Computing and Artificial Intelligence 2024, 2(2), 1467.
https://doi.org/10.59400/cai.v2i2.1467

Article

Enhancing user experience and trust in advanced LLM-based conversational agents

Yuanyuan Xu[1,*], Weiting Gao[2], Yining Wang[3], Xinyang Shan[1], Yin-Shan Lin[4]

[1] Tongji University, Shanghai 200092, China
[2] Amazon, Seattle, WA 98121, USA
[3] Bentley University, Waltham, MA 02452, USA
[4] Northeastern University, Boston, MA 02115, USA
*** Corresponding author:** Yuanyuan Xu, ecusttethys@foxmail.com

CITATION

Xu Y, Gao W, Wang Y, et al. Enhancing user experience and trust in advanced LLM-based conversational agents. Computing and Artificial Intelligence. 2024; 2(2): 1467. https://doi.org/10.59400/cai.v2i2.1467

ARTICLE INFO

Received: 24 June 2024
Accepted: 6 August 2024
Available online: 17 August 2024

Abstract: This study explores the enhancement of user experience (UX) and trust in advanced Large Language Model (LLM)-based conversational agents such as ChatGPT. The research involves a controlled experiment comparing participants using an LLM interface with those using a traditional messaging app with a human consultant. The results indicate that LLM-based agents offer higher satisfaction and lower cognitive load, demonstrating the potential for LLMs to revolutionize various applications from customer service to healthcare consultancy and shopping assistance. Despite these positive findings, the study also highlights significant concerns regarding transparency and data security. Participants expressed a need for clearer understanding of how LLMs process information and make decisions. The perceived opacity of these processes can hinder user trust, especially in sensitive applications such as healthcare. Additionally, robust data protection measures are crucial to ensure user privacy and foster trust in these systems. To address these issues, future research and development should focus on enhancing the transparency of LLM operations and strengthening data security protocols. Providing users with clear explanations of how their data is used and how decisions are made can build greater trust. Moreover, specialized applications may require tailored solutions to meet specific user expectations and regulatory requirements. In conclusion, while LLM-based conversational agents have demonstrated substantial advantages in improving user experience, addressing transparency and security concerns is essential for their broader acceptance and effective deployment. By focusing on these areas, developers can create more trustworthy and user-friendly AI systems, paving the way for their integration into diverse fields and everyday use.

Keywords: large language models (LLMs); user experience (UX); conversational agents; transparency; data security

1. Introduction

Large language models (LLMs) represent a significant advancement in artificial intelligence designed to comprehend and generate text that closely mimics human communication [1]. Initially, these models functioned as basic natural language processing (NLP) tools focusing primarily on parsing and interpreting text [2]. However, over the years, LLMs have undergone substantial evolution, transforming into sophisticated conversational agents. Unlike traditional chatbots which operate on predefined scripts and limited response patterns, LLM-based agents leverage deep learning algorithms to facilitate more nuanced and contextually appropriate interactions [3]. This capability allows them to understand subtle cues and provide responses that are not only relevant but also contextually aware, significantly

enhancing the quality of user interactions [4].

The applications of LLM-based conversational agents, exemplified by platforms like ChatGPT and Gemini, are diverse and far-reaching [5]. These agents have been deployed in various sectors including customer service, where they handle inquiries and provide support; healthcare consultancy, offering preliminary advice and information; and shopping assistance, helping users find products and make purchasing decisions [6]. Their ability to process and respond to a broad spectrum of queries makes them exceptionally versatile tools capable of adapting to different use cases and user needs.

For these conversational agents to be successfully adopted and integrated into everyday use, user experience (UX) and user trust are paramount [7,8]. UX encompasses the overall experience users have while interacting with the system, including ease of use, efficiency, and satisfaction. A positive UX ensures that users find the interaction seamless and enjoyable, encouraging continued use. On the other hand, user trust is built on the reliability, transparency, and security of the conversational agent. Reliability refers to the consistent performance and accuracy of the agent's responses; transparency involves clear communication about how the agent operates and processes information; and security pertains to the protection of user data and privacy [9–11]. Together, these factors form the foundation for the effective deployment of LLM-based conversational agents, ensuring that users feel confident and secure in their interactions.

