(Angara et al. 2017)	P		Y	Y			
(Z. et al. 2016)		Y			Y		
(Memon et al. 2018)		Y				P	P
(de Bayser et al. 2018)		P			P	P	Y
(Alencar and Netto 2014)	Y						
(Pilato et al. 2007)	Y	P					
(Pilato et al. 2011)	P	P		Y	Y		
(Augello et al. 2009)		P					
(Augello et al. 2011)		Y					
(Tarau and Figa 2004)		P					
(Wong et al. 2012)		P		Y			
(Noori et al. 2014)	P	P					Y
(Huang et al. 2008)		P			Y	Y	
(Bentivoglio et al. 2010)				Y	Y		
(Calvaresi et al. 2019)			P		P		P
(Zolitschka 2020)		P				P	
(Thosani et al. 2020)			Y				
(Calbimonte et al. 2019)			Y		Y		
(Shashaj et al. 2019)		Y	P		Y	Y	P
(Kökciyan et al. 2021)			P			P	
(Chapman et al. 2019)			Y			P	
(Calvaresi et al. 2021a)	P	Y	Y	Y	Y	P	Y
(Maher and Gu 2002)	Y						
(Tatai et al. 2003)				P			
(Mori et al. 2003)							Y
(Bosse 2021)	P	Y	P	P	P	Y	Y

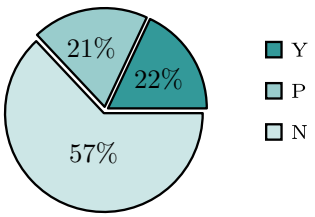


Fig. 14 Overview of strength assessment according to the YPN classification. In particular, Y = the information is explicitly defined/evaluated; P = the information is implicit/stated; N = the truthfulness of the information is not inferable

An example of architectural limitations is (de Bayser et al. 2017), which states performance problems when raising the number of participants in a chat group. To solve this problem, they suggest switching to a micro-service architecture.

Another example is (Calvaresi et al. 2019), emphasizing several limitations of their current system architecture, specifically scaling issues with more complex behaviors, lack of standardized inter-agent communication, and no means of integrating third-party data analysis tools. The solution to these limitations is an entirely new platform based on a MAS. Functional limitations are issues on a functional level that can usually be overcome by exploring alternative approaches to a problem. Examples of functional limitations are (Hettige and K. 2015) and (Jiang et al. 2015), both of which mention limitations related to semantic processing. The proposed solution of (Hettige and K. 2015) is to update the corresponding subsystem, while (Jiang et al. 2015) proposes to analyze the user input with domain-independent analyzers (e.g., linguistic analysis or keyboard analysis).

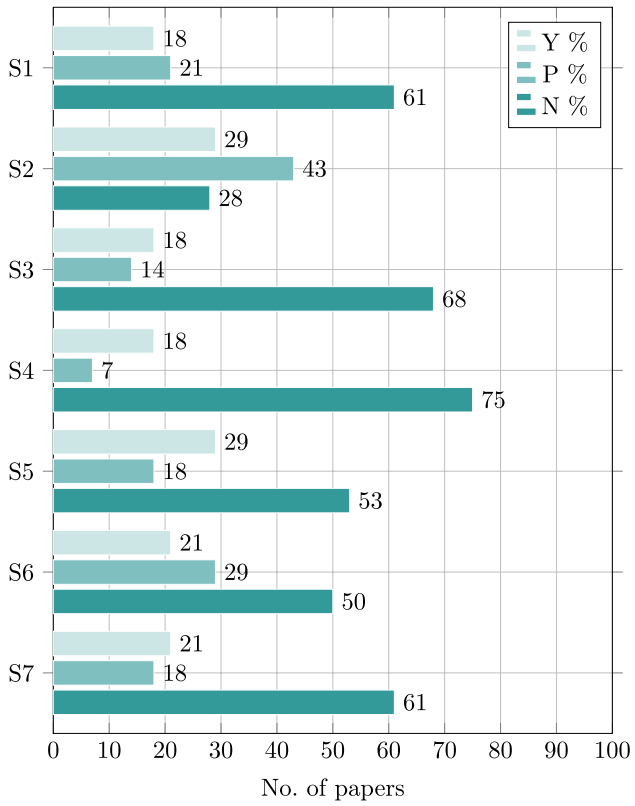


Fig. 15 Qualitative assessment of the strengths (Y-P-N criteria).**S1**: dynamic update of knowledge base; **S2**: adaptability to different domains; **S3**: Profiling (according to user behavior); **S4**: personalization (according to user input); **S5**: reusability of components; **S6**: scalability; **S7**: performance

