Please quote as: Schmitt, A.; Wambsganss, T.; Leimeister, J. M. (2022). Conversational Agents for Information Retrieval in the Education Domain: A User-Centered Design Investigation. Proceedings of the ACM on Human-Computer Interaction (PACM), 6 (CSCW).

Conversational Agents for Information Retrieval in the Education Domain: A User-Centered Design Investigation

Article *in* Computer Supported Cooperative Work (CSCW) · November 2022

DOI: 10.1145/3555587

CITATIONS
0

READS
164

3 authors:

Anuschka Schmitt
University of St.Gallen
9 PUBLICATIONS **20** CITATIONS
SEE PROFILE

Thiemo Wambsganß
École Polytechnique Fédérale de Lausanne
49 PUBLICATIONS **338** CITATIONS
SEE PROFILE

Jan Marco Leimeister
University of St.Gallen
1,032 PUBLICATIONS **12,912** CITATIONS
SEE PROFILE

Conversational Agents for Information Retrieval in the Education Domain: A User-Centered Design Investigation

ANUSCHKA SCHMITT, University of St.Gallen, Switzerland
THIEMO WAMBSGANSS, EPFL, Switzerland
JAN MARCO LEIMEISTER, University of St.Gallen, Switzerland and University of Kassel, Germany

Text-based conversational agents (CAs) are widely deployed across a number of daily tasks, including information retrieval. However, most existing agents follow a default design that disregards user needs and preferences, ultimately leading to a lack of usage and an unsatisfying user experience. To better understand how CAs can be designed in order to lead to effective system use, we deduced relevant design requirements from both literature and 13 user interviews. We built and tested a question-answering, text-based CA for an information retrieval task in an education scenario. Results from our experimental test with 41 students indicate that following a user-centered design has a significant positive effect on enjoyment and trust in a CA as opposed to deploying a default CA. If not designed with the user in mind, CAs are not necessarily more beneficial than traditional question-answering systems. Beyond practical implications for effective CA design, this paper points towards key challenges and potential research avenues when deploying social cues for CAs.

CCS Concepts: • **Applied computing** → **Interactive learning environments**; • **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → *Laboratory experiments.*

Additional Key Words and Phrases: conversational agent, interaction design, user-centered, trust, enjoyment

ACM Reference Format:
Anuschka Schmitt, Thiemo Wambsganss, and Jan Marco Leimeister. 2022. Conversational Agents for Information Retrieval in the Education Domain: A User-Centered Design Investigation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 486 (November 2022), 22 pages. https://doi.org/10.1145/3555587

1 INTRODUCTION

Conversational agents (CAs) have been deployed in a wide range of domains and applications including assisted driving [33], customer service [5], health therapy [4], security [53], and education [32, 77]. Based on natural language processing (NLP) and machine learning (ML) models, such agents can be tailored towards a particular interaction context or target group. They thereby commonly adopt social cues to facilitate acceptance, enjoyment and achievement of interaction goals [20]. A predominant and promising task for CA interaction is information retrieval [56], especially in the fields of healthcare, e-commerce, and education [21, 37]. Traditional information retrieval and question-answering systems including email, websites, or simple FAQ sheets allow users to access relevant information [10, 63, 66]. However, each of these systems comes with certain limitations, including delayed response time, a lack of information and system quality, unstructured and excessive amounts of information, as well as enforcement of personal information disclosure

Authors' addresses: Anuschka Schmitt, anuschka.schmitt@unisg.ch, University of St.Gallen, St.Gallen, Switzerland; Thiemo Wambsganss, thiemo.wambsganss@epfl.ch, EPFL, Lausanne, Switzerland; Jan Marco Leimeister, janmarco.leimeister@unisg.ch, University of St.Gallen, St.Gallen, Switzerland, University of Kassel, Kassel, Germany.

2573-0142/2022/11-ART486
https://doi.org/10.1145/3555587

[2, 68]. The rise and commercial availability of (customized) CAs offer to overcome previously raised issues of personalization, responsiveness, and availability. CAs promise to redefine human–computer interaction (HCI) and can thus change how information is retrieved and understood [53].

More interestingly, CAs offer a more natural and effortless, even human-like interaction. Despite the wide reliance on CAs and their social cues, most existing CAs are not grounded in evidence from theory. Simultaneously, when it comes to natively built CAs or CAs deployed in a particular domain, default versions are commonly relied upon [9, 11, 22]. We believe that grounding the design of text-based CAs in established theory can increase the CA's effectiveness and can contribute significantly to the acceptance of and trust in the provided information [24, 49]. While much research points towards the potential of the anthropomorphization of agents and the embedding of social cues within CAs, little is known about the downstream consequences of such design features. As part of this study, we are interested in understanding and investigating the effect of user-centered, social design cues of a CA on behavioral task outcomes beyond perceptual and intentional variables.

This paper focuses on applying a user-centered, theory-driven dual design science approach to a CA deployed in an information retrieval task. We designed a text-based CA called Hermine that supports students in retrieving course-relevant information and in posing course-related questions. We thereby followed a theory-motivated approach where we systematically reviewed literature in the field of question-answering, education technology, and HCI to derive concrete requirements and principles for the design of a CA for learning-related information retrieval. In a subsequent step, we identified user requirements from 13 semi-structured interviews with students. In consideration of the requirements derived from literature and users, we built a prototype and ultimately a final version of Hermine, which we tested as part of multiple design hypotheses with potential users. To create Hermine, we (1) implemented the derived design principles and functionalities and (2) developed question-answering models through ML-based interaction intents to model a learner–tutor question and answering scenario using Python frameworks flask, spaCy, and Chatterbot. To evaluate our design and assess the impact of our CA, we compared Hermine to a traditional question-answering system and a non-design-driven CA that did not incorporate our derived design principles. As part of an online lab experiment, we asked participants (N = 41) to retrieve information on a university course in order to answer six learning-related and course-related questions. We found that students interacting with the CA designed based on our design principles provided more correct answers to all questions than students from the other two treatment groups (traditional question-answering system and non-design-driven CA). More so, users interacting with Hermine trusted the provided information more and reported higher levels of enjoyment throughout the task. Our study results do not find proof that a non-design-driven CA performs significantly better than a traditional question-answering system.

The main contribution of this paper is twofold: First, we demonstrate the effectiveness of developing a theory-motivated and user-centered CA by comparing Hermine with an instantiation of current default CAs for information retrieval and a traditional question-answering system in an educational scenario. The results demonstrate the benefits of leveraging NLP and ML to foster a user-driven design for CAs. Second, the findings of our experimental evaluation depict that deploying a CA is not always the most desirable option, in particular when it is not explicitly designed with the user in mind.

The paper is organized as follows. We first review related literature on the design of CAs and introduce the concept of social response theory and its relevance for CA design. Next, we describe our text-based CA for information retrieval of course-related content. We then describe

the methodology of our experiment and discuss related results. Lastly, we offer implications for theory, design, and practice.

2 RELATED WORK

This section presents related work on CAs and information retrieval, information retrieval tasks in educational settings, social response theory, and user-centered design for CAs.

2.1 Conversational Agents and Information Retrieval

CAs are dialogue-based interfaces built with NLP that enhance existing tools through a text-based or speech-based conversational interaction [46, 55, 61]. Providing such a conversational experience can be achieved through CAs' ability to identify, understand, and react to user intents, either through natural sentence structures or keyword triggers. CAs can be implemented in a variety of electronic hardware, such as smart personal assistants (e.g., Amazon's Alexa) or wearables, allowing for new forms of HCI. Ultimately, beyond general-purpose deployments (i.e., Apple's Siri or Amazon's Alexa), CAs allow for services to become more accessible (i.e., healthcare) or for companies to improve their operational efficiency [70] and thus can be implemented in a domain-specific context. The increasingly sophisticated interaction quality of CAs raises the question of how particular tasks currently handled by traditional, non-conversational, and static systems can be enhanced through a novel, conversational design to ultimately enhance the user experience and improve task outcomes.

In particular, CAs for question-answering offer to retrieve relevant information easier, faster, and more effectively while enabling a more adaptive, social, and personalized user experience [32]. Embedded in CAs, this novel technology replaces traditional information retrieval systems [63], such as users searching for answers in matching texts using keywords [6]. Furthermore, CAs promise to speed up the information search process by allowing individuals to directly acquire the correct response to their queries rather than browsing through a collection of potential replies [53]. Research has shown that dialogue-based question-answering is more natural to humans than traditional keyword searches, especially for individuals with poor technical skills [71]. As a result, conversational CAs reshape how people access and retrieve data by improving ease of use [49] and user acceptance rates [24] for information retrieval tasks.

2.2 Information Retrieval Tasks in Educational Settings

MOOCs as well as traditional large-scale lectures at universities face the challenge of providing students with individualized support. In fact, public universities exhibit student–educator ratios of 100 to 1, with this ratio even rising to 10,000 to 1 for time-independent and location-independent online formats [73, 80]. With the decreasing capacity to directly address every student and each request, this lack of student support leads to dissatisfaction, increased course dropout rates, and poor learning outcomes [8, 19]. The possibility to address students' simple and personal questions is naturally hindered, and educators are often confronted and overwhelmed with large amounts of repetitive questions.