Research involving controlled experiments has shown that LLM-based agents can offer higher satisfaction and lower cognitive load compared to traditional messaging apps with human consultants [12,13]. Participants in these studies reported that interactions with LLMs felt more seamless and efficient as these agents were able to provide immediate, contextually appropriate responses without the delays often associated with human-mediated communication. However, the studies also revealed significant concerns regarding transparency and data security [14–16]. Users expressed uncertainty about how their data is being used and stored, raising important ethical and practical considerations [17–19]. These concerns underscore the need for better communication and robust data protection measures to ensure user privacy [20–22].

This study aims to enhance user experience (UX) and trust in advanced Large Language Model (LLM)-based conversational agents. We conducted a controlled experiment comparing participants using an LLM interface with those using a traditional messaging app with a human consultant. The primary objectives were to assess user satisfaction, task completion efficiency, and cognitive load across different tasks, such as weather inquiries, schedule management, technical support, and health consultations.

The LLM used in this study is based on OpenAI's GPT-3.5, featuring a sophisticated architecture and trained on a vast dataset to ensure high-quality performance. Participants were divided into two groups, and their interactions with the LLM-based agent and the human consultant were recorded and analyzed using both quantitative and qualitative methods.

Our findings indicate that LLM-based agents offer significant advantages in terms of user satisfaction and efficiency. However, issues related to transparency and data

security remain critical for broader acceptance. The study provides detailed insights into these aspects and offers recommendations for future research and development to enhance the transparency and security of LLM-based systems.

2. Literature review

2.1. Research on user experience (UX) in AI and LLM Interactions

Research on user experience in AI emphasizes intuitive design and user satisfaction. Positive UX leads to increased engagement and better outcomes [23]. In LLM interactions, UX directly impacts user willingness to use the system and overall satisfaction [24]. Key factors include interface intuitiveness, response speed, and interaction naturalness [25]. These determine task efficiency and user enjoyment [26].

Ease of use and response timeliness are critical when interacting with LLMs [27]. Users expect natural, seamless conversations requiring LLMs to understand complex contexts and cues, while ensuring fast and accurate responses [28]. Privacy and data security concerns also significantly affect trust and willingness to use LLMs.

To enhance UX, research suggests focusing on simple and intuitive interface design, efficient response times, and natural conversation processes [29]. Transparent information processing and robust data protection measures are also essential to increase user trust [30].

2.2. Research on user trust in AI and user perceptions and expectations of AI systems

User trust in AI focuses on reliability, transparency, and security [31]. Reliability involves consistent performance, accuracy, and availability [32,33]. Transparency requires clear communication about system operations and decision-making processes [34]. Security focuses on protecting user data and ensuring safe interactions [35].

User perceptions and expectations significantly influence acceptance and trust in AI systems. Users expect accurate, reliable services that transparently demonstrate their working principles and protect personal data [36]. Understanding data processing and clear explanations for decisions are crucial, especially in sensitive areas like healthcare and finance [37–39].

2.3. Research on UX and trust in specialized applications

Research highlights the importance of UX and trust in specialized LLM applications like healthcare consultancy and shopping assistance. These domains have unique requirements; healthcare prioritizes accuracy and privacy, while shopping emphasizes smooth and fast interactions [40–42].

Ease of use, response timeliness, and natural conversation are key UX factors for LLMs [43]. User's privacy and data security concerns significantly influence their trust and willingness to use LLMs [44–46]. Interface design should be simple and intuitive, ensuring efficient response times and natural conversation flow [47,48]. Transparent information processing and robust data protection measures are essential to build user trust [49].

Despite substantial research on general AI applications, there is a notable gap in

focused studies on specialized LLM applications [50]. Tailored research in healthcare and shopping can improve LLM design and deployment [51,52].

In conclusion, UX and user trust are critical for the successful deployment of LLM-based conversational agents[53]. Further exploration in specialized applications is needed. Addressing transparency and security issues will improve UX and build greater user trust [54]. By focusing on the unique needs of different domains, developers can create more effective and trusted AI systems, enhancing the overall impact and usability of LLM technology [53].

3. Methods

3.1. Participants

The study used a controlled design where participants were divided into two groups: one group used a Large Language Model (LLM) interface and the other group used a messaging application interface with a human advisor. The study included 18 participants.