5.9 Future challenges stated in the primary studies

Concerning **SRQ10** giving the heterogeneous perspective of the future challenges are rather disparate. However, generally, future challenges can be divided into three categories:

- **System-related** challenges relate to extending already existing functionalities.
- **Functionality-related** challenges refer to new functionality to be implemented.
- **User-related** challenges refer to collecting user experiences (usually in the form of trials).

The studies were analyzed for these three categories. Figure 16 shows the breakdown of the three categories across all studies. With 57.9%, most studies desire to enhance their current system's stability or expand already implemented functionalities. For example, (Shashaj et al. 2019) see improving the system component stability and interoperability with other FIPA[25]-compliant MAS environments

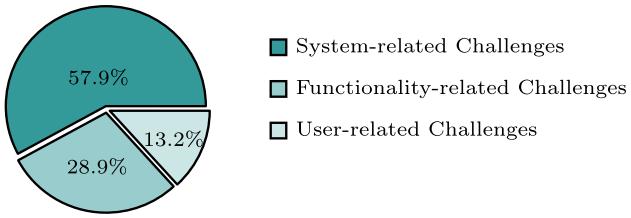
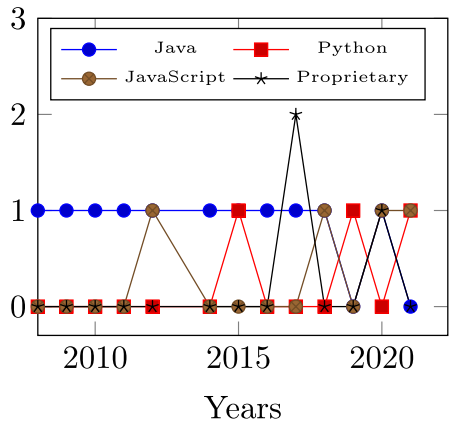
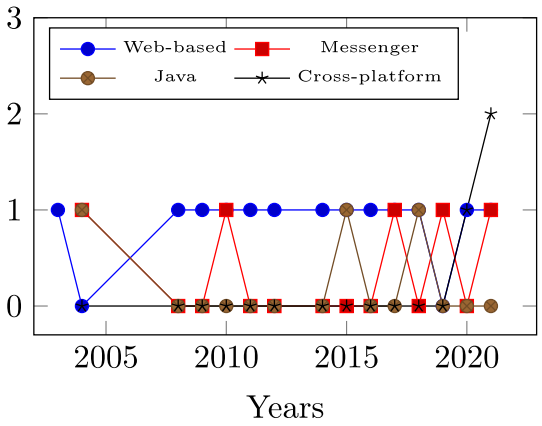


Fig. 16 Distribution of future challenge per category



(a) Utilized back-end technologies.



(b) Utilized front-end technologies

Fig. 17 MAS-based chatbot technologies over the years

as a future goal, whereas (Calvaresi et al. 2019) wish to adapt their architecture to allow distributed computing among several servers to increase performance and to handle agent migration from one server instance to another. A complete list of system-related challenges can be seen in Table 9. At 28.9%, about one-third of studies are endeavoring to add new functionalities to their existing system.

[25] Foundation for Intelligent Physical Agents.