Currently, question-answering tasks to communicate with educators and retrieve relevant learning material are still largely handled through human interactions as well as static interfaces such as websites or FAQ documents [12]. Besides the social component, human–human interaction through email or telephone offers students tailored and personalized responses and the possibility to ask follow-up questions on a more detailed level. At the same time, however, student requests result in an information overload on the educator side, which can ultimately lead to delayed and non-satisfactory responses for students, particularly with regard to educator–learner ratios nowadays [12, 68]. Alternatively, students can obtain course-relevant information from university

websites or FAQ documents, which offer on-demand availability independent of time and place [66]. Nevertheless, website information and FAQ documents remove social interaction and the personalization of answers while easily overwhelming the user with large amounts of information [2].

This is where the potential of CAs comes into play. Extant research demonstrates the opportunities for and applicability of CAs being used in large-scale educational settings within and beyond information retrieval tasks. First empirical results illustrate the positive effect of their deployment on student engagement, participation levels, and course continuation [29, 73]. Ultimately, CAs for information retrieval tasks can be deployed to provide students with precise and personalized information about learning concepts, course contents, or administrative issues. Specific tasks include information retrieval on financial services, study content, enrolment, and admissions, or technical problems.

Studies on CAs for information retrieval and educational scenarios have focused on the technical feasibility of developing and embedding question-answering systems in CAs (e.g., [12, 39, 50]). For example, [44] used a forwarding chaining ontology to create a CA for question-answering tasks for students on Dialogflow. [12] described how to build a Telegram-based question-answering CA using Dialogflow and how to improve its capabilities. To collect question-answering data, [68] evaluated a novel framework for supporting dataset development. However, there is less research on the interaction design, perception, and adoption of Q&A chatbot services for students, as [12, 68, 73] point out.

2.3 Conversational Agent Design and Social Response Theory

Beyond improving systems' response qualities, non-functional aspects such as the design of CAs and their interactions with students have been named as crucial research avenues to be explored [48, 68]. Explicit design cues have to be considered and deployed in order to allow for an effortless, flexible, and natural interaction. CA design is concerned with how design features regarding the embodiment, the interaction, and communication structure, as well as the anthropomorphization of CAs can be tailored towards particular contexts, tasks, and user needs [17].

In fact, current literature on CAs and question-answering agents for education neglects a design perspective that considers a question-answering agent for a particular learning scenario [77]. Empirical findings from other CA deployment contexts illustrate how little modifications of CA design features can have a detrimental impact on users' perception, behavior, and performance [18, 69]. Extant literature has sparsely explored both the use of CAs for information retrieval tasks in an educational setting and, most importantly, the specific design requirements that arise in this context. Compared to traditional information retrieval and question-answering systems, CAs' underlying design enables an adaptive, natural, and social interaction. Thanks to developments in both dialogue system and visual embodiment areas, fluid conversational interactions and visual embodiment of CAs became possible from the 1990s onwards [64]. More recently, advances in ML and artificial intelligence (AI) techniques enhanced further the competencies and skills, as well as perceived naturalness, of CAs, enabling such systems to take on an increasingly human-like interaction and communication style [62, 65].

Past research has demonstrated how design cues, such as visualizations, for instance, make users perceive CAs as a character with personality, which has positive effects on the overall interaction [43]. In fact, HCI studies have revealed that psychological principles come into play when people interact with computers. CAs exhibiting human behavior can improve user acceptance of and trust in CAs [76] as user responses to systems are subconsciously triggered by social signals or behavioral cues. Previous findings point towards peoples' inclination to respond socially to a human-like object, i.e., animals but also technology, a phenomenon coined social response theory [42, 43].

The idea that humans apply social heuristics and respond to ML-based systems such as CAs in an equally "social manner" as compared to humans might also address the perfection schema, which refers to humans' desire for perfect prediction and the assumption that technological artifacts work perfectly. When a machine exhibits human behavior or traits, however, users are more willing to accept mistakes occasionally [7, 47]. As a result, social cues offer the potential to improve users' trust in, acceptance of, and experience with a system [42].

At the same time, it is unclear to what extent social response theory holds for the use case and task at hand, namely an information retrieval task in an educational setting with students as users. While much research assumes that human–human interaction and HCI are comparable [42], certain differences between the two interaction types should be considered [62]. For example, in human-to-human conversations, sender and receiver can rely on past and contextual information as the conversation progresses. In an HCI, this is only possible to a certain extent (i.e., by referring back to information from the chat history). More so, Mou and Xu [40] found that users' self-disclosure in terms of communication style is different for the two interaction types, with users being less extroverted and disclosing less information with CAs. More so, anthropomorphizing a CA can raise feelings of uncanniness and evoke inappropriate expectations regarding agency and capabilities of the CA [28, 34, 48].

Many commercially available CAs rely on social cues to increase user engagement and establish trust. Some of the most widely used social cues include the use of both verbal and non-verbal communication, i.e., emojis [59], and the agent's identity, including an avatar name and face [3]. Applying our understanding from social response theory to CAs, the design and related cues of CAs are crucial in developing and deploying useful, engaging, and trustworthy systems that consider users' cognitive, social educational, and emotional concerns [65]. Social response theory has been successfully applied to user-centered design [16, 27]. Until now, however, research on the potential of embedding social cues in CAs in educational settings has been scarce [79]. For the design of a question-answering agent, these findings hold several implications. As HCIs are usually shorter in terms of duration for information retrieval, CAs should be designed for an efficient procedure, especially for goal-oriented tasks such as information retrieval. Furthermore, a user's goals in terms of task, communication, and relationship should be defined when designing a user-centered CA [40]. In addition, users' mental models or beliefs about an interface should be understood, as they have a critical impact on the user experience and thus influence their perception of and interaction with the CA [62].

3 DESIGNING A USER-CENTERED CONVERSATIONAL AGENT FOR INFORMATION RETRIEVAL IN EDUCATION

To investigate how to develop and implement a user-centered design for CAs and how such a user-centered design affects perceptual and behavioral task outcomes,we designed and built Hermine. Hermine is a CA for information retrieval tasks in educational settings. Hermine exhibits two main components: a user-centered conversational interface and question-answering models embedded in the back end. As illustrated in Figure 1, the basic user interaction with Hermine allows users to retrieve information on course-related content and pose their own questions.

In order to build a CA for an information retrieval task in an educational setting, we followed two subsequent procedures, namely a theory-driven top-down approach and a user-centered bottom-up approach following the build-measure-learn paradigm [52]. The build-measure-learn paradigm encompasses an integrative design approach for creating a certain application (build), rapidly evaluating it with the end user (measure), and deriving theoretical and practical implications from the evaluation (learn) in order to (a) better understand users' needs and interaction behavior, and to (b) address these for an optimized user experience.

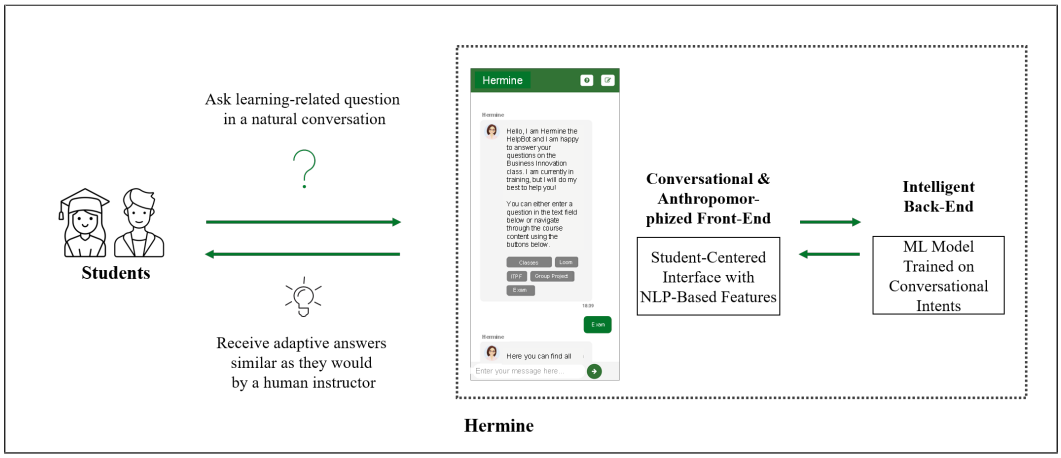

Fig. 1. Basic user interaction concept of Hermine: students ask learning-related questions and receive adaptive answers in a natural conversation based on trained intents, similar as they would by a human instructor.

3.1 Deriving Design Requirements from Literature

For the theory-driven approach, we followed [14] and [72] in conducting a systematic literature review with the aim to derive theory-based requirements for the design of an adaptive CA for question-answering queries in large-scale educational settings. We initially focused our research on studies that demonstrate the successful implementation of question-answering agents in education. Two broad areas for deriving requirements were identified: human–computer interaction and educational technology. We only included literature that deals with or contributes to a conversational knowledge retrieval system in education. On this basis, we selected 41 papers for more intensive analysis. We grouped related requirements in these contributions into four clusters as literary issues, which served as theory criteria for our design. The four clusters enclose the literature streams of (1) social interaction and emotional intelligence of computers [15, 42], (2) usability and design of information retrieval systems (e.g., [45, 78]), (3) information overload and filtering [30], and (4) learner-centered design [67].