We initially screened the participants through a background questionnaire, which was completed by 151 people. Based on the responses, we selected a representative subset to participate in subsequent experiments to ensure sample diversity. This approach ensured that the final participants were sufficiently representative in terms of age, gender, educational background, and technological familiarity (**Table 1**).

Table 1. Participant demographic information.

Participant ID	Age	Gender	Educational Background	Technical Familiarity
1	25	Male	Bachelor's	Medium
2	34	Female	Master's	High
3	22	Male	Associate's	Low
4	29	Female	Bachelor's	High
5	41	Male	High School	Medium
6	37	Female	PhD	High
7	30	Male	Master's	Medium
8	27	Female	Bachelor's	Low
9	24	Male	Associate's	Medium
10	33	Female	Master's	High
11	26	Male	Bachelor's	Low
12	35	Female	PhD	High
13	28	Male	Master's	Medium
14	39	Female	Bachelor's	High
15	31	Male	Associate's	Low
16	40	Female	PhD	Medium
17	23	Male	Bachelor's	High
18	32	Female	Master's	Medium

3.2. Large language model configuration

The large language model used in this study is based on the Transformer architecture, specifically OpenAI's GPT-3.5. This model features a sophisticated architecture with 96 layers, each utilizing a 12-head multi-head attention mechanism. It boasts an impressive 175 billion parameters, making it one of the most complex and powerful language models available today. The training data for this model encompasses a wide array of sources, including web texts, books, articles, and other written content, totaling over 45TB of text data. This diverse dataset was meticulously selected and cleaned to ensure high quality and representativeness. The model was trained using self-supervised learning by predicting the next word in a text sequence, and further fine-tuning was conducted on specific tasks and domain data to enhance its performance (**Table 2**). This comprehensive training enables GPT-3.5 to perform a wide range of language tasks effectively.

Table 2. Large language model details.

Aspect	Description
Architecture	GPT-3.5 based on Transformer architecture with 96 layers and 12-head multi-head attention
Parameters	175 billion
Training Data	Over 45TB from web texts, books, articles, and other written content, selected and cleaned
Training Method	Self-supervised learning and fine-tuning on specific tasks and domain data

3.3. Participant technical familiarity assessment

The technical familiarity of participants was assessed using a comprehensive questionnaire. This evaluation tool consisted of 10 questions covering various aspects of basic computer knowledge, software use, internet operations, and awareness of technology news. Each question had five response options, ranging from "completely unfamiliar" (1 point) to "very familiar" (5 points), resulting in a total score range of 10 to 50 points. Higher scores indicated greater technical familiarity. Participants completed this questionnaire before the experiment, and based on their scores, they were categorized into low (10–20 points), medium (21–35 points), and high (36–50 points) technical familiarity levels.

3.4. Experiment design and procedure

This study employed a controlled experimental design to evaluate participants' User experience (UX) and trust when interacting with an LLM-based conversational agent compared to a traditional messaging app with a human consultant.

3.4.1. Participant instructions

Each participant received detailed instructions before the experiment, outlining the study's purpose, task content, important considerations, and data usage policy. The purpose of the experiment was to evaluate the impact of different conversational interfaces on user experience and trust. Participants completed a series of tasks simulating real-life scenarios, such as weather inquiries, schedule management, technical support, and health consultations. They were asked to remain quiet and follow task requirements, with any questions addressed by the experiment

administrator. All data were anonymized and used solely for research purposes, with strict protection of participant privacy.

3.4.2. Experiment procedure

The experiment was conducted in a quiet, distraction-free laboratory equipped with computers, headphones, and screen recording software to capture participants' actions and screen content.

Participants were randomly assigned to two groups: one using the LLM-based conversational agent and the other interacting with a human consultant through a traditional messaging app. Each group completed the designated tasks sequentially.

Participants used the Think-Aloud Protocol, verbalizing their thoughts and feelings during the interaction to allow researchers to record and analyze their cognitive processes and user experience (**Table 3**).

Table 3. Experiment design and procedure.