Table 8 Study limitations and proposed solutions

Studies	Limitations	Proposed solutions
(Hettige and K. 2015)	Limited semantic processing ability	Update language processing to enhance system intelligence
(de Bayser et al. 2017)	Scalability issues when increasing users	Micro-service implementation to improve scalability
(Vasconcelos et al. 2017)	Predefined test scenarios and predicted results	
(Jiang et al. 2015)	Problems identifying domain based on user context	Analyze user input with domain-independent analyzer
(Z. et al. 2016)	Costly detailed development, usability/layout problems, heterogeneity in a unified system	Unifying agent front end
(Agostaro et al. 2005)	Large amount of data necessary to use LSA	
(Wong et al. 2012)	Limited coverage of content and knowledge base	Automatic content mining to increase accuracy
(Calvaresi et al. 2019)	Complex behaviors compromise scalability, no integration of 3rd party data analysis, missing standardized agent communication	Upgrade to MAS platform is necessary
(Zolitschka 2020)	Only simple parametrization considered, Real-world application missing	
(Calvaresi et al. 2021a)	Synergies with non-agent frameworks unexplored, missing seamless integration of diverse knowledge, functionality for medical personnel needed	
(Maher and Gu 2002)	Components of virtual architecture need to be predefined, programmed and stored online statically	
(Tatai et al. 2003)	No emotional memory, possible emotional overload	Emotional memory to store dominant emotions for longer period
(Bosse 2021)	Short-time and range communication issues, simplified NLP approach leads to reduced dialogue quality	

(Vasconcelos et al. 2017) attempt to implement more test metrics to test more aspects of a chatbot system, and (Memon et al. 2018) seek to expand their chatbot with a graphical user interface and extend its user input capabilities with voice recognition and interpretation. All functionality-related future challenges are listed in Table 10. 13.2% of all future challenges focus on capturing user feedback. (Alencar and Netto 2014) are seeking to test their tutoring system with the help of students and make further improvements to the system based on the feedback collected, and (Kökciyan et al. 2021) are conducting two pilot studies with patients to test different aspects of their system. Table 11 lists all user-related challenges stated in the primary studies.

6 Discussion

Analyzing the primary studies emerges that the application of the MAS' paradigm has slightly increased in the past twenty years, although only moderately. The elaborated works acknowledge the suitability and the intrinsic added value of agent-based systems, including autonomy, goal-setting, and behavior definition. Nevertheless, it appears that these technologies are mostly at an early stage of development. On the one hand, the TRL of most primary studies did not exceed level 3 or 4 (as shown in Fig. 11), and it is questionable whether these early-stage systems would be capable of meeting the requirements of a real-world scenario. On the other hand, a few systems have been studied in real-world scenarios (i.e., (Calvaresi et al. 2021a)—testing the developed chatbot in a physical balance-preserving campaign and (Kökciyan et al. 2021) – letting both experts and real users analyzing the system. However, it still remains to test such systems in fully operational environments.

Several studies focused on aspects revolving around the management and reconciliation of different knowledge bases. However, only one (Calvaresi et al. 2021a) has addressed the topic of data privacy and user consent directly. To date, this is a remarkable concern that practitioners have to address imperatively. Indeed, too many studies addressing topics such as user profiling and the processing of user input to enhance chatbot knowledge have either ignored data privacy or not tackled it explicitly. If people are involved, it is of paramount importance to ensure their control over their data. Due to the