3.2 Deriving Design Requirements from Users

Building upon the derived literature issues and meta-requirements, we subsequently followed a user-centered bottom-up design approach. We conducted 13 semi-structured interviews with students to build an initial understanding of the needs and requirements of potential users regarding course information retrieval in general, and for a CA for question-answering in specific [26]. The interviewees were randomly selected students from our university, and each interview lasted around 20 to 35 minutes. The participants were between 23 and 28 years old; ten were female, three were male. Our interview guideline consisted of 23 questions about the interviewees' experience with CAs, information retrieval in educational settings, and requirements and expectations for a CA in educational settings. Additionally, we asked participants to draw the design and interaction with a CA in an academic setting. The overall goal of the interviews was to understand users' needs and expectations regarding a CA for information retrieval in an academic course setting.

To evaluate the interviews, a qualitative content analysis was conducted. We thereby followed the methodological structure proposed by [25] and [35]. Based on an open coding system, we added

descriptive names, namely "codes" while reviewing the qualitative answers for each interview question along all interviewee answers. The coding hereby enabled us to examine the expectations of the students regarding the design, interaction, and functionality of the CA for relations, similarities, as well as differences. The labeling of information followed constructed codes by the analyzer, mainly relying on academic terms identified in our systematic literature review. Based on the interviewees' answers, we developed user stories [13] and later mapped them onto specific design requirements. While the user stories are exemplary expectations and thoughts of the students, the design requirements represent categories of similar, grouped codes[60]. To ensure the validity of our analysis, user stories and requirements were developed by the researcher who also conducted the interviews, which were later discussed and confirmed by a second researcher. Our results from the user design requirement analysis are in line with the design of extant conversational agents deployed in education [54, 74, 75].

We designed three low-fidelity prototypes of Hermine (e.g., screenshots and click-based mockups) to pretest our design instantiations and to receive an initial understanding of the human–computer interaction with a CA for large-scale educational information retrieval tasks. We pretested these design instantiations of Hermine in qualitative studies with a total of ten students, including eight graduate and two undergraduate students between the ages of 21 and 28. Again, the participants from the pretest exercise were randomly selected students from the university and differed from the individuals who participated in the interviews. This procedure allowed us to receive unbiased feedback on the preliminary design requirements as well as to corroborate expressed expectations regarding the design with actual impressions. In addition, this pretest enabled us to learn more about conversation strategies such as the conversational openings or the design of predefined buttons. For example, we experienced that students prefer to receive predefined answer buttons with possible question categories in order for them to efficiently receive an answer. Hence, we incorporated this design cue in the final design of Hermine. 80% of students mentioned that they would like to use a question-answering system that immediately provides clear and correct answers in natural language, which we incorporated in design principle 2. Eight of the thirteen students stated that they would like to use a chatbot for either administrative or organizational questions to receive quick answers for frequently asked questions (Table 1, design principle 2). From the responses of a majority of students (60%), we derived the following user story: "As a student, I want a question-answering system that can respond to my needs and provide me with the best possible support in answering my questions so that I am not left on my own when looking for information". This implies that the system needs to understand the individual needs of each user and needs to be able to help answer the questions (Table 1, design principle five). Another requirement which we derived, is the applicability of the question-answering system towards different educational topics and different semester levels (Table 1, design principle 3). Five students expressed their desire for a question-answering system that is easily accessible and usable on different devices, which we incorporated in design principle 4.

3.3 Deriving Design Principles

Consolidating our key findings from both our literature search and our empirical user interviews, we arrive at five overarching design principles. These design principles underline users' need for convenience and support and point towards the anthropomorphization of the CA through social cues (displaying human competences such as empathy) or the use of an avatar. We provide an overview of the design principles and the informing literature and user requirements in Table 1. The design principles are instantiated as functionalities in the final version of Hermine.

	Design Principle	Exemplary Literature Requirements	Exemplary User Requirements
(1)	The CA should be adapted to a specific educational setting, empathetic, and displayed as an avatar to provide a user-friendly experience.	[15, 42]	As a student, I want a system that is fun, intuitive, and convenient to use in the context of my specific course.
(2)	The CA should filter the most frequently asked questions, provide predefined "question" buttons, appropriate answers, and further useful information for students.	[30, 45, 78]	As a student, I want a question-answering system that is easy to use so that I can get correct and clear answers quickly. I would like to have a question-answering system for administrative and organizational questions, which can provide frequently asked questions and easy-to-find information so that I receive answers as quickly as possible.
(3)	The CA should be easily accessible for students and should be designed according to the corporate identity of the university to provide a seamless experience.	[67]	As a student, I want to have a question-answering system that is designed for the corresponding semester and for different areas, such as the study program, lecture, and general information so that I am sure to get answers to specific questions.
(4)	The CA should be deployed as a web-based application with an intuitive and responsive UX that is convenient to use on different devices.	[45, 67, 78]	As a student, I want a question-answering system that is easy to find and works on all my devices so that I can use it without additional effort.
(5)	The CA should provide the user with quick answers in natural language, should understand the individual needs, and provide a feedback and help button.	[45, 67, 78]	As a student, I want a question-answering system that can respond to my needs and provide me with the best possible support in answering my questions so that I am not left on my own when looking for information.

Table 1. Overview of derived design principles on how to build a learner-centered CA for question-answering information retrieval in a large-scale educational context.

3.4 Developing a User-Centered Conversational Interface

3.4.1 Translating Design Principles into System Functionalities. Based on our design principles, we built Hermine as a cloud-based web application that can be used on all kinds of devices, including desktop and mobile versions (F1). A screenshot of Hermine and its different functionalities (e.g., *F1 - 6*) can be seen in Figure 2. The interaction with Hermine is initiated by the CA itself by proactively offering students help with finding information on learning and course-related matters. The agent is designed in the corporate identity of the students' university (F2). In general, the persona of Hermine employs a friendly conversational style including colloquial yet sufficiently professional language (F6). Besides giving the agent the human-like name Hermine, we used a human-like avatar to support students in empathizing with the persona of the CA. Hermine provides students with an overview of predefined question clusters (F3), including administrative, content-related and examination-related questions. The predefined answer buttons help the user to receive an overview of different question categories and ensure both flexibility and efficiency regarding information retrieval and interaction with Hermine. The user always has the option to receive an introduction to the interaction structure or additional help when interacting with Hermine (see Figure 2, F4). Also, Hermine is provided with an always-present feedback button (see Figure 2, F5) for users to provide feedback about the course or Hermine itself whenever they want to. Throughout the interaction, the users always have the choice to answer with predefined buttons or with freely written text (e.g., F6). [1]

3.4.2 Developing and Modeling Question-Answering Intents. The web application of our CAs is developed in HTML5 with CSS. The front end JavaScript is connected to a Python script that a) processes incoming user intents and b) provides predefined answers based on the incoming classifications. We rely on the web framework Flask as our back end to easily embed a web application, namely our CA, in Python. For the conversational logic of a student–educator talk, we modeled 70 intents, including the introduction of the conversation, frequently asked questions, and casual dialogue. We collected FAQ questions from the course lectures over the past two years and built a database of core course questions. This database resulted in 49 frequently asked questions that we enriched with our user interviews. With this basis for our intent modeling, we ensured that our intent classification fitted the desired user interaction and dialogue in terms of language and conversation style. For instance, next to an exam button, we included the following queries: "When is the exam taking place?" or "What is the duration of the exam?" to the defined intent "exam". In addition, we added a variation of approximate utterances and user queries for each intent as done in [38, 41]. The main aim of our intent modeling was to cover course content that would otherwise be available via a standard FAQ document of the university's course.

3.4.3 Training and Deploying a Question-Answering Model for a Conversational Agent. The intents were trained on a "Naive Bayes classifier" in combination with semantic similarity matching as, e.g., also done in [54]. The classifier is trained on the 70 developed intents with the confidence score of the classifier set at 90%. As a probabilistic model, answers are predicted on the basis of the probability of an intent (e.g., a user formulating a question in the same or similar way as the intent) with the confidence score determining the quality of the classification model for whether a user statement belongs to a specific intent [41]. Our intent modeling confidence thus represents a barrier to providing an answer to a specific user request. As compared to lower model confidence, our CA classifies answers more correctly (i.e., providing an output that correctly answers the

[1] We are only able to display selected design functionalities of Hermine. For more insights, e.g., into other instantiated functionalities, such as the conversational introduction, the conversational closing, the adaptivity, or the casual chat mode, the interested reader might refer to the interaction videos in the appendix.

user's request), yet might recognize fewer intents (i.e., being unable to provide a suitable answer and thus asking the user to reformulate his or her question). Thanks to semantic similarity, we can define a metric over a set of terms, which allows our conversational agent to identify user requests and questions that might not be fully congruent with our coded intents yet exhibit a high likeness of meaning or semantic content. The conversational back end is implemented by utilizing the frameworks *chatterBot*[2] and *spaCy*[3].We rely on the Python library spaCy for natural language understanding of users' requests, whereas ChatterBot is a natural language processing library that processes the input statements and gives responses to a user's input through so-called "Logic adapters". More specifically, the conversational dialogue is ensured by ChatterBot giving the response with the highest calculated confidence value for a given input statement. Since our course and interaction with students is done in German, ChatterBot's language independency represents a key advantage over other libraries. We created a webpage to run our CAs using HTML5 and CSS code as done in [38] where we configured our Python yml file to activate and run our Python server and the respective CA domains. We validated our training data to ensure that there are no bugs or inconsistencies in our domain, our NLP pipelines, or responses. Two researchers ran an informal test data collection of around 100 sentences to assess the acceptable accuracy of CA responses (i.e., the proportion of correctly understood intents to the total number of intents) [41], with one of those two researchers being familiar with the course content already. The researchers came up with intents independently of each other. During the test runs, the CA accuracy exceeded 90%, which we deemed as acceptable to proceed with our experimental validation of the CA.