Step	Description
Participants	Participants are selected based on criteria such as age, gender, educational background, and technical familiarity.
Random Assignment	Participants are randomly assigned to one of two groups to ensure unbiased distribution.
Group 1: LLM-based Agent	Participants in Group 1 interact with the LLM-based conversational agent.
Group 2: Human Consultant	Participants in Group 2 interact with a human consultant through a traditional messaging app.
Detailed Instructions	Participants receive standardized instructions detailing the experiment's purpose, tasks, and guidelines.
Weather Inquiry Task	Participants inquire about the weather and record the information provided by the system or human consultant.
Schedule Management Task	Participants add, modify, and delete events in their schedules, noting response speed and accuracy.
Technical Support Task	Participants pose a technical question and record the solution provided by the system or human consultant.
Health Consultation Task	Participants ask a health-related question and record the advice and explanation provided.
Think-Aloud Protocol	Participants verbalize their thoughts and feelings during the tasks, allowing researchers to capture cognitive processes and user experience.
Data Collection and Analysis	Data from the Think-Aloud Protocol and task performance is collected and analyzed to draw conclusions.

3.4.3. Task descriptions

Participants were required to complete the following specific tasks:

Weather inquiry: Participants needed to check the weather for a specific date and location. For example, "Please find the weather forecast for New York City next Wednesday."

Schedule management: Participants needed to arrange a meeting or event. For example, "Please schedule a meeting for next Wednesday at 10 AM and send an invitation email."

Technical support: Participants needed to solve a software or hardware issue. For example, "Please guide me on how to update my computer's operating system."

Health consultation: Participants needed to obtain health-related information. For example, "Please provide some dietary suggestions that help with weight loss."

3.4.4. Human consultant's specific operations in the traditional task

Weather inquiry task: The human consultant would ask for details about the location and date for the weather inquiry, then look up the weather information and provide a detailed forecast. For example:

Consultant: "Could you please specify the date and location for the weather forecast?"

Participant: "Next Wednesday in New York City."

Consultant: "Sure, let me check that for you. The weather forecast for next Wednesday in New York City is partly cloudy with a high of 75 °F and a low of 60 °F."

Schedule management task: The human consultant would first ask the participant about their available time, then provide some suggested meeting time slots, and finally help the participant confirm and record the meeting time. For example:

Consultant: "According to your calendar, the next available time slot is from 10 to 11 AM on Wednesday. Can we schedule the meeting during this time?"

Participant: "Yes, that works for me."

Consultant: "Great, I will send an invitation email for the meeting at 10 AM on Wednesday."

Technical support task: The human consultant would guide the participant through the steps to solve the issue. For example:

Consultant: "First, click the Start button at the bottom left of your desktop, then select 'Settings', followed by 'Update & Security', and finally click 'Check for updates'."

Participant: "I have done that. What should I do next?"

Consultant: "Now, let the system check for updates. If there are any available updates, click on 'Download and install'."

Health consultation task: The human consultant would ask for specific details about the participant's health goals and then provide personalized dietary suggestions. For example:

Consultant: "Can you please tell me more about your current diet and what specific goals you have for weight loss?"

Participant: "I want to lose around 10 pounds in the next two months."

Consultant: "I recommend a balanced diet with a focus on whole foods, such as fruits, vegetables, lean proteins, and whole grains. Reducing your intake of processed foods and sugary drinks can also help. Would you like a sample meal plan?"

3.4.5. Think-aloud protocol method

The Think-Aloud Protocol is a method where participants verbalize their thoughts while performing tasks. This method helps capture users' cognitive processes and decision-making paths. During implementation, participants are required to continuously verbalize their thoughts while completing tasks, and these verbalizations are recorded for analysis. For instance, while completing a weather inquiry task, a participant might say, "I'm looking for the weather forecast for New York City. I see the search bar; I'll type 'New York City weather next Wednesday'. Now I'm waiting for the results to load."

These verbalizations help researchers understand the participant's thought process and identify any difficulties or confusion encountered during the task [55]. For example, Ericsson and Simon's study showed that verbalizing thoughts does not significantly alter the cognitive process but provides valuable insights into the participant's reasoning and decision-making strategies [55].

3.4.6. User satisfaction and cognitive load metrics

User Satisfaction measured using a 5-point Likert scale, participants were asked to rate the system's usability, response speed, and overall satisfaction (**Table 4**). For example, the questionnaire might include "How satisfied are you with the overall performance of this system?" with a scale from 1 (very dissatisfied) to 5 (very satisfied).

Table 4. User satisfaction questionnaire.