Table 9 Future challenges: system-related

Study	System-related future challenges
(Hettige and K. 2015)	Passing the Turing test
(de Bayser et al. 2017)	Support for decoupled interaction norms specifications
(Kalia et al. 2017)	Implement tooling to guide users through the methodology
(Angara et al. 2017)	Connect Foodie to smart appliances and smart services, integrate it into digital assistants (e.g., Siri), establish Foodie a "Cognitive Internet of Things (IoT) Recipe Maven"
(Z. et al. 2016)	Link the system to other spoken dialogue systems to attract users; pass stability tests when operating several remote agents
(de Bayser et al. 2018)	Implement learning process to derive conversation rules from dialogue corpus
(Agostaro et al. 2005)	Implement bots using the LSA application for all user interactions to take dialogue history and context into account
(Pilato et al. 2007)	Enhance the "associative" interaction and update mechanisms of the knowledge base
(Augello et al. 2009)	Further explore opportunities offered by the proposed architecture
(Tarau and Figa 2004)	Improve extraction methods from query/answering transcripts, for agent scripts and story-specific metadata. Extract complex conversational intelligence for deep natural language query pattern matching
(Wong et al. 2012)	Develop a proactive engagement model where the bot actively monitors user engagement and applies conversational strategies when required
(Noori et al. 2014)	Compare current pattern matching technique with newly implemented semantic similarity technique
(Huang et al. 2008)	Move animation synchronization part from animator to real-time driving message passing to allow wider range of components
(Calvaresi et al. 2019)	Allow distributed computing among several servers to increase performance. Handle agent migration from one server instance to another
(Zolitschka 2020)	Enhance the system to demonstrate and evaluate the framework in terms of sequential dependent topics
(Thosani et al. 2020)	Implement reinforced learning mechanisms to use the system in additional domains
(Calbimonte et al. 2019)	Implement Key Performance Indicators (KPIs) to evaluate persuasion metrics. Inclusion of vocabularies and ontologies to formalize domain knowledge models and argumentation graphs. Specify mechanisms for agent coordination and negotiation to incorporate computational persuasion. Study privacy concerns and adopt ethics-by-design methodologies
(Shashaj et al. 2019)	Improve component usability as well as interoperability with other FIPA-compliant MAS environments
(Calvaresi et al. 2021a)	Seamless integration of contextually diverse knowledge. Dynamically integrate user-groups. Study automation of feedback classification and place autonomous logic triggers for sensitive feedback. Investigate run-time definition of agent behaviors
(Tatai et al. 2003)	Refinement of the models
(Mori et al. 2003)	Increase functionality of the module, such as using more efficient state transition techniques
(Bosse 2021)	Extend the simplified NLP model to increase the quality of dialogues

implementation of more rigid data privacy laws such as GDPR, next-generation systems must have no room to neglect this topic.

The analysis of the technologies' distribution within the primary studies suggests some trends to be observed. Figure 17a shows the back-end technologies used over the years. It is possible to notice that Java-based systems have been used extensively. However, since 2015, Python-based systems have emerged, likely due to Python's prevalence in areas such as machine learning and data science libraries. Moreover, since 2017, the employment of proprietary systems (e.g., IBM Watson) has been increasingly considered. Although initially rather rudimentary, such platforms now offer a wide range of possibilities, such as integrating machine learning modules or extensive analytical capacities. Figure 17b shows that a shift occurred in the area of front-end technologies too. In addition to the increasing prevalence of web-based solutions, messaging services such as Facebook Messenger or Telegram have become increasingly popular since 2015. Nevertheless, in recent years, the use of cross-platform frameworks became a consistent practice. Cross-platform frameworks such as Ionic or Flutter make it possible to develop front-end solutions for mobile phones and web browsers using a single code base. Moreover, it can be observed a trend to use more complex multi-agent chatbots (e.g., (Bosse. 2021)) to blend in IoT and micro-services domain with highly scalable multi-agent chatbot networks.

Most studies have used MAS enabling agents to abstract individual components such as language processing or output composition. (Calvaresi et al. 2019) and (Calvaresi et al. 2021a) have taken a different approach by coupling

Table 10 Future challenges: functionality-related

Study	Functionality-related Future Challenges
(de Bayser et al. 2017)	Develop a multi-party governance service to enforce exchange of compliant utterances
(Vasconcelos et al. 2017)	Add similarity metrics (e.g. perplexity and distance measures) to account for partially correct answers
(Jiang et al. 2015)	Implement new plug-ins to support other types of knowledge sources. Enhance currently implemented plug-ins to fully support the specific knowledge sources
(Memon et al. 2018)	Implement a graphical user interface. Add Natural Language Understanding (NLU)/NLP, so the system can understand spoken language and translate it into text. Develop a multi-party chatbot system supporting emojis, animations etc
(Alencar and Netto 2014)	Create new intelligent agents that monitor other activities. Improve avatar gestures
(Tarau and Figa 2004)	Find creative uses for the new Google metasearch API
(Huang et al. 2008)	Develop deliberate phase with internal context state instead of current simple AIML script executor. Extend system to support simultaneous multiple sessions to run in a web-environment
(Bentivoglio et al. 2010)	Implement a dashboard that merges statistical analysis and performance indicators to allow teachers and tutors to monitor course activities and participation and schedule interventions. Implement user profiling and customization of the learning process
(Calvaresi et al. 2019)	Analyze the data of previous cessation programs to adapt decisions and therefore provide better service to users. Implement explanatory behavior, allowing to explain the rationale behind treatment decisions to the user to increase user trust
(Chapman et al. 2019)	Integrate wireless sensor data collection, patient personalized treatment plan etc. into the platform to dynamically adapt the treatment based on that data. Capture patient decisions about and responses to daily care in a standardized way to enable treatment effectiveness statistics
(Tatai et al. 2003)	Introduce more automated features for emotion-message act assignment