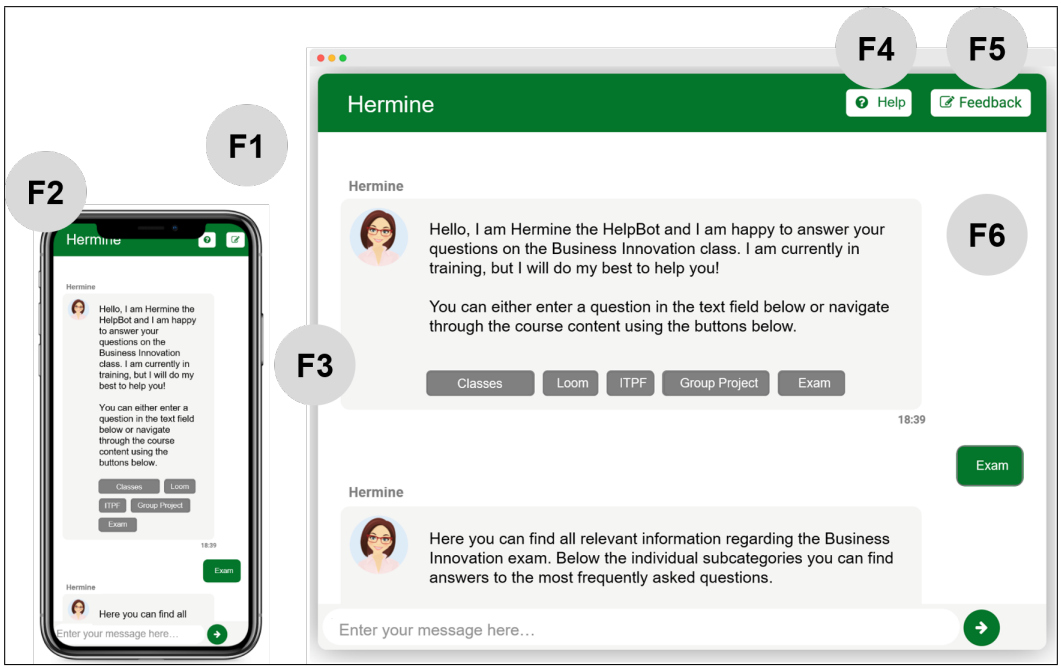

Fig. 2. Screenshots of our adaptive conversational question-answering system for education: a learner receives answers to learning-related questions through natural and adaptive interaction.

[2]https://chatterbot.readthedocs.io/en/stable/
[3]spacy.io

4 EXPERIMENT

To evaluate (1) our approach for a user-centered CA design for an information retrieval task in an educational setting and (2) our hypothesis that a learner-centered CA will improve students' ability to retrieve learning-relevant information and influence their interaction experience, we conducted an online lab experiment as part of which participants were asked to answer learning- and course-related questions. Our empirical evaluation was concerned with how effective a user-centered CA is in helping learners retrieve course-relevant information and how learners perceive the interaction with such a CA. Deploying a 3 (traditional question-answering system, non-design-driven CA, learner-centered CA Hermine) x 1 between-subject design, participants thus had access to different tools to retrieve relevant information dependent on the treatment they were randomly assigned to. The traditional question-answering system, as well as the two CAs, were all designed in German. To allow for a better understanding, we translated visualizations and examples of the respective user interfaces into English (e.g., see Figure 2). Specifically, we hypothesized that

H1: *Interacting with a user-centered CA as compared to a traditional question answering system and a non-design driven CA has a positive effect on users' performance in multiple choice and open-ended information retrieval questions.*

H2: *Interacting with a user-centered CA as compared to a traditional question answering system and a non-design driven CA leads to greater levels of perceived trust, social presence, and enjoyment.*

4.1 Design and Treatment Groups

To evaluate Hermine, we compared it with a traditional question-answering system in the form of an overview of frequently asked questions (FAQ) on a website and an alternative, non-design-driven CA. This allowed us to compare our conversational and adaptive CA Hermine to both a non-conversational question-answering tool and a CA that did not consider the user-centered design principles. To control for differences in the designs of the tools, we also built the traditional question-answering tool and the second CA ourselves. The two CAs were built based on the same back end and relied upon the same ML model trained on 70 intents, allowing the two conversational tools to provide adaptive and personalized answers to the user. We only manipulated the instantiation of our five design principles of the human-computer interaction in the front end. The traditional question-answering system listed frequently asked questions with subsequent explanations for those questions in a flowing text format. The provided question and answer content of all three tools was exactly the same. The traditional question-answering system and both CAs were accessible via a website link each.

4.2 Study Procedure and Perceptual Measures

The experiment consisted of three key parts: (1) a pretest collecting user dispositions, (2) an information retrieval task, and (3) a post-test measuring users' perception of the interaction. The pretest and post-test phases were consistent for all participants. As part of the information retrieval task, TG1 used Hermine to find specific information on learning- and course-related questions, whereas participants of the CG used a FAQ-document-based tool, the standard tool for students to find information. TG2 used an alternative CA that did not include our user-centered design principles.

(1) Pretest: The experiment started with a short introduction of the overall procedure of the experiment and a pre-survey of seven questions including an attention check. Participants had to explicitly agree to the nature of and involvement with the experiment before proceeding with the pre-survey questionnaire items. Here, we tested two different constructs. First, we asked for users' personal innovativeness regarding information technology based on four items according

to [1] (sample items: *"I like to experiment with new information technologies"* or *"If I heard about a new information technology, I would look for ways to experiment with it,"*; 7-point scale, from 1: "Strongly Disagree" to 7: "Strongly Agree", $\alpha_{PersonalInnovativeness}$ = .8). Second, we tested users' dispositional trust in technology following [23], (sample item: *"I usually trust a technology until it gives me a reason not to trust it"*; 7-point scale, from 1: "Strongly Disagree" to 7: "Strongly Agree", $\alpha_{TrustingDispositions}$ = .93).

(2) Information Retrieval Task: As part of the experiment task, participants had to answer questions about the structure and content of a university course. In total, the information retrieval task was compromised of four multiple-choice questions and two open-ended questions (sample questions: *"What date is the final exam?"* and *"Which course deliverables are graded? Please provide a short overview of all deliverables and how they are composed."*). TG1 used our learner-centered CA Hermine to find course-relevant information, whereas TG2 had access to an alternative CA that was not built on our developed design principles. The CG used a traditional question-answering tool. We did not provide any introduction to any of the tools. Students using one of the two CAs retrieved individual answers by interacting with the respective system. Participants using the traditional question-answering system retrieved information by searching through the document. The mean time of task completion was 16.25 minutes (SD = 10.12).

3) Post-test: As part of the post-survey, we included 15 items to assess participants' perception of the question-answering system and the information retrieval interaction, and to control for manipulation. We measured trust in the information provided by the question-answering system (scale adapted from [36], sample item: "To me, the question-answering system is generally accurate in providing information."; 7-point scale, from 1: "Strongly Disagree" to 7:"Strongly Agree", $\alpha_{TrustinInformation}$ = .86). Moreover, we asked participants to report how much they enjoyed the interaction with the respective technology (scale adapted from [31], sample items: *"The interaction with the question-answering system is exciting"* and *"I enjoy finding information with this question-answering system compared to another tool"*; 7-point scale, from 1: "Strongly Disagree" to 7:"Strongly Agree", $\alpha_{Enjoyment}$ = .91). We measured users' perceived social presence of the question-answering system (scale adapted from [51], sample items: *"There is a sense of intimacy in the question-answering system"* and *"There is a sense of sociability in the question-answering system"*; 7-point scale, from 1: "Strongly Disagree" to 7:"Strongly Agree", $\alpha_{SocialPresence}$ = .94). Lastly, we posed three qualitative questions: "*What did you particularly like about the use of the question-answering system?*", "*What else could be improved?*", and "*Do you have any other ideas?*", and captured participants' demographics including gender, age, and nationality.

4.3 Behavioral Measures

Besides measuring self-reported user perceptions in the post-test, our objective was to measure the behavioral performance of the participants, namely the correctness of answers in the information retrieval task in order to evaluate our main hypothesis. To do so, we considered participants' answers to the course-related questions.

1) Correctness of multiple-choice questions: The closed multiple-choice questions were analyzed for the correct answer provided. Since only one out of multiple answer options was correct, answer correctness was measured as either correct (1) or incorrect (0). This classification was conducted for each of the four multiple-choice questions in order to identify potential differences in answer correctness among the three treatments. In addition, an overall mean score (range 0-1, 1 = highest) across the four multiple-choice questions was calculated.