Table 11 Future challenges: user-related

Study	User-related future challenges
(Vasconcelos et al. 2017)	Evaluate usability with real users
(Kalia et al. 2017)	Conduct developer studies to evaluate framework effectiveness
(Alencar and Netto 2014)	Conduct experiments with a class of students
(Zolitschka 2020)	Analyze feedback of real users to evaluate the proposed approach in a real-world application
(Kökciyan et al. 2021)	Pilot study where stroke patients will use the platform to self-manage their health conditions; General-practitioner experts will evaluate the generated recommendation derived from clinical guidelines

the users themselves with personalized agents. According to such studies, the goal of this 1:1 relation is to facilitate user profiling, data management, privacy preservation, and personalization. Indeed, by interacting with the user, the respective agent is expected to increase its knowledge and enhance the personalization's accuracy level over time.

Looking at the evaluation of the strengths of the primary studies in Fig. 14, it is noticeable that S2 (i.e., adaptability to different domains) and S6 (i.e., scalability) have an above-average number of implicitly defined and evaluated strengths. In the case of S2, this is primarily due to studies having justified their system's adaptability with the implementation of a single case study to conclude that the system can also be applied to other domains. This is not necessarily a wrong assumption, but the implementation of several distinct scenarios would have been more effective to show this strength explicitly. Compared to S2, S6 is a more generic strength. Since most of the studies are in an early prototype stage, even if the respective systems' scalability was reported as a strength, this strength was mostly not evaluated. This leads to the question of what methods can be used to evaluate a chatbot platform's scalability. All studies use the term scalability as a synonym for *Size Scalability* as defined by (Neuman 1994). *Size Scalability* defines that a system scales easily with the number of users and resources without noticeable loss of performance. To implicitly define this aspect, a load test with several simulated users in which the response times and hardware load of the system are analyzed could be theoretically sufficient.

7 Conclusions

This paper has analyzed the current state of the art of chatbot solutions leveraging the multi-agent approach and agent-based frameworks by performing an SLR. In particular, it employs a well-established methodology characterized by ten structured research questions. Such an investigation focused on aspects including application domains, end-users, requirements, objectives, technology readiness level, designs, strengths, limitations, and future challenges of the solutions found in the literature. Such aspects have been analyzed "per-feature" and overall aggregated in a reconciling discussion. The insights elicited in this work can be beneficial for both theoretical and practical future research.

Acknowledgements The work is supported by: ◇ HES-SO RCSO ISNet PERSA project: Computational Persuasion in eHealth Support Applications. ◇ Chist-Era grant CHIST-ERA19-XAI-005, and by (i) the Swiss National Science Foundation (G.A. 20CH21_195530), (ii) the Italian Ministry for Universities and Research, (iii) the Luxembourg National Research Fund (G.A. INTER/CHIST/19/14589586), (iv) the Scientific, and Research Council of Turkey (TÜBİTAK, G.A. 120N680) ◇ The authors thank Diego Collarana from Universidad Privada Boliviana for his insights. This collaboration was supported by the Research Partnership Grant RPG2106 funded by the Swiss Leading House for Latin America.

Funding Open access funding provided by University of Applied Sciences and Arts Western Switzerland (HES-SO).

References

Adamopoulou E, Moussiades L (2020) Chatbots: History, technology, and applications. Machine Learning with Applications 2(100):006

Agostaro F, Augello A, Pilato G et al (2005) A conversational agent based on a conceptual interpretation of a data driven semantic space. Congress of the Italian Association for Artificial Intelligence. Springer, New York, pp 381–392

Alencar M, Netto JF (2014) Tutor collaborator using multi-agent system. International Conference on Collaboration Technologies. Springer, New York, pp 153–159

Angara P, Jiménez M, Agarwal K, et al (2017) Foodie fooderson a conversational agent for the smart kitchen. In: Mindel M, Lyons KA, Wigglesworth J (eds) Proceedings of the 27th Annual International Conference on Computer Science and Software Engineering, CASCON 2017, Markham, Ontario, Canada, November 6-8, 2017. IBM / ACM, pp 247–253, http://dl.acm.org/citation.cfm?id=3172825

Anjomshoae S, Najjar A, Calvaresi D, et al (2019) Explainable agents and robots: Results from a systematic literature review. In: 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019, International Foundation for Autonomous Agents and Multiagent Systems, pp 1078–1088

Augello A, Pilato G, Vassallo G et al (2009) A semantic layer on semi-structured data sources for intuitive chatbots. 2009 International Conference on Complex. Intelligent and Software Intensive Systems, IEEE, pp 760–765

Augello A, Scriminaci M, Gaglio S, et al (2011) A modular framework for versatile conversational agent building. In: 2011 International Conference on Complex, Intelligent, and Software Intensive Systems, IEEE, pp 577–582

Augello A, Gentile M, Dignum F (2017) An overview of open-source chatbots social skills. International conference on internet science. Springer, New York, pp 236–248

Bentivoglio C, Bonura D, Cannella V et al (2010) Intelligent agents supporting user interactions within self regulated learning processes. J E-learn Knowl Soc 6(2):27–36

Bosse. S (2021) Distributed serverless chat bot networks using mobile agents: A distributed data base model for social networking and data analytics. In: Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART,, INSTICC. SciTePress, pp 398–405, https://doi.org/10.5220/0010319503980405

Calbimonte JP, Calvaresi D, Dubosson F et al (2019) Towards profile and domain modelling in agent-based applications for behavior change. International Conference on Practical Applications of Agents and Multi-Agent Systems. Springer, New York, pp 16–28

Calvaresi D, Dubovitskaya A, Calbimonte JP et al (2018) Multi-agent systems and blockchain: results from a systematic literature review. International conference on practical applications of agents and multi-agent systems. Springer, New York, pp 110–126

Calvaresi D, Calbimonte JP, Dubosson F, et al (2019) Social network chatbots for smoking cessation: agent and multi-agent frameworks. In: 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, pp 286–292

Calvaresi D, Calbimonte JP, Siboni E et al (2021) Erebots: Privacy-compliant agent-based platform for multi-scenario personalized health-assistant chatbots. Electronics. https://doi.org/10.3390/electronics10060666

Calvaresi D, Ibrahim A, Calbimonte JP et al (2021) The evolution of chatbots in tourism: a systematic literature review. Inf Commun Technol Tour 2021:3–16

Chapman M, Balatsoukas P, Ashworth M, et al (2019) Computational argumentation-based clinical decision support. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '19, p 2345-2347

Cui L, Huang S, Wei F, et al (2017) Superagent: A customer service chatbot for e-commerce websites. In: Proceedings of ACL, System Demonstrations, pp 97–102

de M. Batista AF, Marietto MdGB, Barbosa GCdO, et al (2009) Multi-agent systems to build a computational middleware: A chatterbot case study. In: 2009 International Conference for Internet Technology and Secured Transactions,(ICITST), IEEE, pp 1–2

de Bayser MG, Cavalin PR, Souza R, et al (2017) A hybrid architecture for multi-party conversational systems. CoRR arXiv: abs/1705.01214

de Bayser MG, Pinhanez C, Candello H, et al (2018) Ravel: a mas orchestration platform for human-chatbots conversations. In: The