2) Correctness of open-ended questions: The correctness of the open-ended course questions was analyzed for accuracy and completeness of the answer provided. Participants were not limited in the content and the amount of the written answers. We applied a grading scheme with four levels

(*(1) completely incorrect, (2) partially correct, partially incorrect, (3) correct, yet missing information, (4) completely correct*). An example classification of potential answers was developed by a first annotator and confirmed by a second one to arrive at a shared understanding of what type of answer would apply for a *"completely correct"* grading, for instance. For each open-ended question, the annotators agreed upon which keywords should be mentioned. An answer was *completely correct* when all keywords were listed and when a supporting explanation was given. If no full explanation was given, annotators labeled the answer as *correct, yet missing information.* Third, an answer was classified as *partially correct, partially incorrect* if false information was given or numerous keywords were missing. Anything else was considered *completely incorrect.* We relied on the same two annotators who independently judged the two open-ended questions on the developed four-level grading scheme. Grading was cross-checked for a consistent evaluation. Answers for which the two annotators diverged on in their respective gradings were revisited and discussed until a final grading was arrived at. Similar to our procedure for the multiple-choice questions, we also calculated the mean average score for the two open-ended questions (range (1-4, 4 = highest).

4.4 Data Collection and Cleaning

We recruited 82 students from our university to take part in our experiment. The experiment was conducted as a web experiment designed and executed according to the ethical guidelines of the university. Participation in the experiment was voluntary and not (financially) rewarded. Data points that failed the attention or manipulation check, or participants who did not complete the experiment task fully, were removed. Ultimately, we counted 41 total valid results, with 13 in treatment group 1 (TG1 - Hermine) and 14 each in the control group (CG - traditional question-answering system) and treatment group 2 (TG2 - basic CA). Participants had an average age of 27.71 years (SD = 7.96), 19 were female, 22 were male. There were no significant differences among the three groups regarding age or gender ($p > .1$).

Before running our main statistical analysis, we conducted two relevant data assumption steps. First, to ensure randomization and to control for potential effects of confounding variables with our small sample size, we ran additional analyses on the control and demographic variables. There were no significant differences in trusting disposition and personal innovativeness among the three treatments (all $p > .1$).

Our study and related statistical analysis exhibit a small sample size N. We discarded a large number of data points within our initial sample to prioritize a rigorous data cleaning over a large sample size. The relatively large number of removed data points may be due to voluntary participation in the experiment. We evaluated our CA and its design principles with other participants beforehand and tested for a normal distribution of our final data, thereby strengthening the listed statistical results. We controlled for normal distribution by visually checking the distribution of our data through density and q-q plots. In addition, Shapiro Wilk tests confirm assumptions of normality for the distribution of our data ($p > .05$).

5 RESULTS

To assess our hypothesis that a user-centered CA in comparison to a non-design-driven CA and a traditional question-answering system will improve students' ability to retrieve course-relevant information and influence their interaction experience, we conducted a number of statistical tests.

5.1 Results Regarding Learners' Ability to Retrieve Course-Related Information

To evaluate our first hypothesis, we compared the correctness of answers provided for each course content question among the three treatments. Moreover, we studied the answer correctness across all four quantitative and across the two qualitative questions.

Group	N	**MC Q1** % correct (SD)	**MC Q2** % correct (SD)	**MC Q3** % correct (SD)	**MC Q4** % correct (SD)	**O-E Q1** mean; 1-4 (SD)	**O-E Q2** mean; 1-4 (SD)
CG: Traditional question-answering system	14	100.0% (0.0)	64.28% (0.497)	50.0% (0.519)	42.86% (0.514)	3.64 (0.49)	2.57 (0.94)
TG1: User-centered CA Hermine	13	100.0% (0.0)	100.0% (0.0)	92.31% (0.277)	84.62% (0.376)	3.54 (0.97)	3.62 (0.51)
TG2: Non-design-driven CA	14	42.86% (0.51)	78.57% (0.426)	57.14% (0.514)	57.14% (0.514)	2.86 (0.66)	2.71 (0.73)
Significance		*** p < .001	. p < .1	* p < .05	. p < .1	* p < .05	** p < .01
Tukey post hoc test	41	TG2-CG *** p < .001, TG1-TG2 *** p < .001	TG1-CG . p < .1	TG1-CG . p < .1	TG1-CG . p < .1	TG2-CG ** p < .01, TG1-TG2 . p < .1	TG1-CG ** p < .01, TG1-TG2 ** p < .01

Table 2. Overview of results (mean and standard deviation (SD)) on learners' ability to retrieve learning- and course-related question in terms of answer correctness for multiple choice (MC Q) and open-ended questions (O-E Q) across the three groups

We found that students who used Hermine had a significantly higher level of correct answers for both the multiple-choice and the open-ended questions. More specifically, we conducted an ANOVA comparing participants in the three conditions regarding their performance across the first four multiple-choice questions. The test revealed a significant main effect ($F(2, 38) = 14$, $p < .001$, $\eta^2 = 0.42$). In fact, participants who interacted with the conversational and adaptive question-answering system Hermine performed significantly better across the multiple-choice questions compared to the traditional question-answering tool ($M_{UserCenteredCA} = .942$, $M_{TraditionalQA} = .643$, $t = 0.30$, Cohen's $d = -1.62$, $p < .001$) and the CA that did not incorporate any of our design principles ($M_{UserCenteredCA} = .942$, $M_{NonDesignCA} = .589$, $t = 0.35$, Cohen's $d = -2.31$, $p < .001$) with a scale from 0 (incorrect) to 1 (correct) and large effect sizes for both t-tests.

Our subsequent ANOVA analysis shows a significant effect of our question-answering system manipulation on correctness for open-ended questions, too ($F(2, 38) = 6.278$, $p < .01$, $\eta^2 = 0.24$). On the overall correctness scale for the open-ended questions (1: Completely correct; 4: Completely incorrect), a post Bonferroni test reveals that participants interacting with the TG1 question-answering system, Hermine, performed significantly better compared to participants interacting with the non-design-driven question-answering system of TG2, with this t-test having a large effect size ($M_{UserCenteredCA} = 3.58$, $M_{NonDesignCA} = 2.79$, $t = 0.79$, Cohen's $d = -1.47$, $p < .01$).

Beyond the previously reported results, Table 2 demonstrates the same statistical analysis for each question. Participants interacting with Hermine also performed significantly better in retrieving course-related information for each of the individual six questions. For both the multiple-choice and the open-ended questions, our results do not exhibit any significant differences between TG2 and the traditional question-answering tool, except for the first open-ended question. The results demonstrate that users interacting with Hermine performed significantly better in both types of information retrieval questions. There is no significant difference among groups with regard to task

Group	N	**Trust in Provided Information** Mean; scale: 1-7 (SD)	**Level of Enjoyment** Mean; scale: 1 - 7 (SD)	**Social Presence** Mean; scale: 1 - 7 (SD)
CG: Traditional question-answering system	14	4.55 (1.15)	2.97 (0.81)	2.57 (0.79)
TG1: User-centered CA Hermine	13	5.59 (1.03)	5.65 (0.92)	5.23 (0.79)
TG2: Non-design-driven CA	14	4.40 (0.68)	3.77 (1.39)	3.34 (1.49)
Tukey post hoc test	41	TG1-CG * p = .05, TG1-TG2 ** p < .01	TG1-CG *** p < .001, TG1-TG2 *** p < .001	TG1-CG *** p < .001, TG1-TG2 *** p < .001

Table 3. Overview of results regarding user experience for perceived trust in information and level of enjoyment among the three groups.

completion time ($p > .1$), demonstrating that with a comparable task completion time, participants using Hermine were more efficient in retrieving relevant information.

5.2 Results Regarding User Experience in Terms of Enjoyment, Trust, and Social Presence

To evaluate users' subjective perception of the question-answering systems and to test hypothesis 2, we compared the constructs of trust in information and level of enjoyment between participants using the learner-centered CA Hermine and participants using the alternative two tools.

As can be seen in Table 3, participants' perceptions regarding those two constructs varied significantly across the three treatment groups. We compared users' trust in the information provided by the respective question-answering system, ultimately finding a significant main effect ($F(2, 38) = 5.89$, $p < .01$, $\eta^2 = 0.24$). In fact, users interacting with Hermine to retrieve specific course information trusted the provided information significantly more compared to users interacting with the conversational question-answering that was not designed according to our design principles ($M_{UserCenteredCA} = 5.59$, $M_{NonDesignCA} = 4.4$, $t = 1.18$, Cohen's $d = -1.36$, $p < .01$) and compared to users interacting with the traditional question-answering tool ($M_{UserCenteredCA} = 5.59$, $M_{TraditionalQA} = 4.55$, $t = 1.04$, Cohen's $d = -0.95$, $p < .05$). Reported post hocs are of large effect size. In a similar vein, we find a significant effect of the question-answering system treatment on users' level of enjoyment of the respective system ($F(2, 38) = 21.78$, $p < .001$, $\eta^2 = 0.53$). A post hoc test of the CAs with and without consideration of our design principles ($M_{UserCenteredCA} = 5.65$, $M_{NonDesignCA} = 3.77$, $t = 1.87$, Cohen's $d = -1.57$, $p < .001$). as well as of the CA that was made according to our design principles ($M_{UserCenteredCA} = 5.65$, $M_{TraditionalQA} = 2.97$, $t = 2.67$, Cohen's $d = -3.09$, $p < .001$) are statistically significant and of large effect size. A third ANOVA ($F(2, 38) = 21.57$, $p < .001$, $\eta^2 = 0.53$) and subsequent post hocs illustrate that users interacting with Hermine rated the CA as higher in social presence compared to participants interacting with the traditional question-answering tool ($M_{UserCenteredCA} = 5.23$, $M_{TraditionalQA} = 2.57$, $t = 2.66$, Cohen's $d = -3.4$, $p < .001$) and the CA that did not incorporate the design principles ($M_{UserCenteredCA} = 5.23$, $M_{NonDesignCA} = 3.34$, $t = 1.89$, Cohen's $d = -1.57$, $p < .001$). Both tests depict large effect sizes.