6th International Workshop on Engineering Multi-Agent Systems. Stockholm, Sweden
DeepLink (2022) Deeplink.ai: Artificial intelligence to boost your customer relationship. https://www.deeplink.ai/en/, Accessed: 2023-02-24
Estes TW (2011) Knowledge discovery agent system and method. US Patent 8,015,143
Etherington D (2014) Amazon echo is a \$199 connected speaker packing an always-on siri-style assistant - techcrunch. https://techcrunch.com/2014/11/06/amazon-echo/, Accessed: 2023-02-24
European Commission (2017) Appendix g. technology readiness levels (trl). https://ec.europa.eu/research/participants/data/ref/h2020/other/wp/2016_2017/annexes/h2020-wp1617-annex-g-trl_en.pdf
Fadhil A, Gabrielli S (2017) Addressing challenges in promoting healthy lifestyles: the al-chatbot approach. In: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, ACM, pp 261–265
Galster M, Weyns D, Tofan D et al (2014) Variability in software systems-a systematic literature review. IEEE Trans Software Eng 40(3):282–306. https://doi.org/10.1109/TSE.2013.56
Guzzoni D (2008) Active: a unified platform for building intelligent applications (phd thesis) pp 1–263. https://doi.org/10.5075/epfl-thesis-3990, http://infoscience.epfl.ch/record/114758
Hettige B, K. A (2015) Octopus: a multi agent chatbot. In: 8th International Research Conference, KDU, November 2015
Huang HH, Cerekovic A, Tarasenko K et al (2008) Integrating embodied conversational agent components with a generic framework. Multiagent Grid Syst 4(4):371–386
Huddar MG, Sannakki SS, Rajpurohit VS (2021) Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN. Int J Interact Multim Artif Intell 6(6):112–121. https://doi.org/10.9781/ijimai.2020.07.004
Hung V, Elvir M, Gonzalez A, et al (2009) Towards a method for evaluating naturalness in conversational dialog systems. In: 2009 IEEE international conference on systems, man and cybernetics, IEEE, pp 1236–1241
Jiang R, Banchs RE, Kim S, et al (2015) Configuration of dialogue agent with multiple knowledge sources. In: 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), IEEE, pp 840–849
Kalia AK, Telang PR, Xiao J, et al (2017) Quark: a methodology to transform people-driven processes to chatbot services. In: International Conference on Service-Oriented Computing, Springer, pp 53–61
Kitchenham B, Pearl Brereton O, Budgen D et al (2009) Systematic literature reviews in software engineering—a systematic literature review. Inf Softw Technol 51(1):7–15. https://doi.org/10.1016/j.infsof.2008.09.009
Kitchenham B, Brereton P, Turner M et al (2010) Refining the systematic literature review process-two participant-observer case studies. Empir Softw Eng 15(6):618–653. https://doi.org/10.1007/s10664-010-9134-8
Kökciyan N, Sassoon I, Sklar E et al (2021) Applying metalevel argumentation frameworks to support medical decision making. IEEE Intell Syst 36(2):64–71. https://doi.org/10.1109/MIS.2021.3051420
Landauer TK, Foltz PW, Laham D (1998) An introduction to latent semantic analysis. Discourse Process 25(2–3):259–284. https://doi.org/10.1080/01638539809545028
Maher ML, Gu N (2002) Design agents in virtual worlds-a user-centred virtual architecture agent. Agents in Design. pp 23–38
Mauldin ML (1994) Chatterbots, tinymuds, and the turing test entering the loebner prize competition. In: Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence. AAAI Press, AAAI'94, p 16-21
Memon Z, Jalbani AH, Shaikh M et al (2018) Multi-agent communication system with chatbots. Mehran Univ Res J Eng Technol 37(3):663
Mori K, Jatowt A, Ishizuka M (2003) Enhancing conversational flexibility in multimodal interactions with embodied lifelike agent. In: Proceedings of the 8th international conference on Intelligent user interfaces, pp 270–272
Mualla Y, Najjar A, Daoud A et al (2019) Agent-based simulation of unmanned aerial vehicles in civilian applications: a systematic literature review and research directions. Futur Gener Comput Syst 100:344–364. https://doi.org/10.1016/j.future.2019.04.051