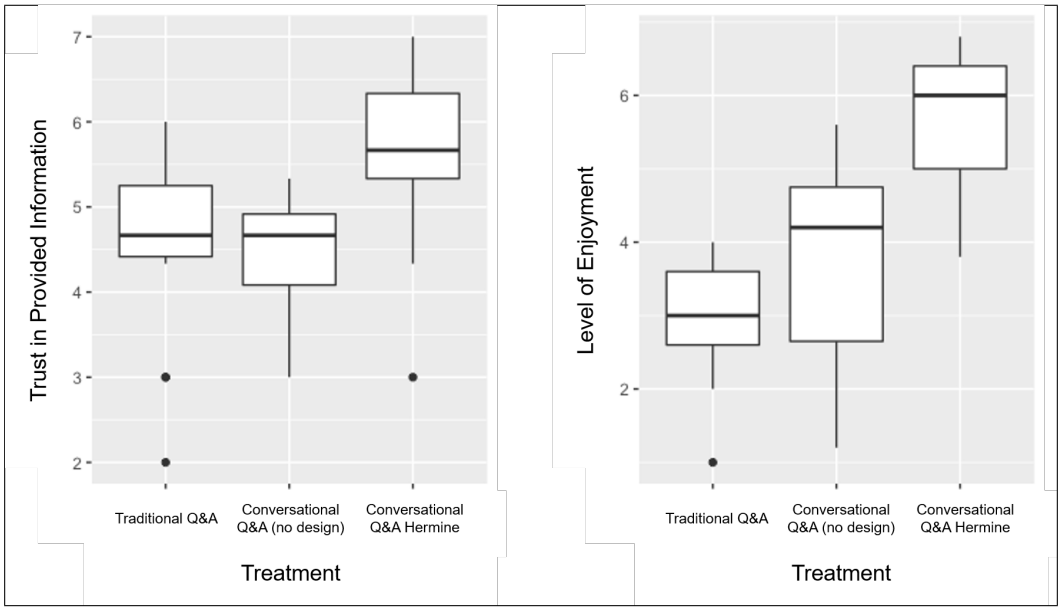

Fig. 3. Results regarding user-reported trust in provided information (left) and level of enjoyment (right) across the three different question-answering systems deployed in the experiment.

Interaction Aspect	Feature
On engagement and interactivity	*"The interaction with the system was very playful. I was guided through the questions yet could come back to the main menu at any time. The system is very intuitive and clear."*
On speed of the agent	*"I liked the fast response it provides. If I have questions that I can't solve with other resources, I get quick solutions to my problems."*
On agent and conversation style	*"I like the name [Hermine] and think that the question-answering system is using friendly language."*
On graphics	*"I liked that you had the option to either click predefined buttons or freely enter your questions."*
On context of use	*"I would like to use such a tool in my upcoming classes."*

Table 4. Representative user responses on interacting with CA Hermine.

5.3 Qualitative Results

For a more nuanced understanding of users' perception of and thoughts on the learner-centered CA Hermine, we further surveyed all students' opinion on Hermine through a few open questions. Participants positively mentioned the design that was adapted according to the educational institution's branding as well as the simple and effective answer provision of the CA. While Hermine provided answers quickly, one user criticized the loading time at the beginning of the interaction. In general, users enjoyed interacting with Hermine and the interaction options it provided, as illustrated by representative user responses in Table 4. In summary, these findings indicate an effective design instantiation of our natively built CA.

6 DISCUSSION

Our study illustrates the importance of conversational question-answering for an educational setting, namely an information retrieval task on course-related content. We aimed to address previously illustrated research gaps by deriving user-centered and theory-motivated design principles for the interaction design of question-answering agents in large-scale educational settings. Beyond such requirements, we empirically investigated our design principles on students' behavior and user perception.

6.1 Theoretical and Practical Contributions

Our results demonstrate a successful implementation of the developed design principles, with participants being satisfied with the interaction. The natively designed question-answering system not only positively affects users' ability to retrieve learning-related information (H1) but also positively influences students' perception of the question-answering system (H2). Interacting with Hermine improved learners' performance in the multiple choice questions as compared to interacting with the non-design-driven CA and the traditional question-answering system, whereas we could find a significant difference in open-ended question performance between the two CA treatments only (TG1 and TG2). Trust in information, enjoyment, and social presence attributed to the system were all significantly higher for the users interacting with the user-centered CA Hermine as compared to the other two treatments. Ultimately, our findings illustrate how a user-centered CA can enhance task outcomes and user experience for an educational task currently still relying on impersonal and non-adaptive systems.

Our question-answering system Hermine and its empirical evaluation make several contributions to research. First, the CA offers the potential to overcome limitations associated with traditional question-answering tools by allowing students to receive specific information faster while reducing the burden of educators having to answer large amounts of repetitive questions. We expect this to positively affect student satisfaction and learning performance, as well as to reduce dropout rates in large-scale learning scenarios where individual support is only possible to a limited extent due to organizational and financial resources. Second, our study contributes to design knowledge on how to design and deploy CAs in education as question-answering systems. Developed design principles and features confirm the current understanding from literature on social response theory and HCI. Our instantiation of these design principles could enable educators and institutions to develop CAs. More so, our derived design understanding could be extended to domains other than education, i.e., organizational or customer service settings, where fast access to specific information is relevant. It is important to note, however, which principles would need to be considered or adapted for other domains.

6.2 Limitations and Future Work

Our findings should be interpreted with caution as the current study suffers from several limitations. First, participants were able to interact with the agent without boundary conditions such as time pressure or personal involvement in the task. Although study participation was voluntary, motivation among participants was not necessarily consistent which represents a potential confound of our study. Second, we interpreted the effectiveness of our CA based on the correctness of six answers to course-related questions. It begs the question to what extent our results hold when deploying Hermine for more complex questions, varying contexts, and with different users. In addition, it is unclear what the long-term effects of the CA are when being used repetitively or over a longer period of time beyond a single interaction. Our conducted experiment is associated with a homogeneous participant pool. Participants were all students and thus represented potential actual

users of the CA, with 87.8% of study participants stating that they have used a CA (i.e., Facebook Messenger Bot) before. We do not know how our question-answering system is perceived by a user group less familiar with contemporary CAs. Lastly, while we controlled for a normal distribution of our data, subsequent empirical investigations of our tool should be conducted with a larger sample size, especially when aiming to explore the deployment with varying tasks, contexts, and manipulations.

The embedding of social cues in CAs presents a double-edged sword or even a more general paradox that lies at the very heart of agency in machines. Anthropomorphization through human-like language or appearance promises to enhance human-computer interaction by providing a more natural, effortless, and personalized interaction with the user. Simultaneously, increasing anthropomorphization of a CA through social cues can influence and improve the perceived agency of a CA. Inappropriate amounts of agency might not properly reflect the CA's capabilities and competencies [48]. Reflecting on our work, we argue that this paradox is a result of the fundamental conflict between task achievement and overarching ethical questions, which should guide the design process. So far, however, this tension has largely remained abstract and unresolved, including in our work. Design features fostering system trustworthiness, e.g., explanatory statements or transparency, present an opportunity to overcome or counterbalance the multifaceted consequences of anthropomorphism. However, extant empirical work demonstrates that the implications of trustworthiness design are not as straightforward and require further exploration, in particular for CAs [57, 58].

In the context of our study, we relied on a low-stake setting with limited risk for the user and an experiment task where an increase in user trust is desirable in order to drive the adoption of novel technology. The establishment of trust as a key goal in this interaction is also based on the assumption that the information provided by the CA is fully correct and understandable. In a high-stake context and when the user should not blindly trust but challenge the provided information (e.g., when using a CA as an educational tutor), the interaction goal might change (e.g., not simply increasing but calibrating appropriate trust), and as a result also imply relevant design implications regarding the anthropomorphization of such agents (e.g., reducing agency attributed to a system).

7 CONCLUSION

As part of this study, we designed, built, and evaluated Hermine, a user-centered CA that provides students with answers to course-related questions by leveraging contemporary models of ML algorithms. We compared Hermine to both a standard question-answering tool and a basic, non-design-driven CA in an online lab experiment with 41 participants. We found that users interacting with Hermine performed significantly better in the information retrieval task as compared to the ones interacting with alternative question-answering tools. The significant trust in the information and level of enjoyment when interacting with Hermine hint at the importance of considering users' expectations and interaction experience when designing CAs for specific use contexts and domains. Overall, our findings underline the opportunity CAs present for information retrieval tasks in education, which currently still rely on cumbersome and unresponsive question-answering systems. Our results also offer design suggestions to drive trust in CAs deployed in educational and learning settings. While a certain level of social cues provides a foundation for effective interaction with CAs, future research should look into how these cues can be balanced, i.e., through system trustworthiness to ensure that a CA is not blindly used and trusted. With ML-based systems becoming more omnipresent in daily life and in learning scenarios specifically, we hope that our work will attract other researchers to design and build CAs in consideration of users' needs and interaction experience.