Neuman B (1994) Scale in distributed systems. Inf. Sci. Inst., Univ. Southern California (ISI/USC), Los Angeles, CA, USA. 68
Noori Z, Bandarl Z, Crockett K (2014) Arabic goal-oriented conversational agent based on pattern matching and knowledge trees. In: Proceedings of the World Congress on Engineering 2014 Vol I. Newswood/International Association of Engineers, July 2 - 4, 2014, London, UK. ISSN 2078-0958, https://e-space.mmu.ac.uk/id/eprint/609597
Palmarini R, Erkoyuncu JA, Roy R et al (2018) A systematic review of augmented reality applications in maintenance. Robot Comput-Integr Manuf 49:215–228
Pereira J, Díaz Ó (2019) Using health chatbots for behavior change: a mapping study. J Med Syst 43(5):135
Pilato G, Augello A, Vassallo G, et al (2007) Sub-symbolic semantic layer in cyc for intuitive chat-bots. In: International Conference on Semantic Computing (ICSC 2007), IEEE, pp 121–128
Pilato G, Augello A, Gaglio S (2011) A modular architecture for adaptive chatbots. In: 2011 IEEE Fifth International Conference on Semantic Computing, IEEE, pp 177–180
Radianti J, Majchrzak TA, Fromm J et al (2020) A systematic review of immersive virtual reality applications for higher education: design elements, lessons learned, and research agenda. Comput Educ 147(103):778
Rao SBP, Agnihotri M, Babu Jayagopi D (2021) Improving asynchronous interview interaction with follow-up question generation. Int J Interact Multim Artif Intell 6:79–89. https://doi.org/10.9781/ijimai.2021.02.010
Rollo C (1997) jabberwacky - about thoughts - an artificial intelligence ai chatbot, chatterbot or chatterbox, learning ai, database, dynamic - models way humans learn - simulate natural human chat - interesting, humorous, entertaining. http://www.jabberwacky.com/j2about, Accessed: 2022-01-10
Sadin SR, Povinelli FP, Rosen R (1989) The nasa technology push towards future space mission systems. Acta Astronaut 20:73–77. https://doi.org/10.1016/0094-5765(89)90054-4
Shashaj A, Mastrorilli F, Stingo M et al (2019) An industrial multi-agent system (mas) platform. International Conference on P2P Parallel, Grid, Cloud and Internet Computing, Springer, New York. pp 221–233
Tarau P, Figa E (2004) Knowledge-based conversational agents and virtual storytelling. In: Proceedings of ACM symposium on Applied computing, pp 39–44
Tatai G, Csordás A, Kiss Á, et al (2003) The chatbot who loved me. In: Proc. ECA Workshop of AAMAS, Melbourne, Australia
Thosani P, Sinkar M, Vaghasiya J, et al (2020) A self learning chat-bot from user interactions and preferences. In: 4th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, pp 224–229
Vasconcelos M, Candello H, Pinhanez C, et al (2017) Bottester: testing conversational systems with simulated users. In: Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems, pp 1–4
Voigt P, Von dem Bussche A (2017) The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed, vol 10, issue 3152676. Springer International Publishing, Cham. pp 10–5555

Weizenbaum J (1966) Eliza-a computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45

Winkler R (2018) Söllner M (2018) Unleashing the potential of chatbots in education: a state-of-the-art analysis. Acad Manag Proc 15:903. https://doi.org/10.5465/AMBPP.2018.15903abstract

Wong W, Cavedon L, Thangarajah J et al (2012) Flexible conversation management using a bdi agent approach. International Conference on Intelligent Virtual Agents. Springer, New York, pp 464–470

Xu A, Liu Z, Guo Y, et al (2017) A new chatbot for customer service on social media. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, ACM, pp 3506–3510

Yang ECL, Khoo-Lattimore C, Arcodia C (2017) A systematic literature review of risk and gender research in tourism. Tour Manage 58:89–100

Zhao T, Lee K, Eskenazi M (2016) The dialport portal: grouping diverse types of spoken dialog systems. In: Workshop on Chatbots and conversational agents

Zolitschka JF (2020) A novel multi-agent-based chatbot approach to orchestrate conversational assistants. International Conference on Business Information Systems. Springer, New York, pp 103–117