ACKNOWLEDGMENTS

We thank the Swiss National Science Foundation for supporting this research (grant 192718). Furthermore, we thank Leonie Haas for supporting the tool development with her thesis.

REFERENCES

[1] Ritu Agarwal and Elena Karahanna. 2000. Time Flies When You're Having Fun: Cognitive Absorption and Beliefs about Information Technology Usage. *MIS Quarterly* 24, 4 (12 2000), 665. https://doi.org/10.2307/3250951

[2] Mutaz M Al-Debei. 2014. The quality and acceptance of websites: An empirical investigation in the context of higher education. *International Journal of Business Information Systems* 15, 2 (2014), 170–188. https://doi.org/10.1504/IJBIS.2014.059252

[3] T. Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior* 85 (2018), 183–189.

[4] Z. Ashktorab, M. Jain, V. Q. Liao, and J. D. Weisz. 2019. Perceptions of chatbots in therapy. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. Glasgow, Scotland.

[5] Z. Ashktorab, M. Jain, V. Q. Liao, and J. D. Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. Glasgow, Scotland.

[6] Nicholas Belkin. 1993. Interaction with Texts: Information Retrieval as Information-Seeking Behavior. (1993), 55–66. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.6725

[7] Benedikt Berger, Martin Adam, Alexander Rühr, and Alexander Benlian. 2021. Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. *Business and Information Systems Engineering* 63, 1 (2021), 55–68. https://doi.org/10.1007/s12599-020-00678-5

[8] Christopher G. Brinton, Ruediger Rill, Sangtae Ha, Mung Chiang, Robert Smith, and William Ju. 2015. Individualization for education at Scale: MIIC design and preliminary evaluation. *IEEE Transactions on Learning Technologies* 8, 1 (3 2015), 136–148. https://doi.org/10.1109/TLT.2014.2370635

[9] J. Cambre and C. Kulkarni. 2019. One voice fits all? Social implications and research challenges of designing voices for smart devices. In *Proceedings of the ACM on Human-Computer Interaction, 3(CSCW)*.

[10] Sofy Carayannopoulos. 2018. Using chatbots to aid transition. *International Journal of Information and Learning Technology* 35, 2 (2018), 118–129. https://doi.org/10.1108/IJILT-10-2017-0097

[11] R. C. S. Chang, H. P. Lu, and P. Yang. 2018. Stereotypes or golden rules? Exploring likable voice traits of social robots as active aging companions for tech-savvy baby boomers in Taiwan. *Computers in Human Behavior* 84 (2018), 194–210. https://doi.org/10.1016/j.chb.2018.02.025

[12] Sue Inn Ch'ng, Lee Seng Yeong, and Xin Yean Ang. 2019. Preliminary Findings of using Chat-bots as a Course FAQ Tool. In *2019 IEEE Conference on e-Learning, e-Management and e-Services, IC3e 2019*. Institute of Electrical and Electronics Engineers Inc., 52–56. https://doi.org/10.1109/IC3e47558.2019.8971786

[13] Mike Cohn. 2004. *User Stories Applied For Agile Software Development*. Technical Report.

[14] Harris M. Cooper. 1988. Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society* 1, 1 (1988), 104–126. https://doi.org/10.1007/BF03177550

[15] Chris Creed and Russell Beale. 2008. Emotional Intelligence: Giving Computers Effective Emotional Skills to Aid Interaction. In *Computational Intelligence: A Compendium*.

[16] Stephan Diederich, Alfred Benedikt Brendel, and Lutz M. Kolbe. 2020. Designing Anthropomorphic and Communicative Enterprise Conversational Agents. *Business & Information Systems Engineering* 4 (2020). https://doi.org/10.1007/s12599-020-00639-y

[17] Stephan Diederich, Alfred Benedikt Brendel, Stefan Morana, and Lutz Kolbe. 2022. On the Design of and Interaction with Conversational Agents: An Organizing and Assessing Review of Human-Computer Interaction Research. *JAIS* 23 (2022), 96–138. https://doi.org/10.17705/1jais.00724

[18] Aaron C Elkins and Douglas C Derrick. 2013. The Sound of Trust: Voice as a Measurement of Trust During Interactions with Embodied Conversational Agents. *Group Decision and Negotiation* 22, 5 (2013), 897–913. https://doi.org/10.1007/s10726-012-9339-x

[19] Sean B. Eom, H. Joseph Wen, and Nicholas Ashill. 2006. The Determinants of Students' Perceived Learning Outcomes and Satisfaction in University Online Education: An Empirical Investigation*. *Decision Sciences Journal of Innovative Education* 4, 2 (7 2006), 215–235. https://doi.org/10.1111/j.1540-4609.2006.00114.x

[20] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human Computer Studies* 132 (12 2019), 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

[21] L. K. Fryer, M. Ainley, A. Thompson, A. Gibson, and Z. Sherlock. 2017. Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. *Computers in Human Behavior* 75 (2017),

461–46.
[22] A. Følstad and P. B. Brandtzæg. 2017. Chatbots and the new world of HCI. *Interactions* 24, 4 (2017), 38–42. https://doi.org/10.1016/j.chb.2018.02.025
[23] David Gefen and Detmar W Straub. 2004. Consumer trust in B2C e-Commerce and the importance of social presence: Experiments in e-Products and e-Services. *Omega* 32, 6 (2004), 407–424. https://doi.org/10.1016/j.omega.2004.01.006
[24] Justin Scott Giboney, Susan A. Brown, Paul Benjamin Lowry, and Jay F. Nunamaker. 2015. User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit. *Decision Support Systems* 72 (4 2015), 1–10. https://doi.org/10.1016/j.dss.2015.02.005
[25] J. Glaeser and G. Laude. 2010. *Experteninterviews und qualitative Inhaltsanalyse : als Instrumente rekonstruierender Untersuchungen*. VS Verlag fuer Sozialwissenschaften. https://link.springer.com/book/9783531172385
[26] Jochen. Glaser and Grit. Laudel. 2010. *Experteninterviews und qualitative Inhaltsanalyse : als Instrumente rekonstruierender Untersuchungen*. VS Verlag fur Sozialwiss. http://www.springer.com/de/book/9783531172385
[27] U Gnewuch, S Morana, M Adam, and A Maedche. 2017. This is the author ' s version of a work that was published in the following source Please note : Copyright is owned by the author and / or the publisher . Commercial use is not allowed . Institute of Information Systems and Marketing (IISM) The psychop. *Thirty Eighth International Conference on Information Systems, South Korea 2017* December (2017), 0–11.
[28] U. Gnewuch, S. Morana, M. T . P . Adam, and A. Maedche. 2018. Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. In *Proceedings of the European Conference on Information Systems (ECIS)*. Portsmouth, United Kingdom.
[29] Nicola Guarino and Christopher A Welty. 2004. An overview of OntoClean. In *Handbook on ontologies*. Springer, 151–171.
[30] Paul Hemp. 2009. Death by information overload. *Harvard business review* 87, 9 (2009), 82–9.
[31] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys effects of platform and conversational style on survey response quality. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–12. https://doi.org/10.1145/3290605.3300316
[32] Bernhard Kratzwald and Stefan Feuerriegel. 2019. Putting Question-Answering Systems into Practice. *ACM Transactions on Management Information Systems* 9, 4 (2019), 1–20. https://doi.org/10.1145/3309706
[33] S. Laumer, A Racheva, F. Gubler, and C. Maier. 2019. Use Cases for Conversational Agents : An Interview-based Study. In *Proceedings of the Americas Conference on Information Systems (AMCIS)*. Cancun, Mexico.
[34] Mori. M. 1970. The Uncanny Valley. *Energy* 7, 4 (1970), 33–35.
[35] Philipp Mayring. 2000. Qualitative content analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [On-line Journal]* 2, 1 (2000).
[36] D Harrison McKnight, Peng Liu, and Brian T Pentland. 2020. Trust Change in Information Technology Products. *Journal of Management Information Systems* 37, 4 (2020), 1015–1046. https://doi.org/10.1080/07421222.2020.1831772
[37] P. Meier, J. H. Beinke, C. Fitte, and F. Behne, A.and Teuteberg. 2019. FeelFit – Design and Evaluation of a Conversational Agent to Enhance Health Awareness. In *Proceedings of the International Conference on Information Systems (ICIS)*.
[38] Siddhant Meshram, Namit Naik, Megha VT, Tanmay More, and Shubhangi Kharche. 2021. College Enquiry Chatbot using Rasa Framework. In *Asian Conference on Innovation in Technology (ASIANCON)*. Pune, India.
[39] Fernando A. Mikic-Fonte, Martin Llamas-Nistal, and Manuel Caeiro-Rodriguez. 2019. Using a Chatterbot as a FAQ Assistant in a Course about Computers Architecture. *Proceedings - Frontiers in Education Conference, FIE* 2018-Octob (2019), 2018–2021. https://doi.org/10.1109/FIE.2018.8659174
[40] Yi Mou and Kun Xu. 2017. The media inequality: Comparing the initial human-human and human-AI social interactions. *Computers in Human Behavior* 72, 2017 (2017), 432–440. https://doi.org/10.1016/j.chb.2017.02.067
[41] Sergazy Narynov, Zhandos Zhumanov, Aidana Gumar, Mariyam Khassanova, and Matyrkhan Omarov. 2021. Development of Chatbot Psychologist Applying Natural Language Understanding Techniques. In *Twentyfirst International Conference on Control, Automation and Systems (ICCAS 2021)*.
[42] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. https://doi.org/10.1111/0022-4537.00153
[43] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems celebrating interdependence - CHI '94*. ACM Press, New York, New York, USA, 72–78. https://doi.org/10.1145/191666.191703
[44] Implementation Of, A Chatbot In, Online Higher, and Education Settings. 2018. Issues in Information Systems DESIGN AND IMPLEMENTATION OF A CHATBOT IN ONLINE HIGHER Issues in Information Systems. 19, 4 (2018), 44–52.
[45] R. Othman. 2006. The Quality Indicators for an Information Retrieval System: User's Perspective. In *2006 2nd International Conference on Information Communication Technologies*, Vol. 1. 1738–1744. https://doi.org/10.1109/ICTTA.2006.1684648

[46] Nicolas Pfeuffer, Alexander Benlian, Henner Gimpel, and Oliver Hinz. 2019. Anthropomorphic Information Systems. *Business & Information Systems Engineering* 61, 4 (2019), 523–533. https://doi.org/10.1007/s12599-019-00599-y

[47] Aleksandra Przegalinska, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. 2019. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons* 62, 6 (2019), 785–797. https://doi.org/10.1016/j.bushor.2019.08.005

[48] P. Puranam and B.S. Vanneste. 2021. Artificial Intelligence, Trust, and Perceptions of Agency. *Working Paper* (2021). https://doi.org/abstract=3897704

[49] Dragomir Radev, Weiguo Fan, Hong Qi, Harris Wu, and Amardeep Grewal. 2005. Probabilistic question answering on the Web. *Journal of the American Society for Information Science and Technology* 56, 6 (4 2005), 571–583. https://doi.org/10.1002/asi.20146

[50] Bhavika R. Ranoliya, Nidhi Raghuwanshi, and Sanjay Singh. 2017. Chatbot for university related FAQs. *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017* 2017-Janua (2017), 1525–1530. https://doi.org/10.1109/ICACCI.2017.8126057

[51] John T.E. Richardson. 2005. Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education* 30, 4 (2005), 387–415. https://doi.org/10.1080/02602930500099193

[52] Eric Ries. 2011. The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses.

[53] Dmitri Roussinov and José A Robles-Flores. 2007. Applying question answering technology to locating malevolent online content. *Decision Support Systems* 43, 4 (2007), 1404–1418. https://doi.org/10.1016/j.dss.2006.04.006

[54] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-based Adaptive Learning System System for Factual Knowledge. Chi (2019), 1–13. https://doi.org/10.1145/3290605.3300587

[55] Victoria L. Rubin, Yimin Chen, and Lynne Marie Thorimbert. 2010. Artificially intelligent conversational agents in libraries. *Library Hi Tech* 28, 4 (2010), 496–522. https://doi.org/10.1108/07378831011096196

[56] I. Scarpellini and Y. Lim. 2020. Role-Based Design of Conversational Agents: Approach and Tools. In *HCI International 2020 - Late Breaking Posters*.

[57] Anuschka Schmitt, Wambsganss, and Andreas Janson. 2022. Designing for Conversational System Trustworthiness: The Impact of Model Transparency on Trust and Task Performance. In *Thirtieth European Conference on Information Systems*. Timisoara, Romania, 1–18.

[58] Anuschka Schmitt, Thiemo Wambsganss, Matthias Soellner, and Andreas Janson. 2021. Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. In *Forty-Second International Conference on Information Systems*. Austin, Texas, USA, 1–17.

[59] A.-M. Seeger, J. Pfeiffer, and A. Heinzl. 2018. Designing Anthropomorphic Conversational Agents: Development and Empirical Evaluation of a Design Framework. In *Proceedings of the International Conference on Information Systems (ICIS)*. San Francisco, USA.

[60] John V. Seidel. 1998. *Qualitative data analysis*. http://eer.engin.umich.edu/wp-content/uploads/sites/443/2019/08/Seidel-Qualitative-Data-Analysis.pdf

[61] Bayan Abu Shawar and Eric Steven Atwell. 2005. Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics* 10, 4 (2005), 489–516. https://doi.org/10.1075/ijcl.10.4.06sha

[62] Nicole Shechtman and Leonard M Horowitz. 2003. Media inequality in conversation. 5 (2003), 281. https://doi.org/10.1145/642659.642661

[63] R. F. Simmons. 1965. Answering English questions by computer: A survey. *Commun. ACM* 8, 1 (1 1965), 53–70. https://doi.org/10.1145/363707.363732

[64] Ronnie W Smith. [n.d.]. Integration of domain problem solving with natural language dialog : the missing axiom theory. 1 ([n. d.]).

[65] Pavel Smutny and Petra Schreiberova. 2020. Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers and Education* 151, June 2019 (2020), 103862. https://doi.org/10.1016/j.compedu.2020.103862

[66] Eriks Sneiders. 1999. Automated FAQ Answering: Continued Experience with Shallow Language Understanding. *Question Answering Systems. Papers from the 1999 AAAI Fall Symposium* Bubenko 1994 (1999), 97–107. http://www.aaai.org/Papers/Symposia/Fall/1999/FS-99-02/FS99-02-017.pdf

[67] Elliot Soloway, Mark Guzdial, and Kenneth E Hay. 1994. Learner-Centered Design The Challenge For WC1 In The Xst Century. *Interactions* (1994), 36–48.

[68] Yasunobu Sumikawa, Masaaki Fujiyoshi, Hisashi Hatakeyama, and Masahiro Nagai. 2019. Supporting creation of FAQ dataset for e-learning Chatbot. *Smart Innovation, Systems and Technologies* 142, February (2019), 3–13. https://doi.org/10.1007/978-981-13-8311-3{_}1

[69] Ilaria Torre, Jeremy Goslin, and Laurence White. 2020. If your device could smile: People trust happy-sounding artificial agents more. *Computers in Human Behavior* 105, December 2019 (2020), 106215. https://doi.org/10.1016/j.chb.2019.

106215

[70] Gabriele Trovato, Alexander Lopez, Renato Paredes, and Francisco Cuellar. 2017. Security and guidance: Two roles for a humanoid robot in an interaction experiment. *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication* 2017-Janua (2017), 230–235. https://doi.org/10.1109/ROMAN.2017.8172307

[71] Shahper Vodanovich, David Sundaram, and Michael Myers. 2010. Digital natives and ubiquitous information systems. *Information Systems Research* 21, 4 (2010), 711–723. https://doi.org/10.1287/isre.1100.0324

[72] Jan vom Brocke, Alexander Simons, Kai Riemer, Bjoern Niehaves, Ralf Plattfaut, and Anne Cleven. 2015. Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems* 37, 1 (8 2015), 205–224. https://doi.org/10.17705/1cais.03709

[73] Raphael von Wolff, Jonas Nörtemann, Sebastian Hobert, and Matthias Schumann. 2020. Chatbots for the Information Acquisition at Universities – A Student's View on the Application Area. In *International Workshop on Chatbot Research and Design*, Vol. 11970 LNCS. Springer, 231–244. https://doi.org/10.1007/978-3-030-39540-7{\{}{\T1\textbackslash}{_}{\}}16

[74] Thiemo Wambsganß, Leonie Haas, and Matthias Söllner. 2021. Towards the Design of a Student-Centered Question-Answering System in Educational Settings. In *Twenty-Ninth European Conference on Information Systems (ECIS 2021)*. Virtual Conference, 1–8.

[75] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445781

[76] Liz C Wang, Julie Baker, Judy A Wagner, and Kirk Wakefield. 2007. Can a retail Web Site be social? *Journal of Marketing* 71, 3 (2007), 143–157. https://doi.org/10.1509/jmkg.71.3.143

[77] Florian Weber, Thiemo Wambsganss, Dominic Rüttimann, and Matthias Söllner. 2021. Pedagogical Agents for Interactive Learning: A Taxonomy of Conversational Agents in Education. In *Forty-Second International Conference on Information Systems*. Austin, Texas, USA, 1–17.

[78] Tom D Wilson. 1999. Models in information behaviour research. *Journal of documentation* (1999).

[79] R. Winkler and M. Söllner. 2018. Unleashing the Potential of Chatbots in Education : A State-Of-The-Art Analysis . In : Academy of Management. *Meeting, Annual Chicago, A O M* (2018). https://www.alexandria.unisg.ch/254848/1/JML_699.pdf

[80] Rainer Winkler and Matthias Söllner. 2018. Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis. *Academy of Management Proceedings* (2018). https://doi.org/10.5465/ambpp.2018.15903abstract

Received January 2022; revised April 2022; accepted August 2022