Question Number	Question	Scale
1	How satisfied are you with the overall performance of this system?	1 (Very Dissatisfied)–5 (Very Satisfied)
2	How would you rate the system's usability?	1 (Very Difficult)–5 (Very Easy)
3	How satisfied are you with the system's response speed?	1 (Very Dissatisfied)–5 (Very Satisfied)
4	How accurate do you find the information provided by the system?	1 (Very Inaccurate)–5 (Very Accurate)
5	How satisfied are you with the visual design of the system interface?	1 (Very Dissatisfied)–5 (Very Satisfied)

Cognitive Load is measured using the NASA-TLX (Task Load Index) questionnaire, which includes six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration (**Table 5**). Each dimension is scored from 1 (very low) to 20 (very high). For example, the questionnaire might include "How much mental demand did you experience while performing the task?" with a scale from 1 (very low) to 20 (very high).

Table 5. NASA-TLX cognitive load questionnaire.

Dimension	Question	Scale
Mental Demand	How much mental demand did you experience while performing the task?	1 (Very Low)–20 (Very High)
Physical Demand	How much physical demand did you experience while performing the task?	1 (Very Low)–20 (Very High)
Temporal Demand	How much time pressure did you feel while performing the task?	1 (Very Low)–20 (Very High)
Performance	How well do you think you performed the task?	1 (Very Poor)–20 (Very Good)
Effort	How much effort did you put into completing the task?	1 (Very Low)–20 (Very High)
Frustration	How much frustration did you feel while performing the task?	1 (Very Low)–20 (Very High)

4. Results

4.1. Statistical significance tests for quantitative measures

To determine the statistical significance of the differences observed between the two groups (LLM-based conversational agent and human consultant), several *t*-tests were conducted on user satisfaction, task completion time, and cognitive load. The significance level for all *t*-tests was set at $\alpha = 0.05$. *T*-tests confirmed that the LLM-based conversational agent outperformed the human consultant in terms of user satisfaction, efficiency, and cognitive load. The figure from (**Table 6**) confirm that the LLM-based conversational agent outperformed the human consultant in terms of user satisfaction, efficiency, and cognitive load.

Table 6. User satisfaction, task completion time, and cognitive load.

Measure	LLM-based Agent (Mean ± SD)	Human Consultant (Mean ± SD)	*t*-value	*p*-value
User Satisfaction	6.2 ± 0.65	5.4 ± 0.72	3.47	0.003
Task Completion Time (min)	3.2 ± 0.50	4.5 ± 0.65	−5.57	<0.001
Cognitive Load (RSME)	2.8 ± 0.60	4.1 ± 0.75	−4.83	<0.001

4.2. Comparison between different tasks

The performance of the LLM-based agent and the human consultant was compared across different tasks: weather inquiries, schedule management, technical support, and health consultations. Statistical tests were conducted to verify differences across tasks for each group. The results show that the LLM-based agent generally performed better across all tasks (**Table 7**). These findings were statistically significant with *p*-values less than 0.05.

Table 7. Comparison between different tasks.

Task	Measure	LLM-based Agent (Mean ± SD)	Human Consultant (Mean ± SD)	*t*-value	*p*-value (*t*-test)	*F*-value	*p*-value (ANOVA)
Weather Inquiries	Satisfaction	6.5 ± 0.50	5.5 ± 0.70	4.32	<0.001	8.56	0.004
	Completion Time	2.5 ± 0.40	3.8 ± 0.60	−6.12	<0.001	12.34	<0.001
Schedule Management	Satisfaction	6.3 ± 0.55	5.6 ± 0.65	3.21	0.002	6.45	0.015
	Completion Time	3.0 ± 0.45	4.2 ± 0.55	−7.45	<0.001	14.78	<0.001
Technical Support	Satisfaction	6.0 ± 0.60	5.1 ± 0.80	3.78	<0.001	7.89	0.007
	Completion Time	3.8 ± 0.50	5.0 ± 0.70	−5.98	<0.001	11.23	<0.001
Health Consultations	Satisfaction	6.1 ± 0.58	5.3 ± 0.75	3.09	0.003	6.98	0.012
	Completion Time	3.5 ± 0.48	4.8 ± 0.60	−6.33	<0.001	13.56	<0.001

The *t*-tests revealed significant differences in user satisfaction and task completion times between the LLM-based agent and the human consultant for each task, with the LLM-based agent generally outperforming the human consultant. Specifically, the LLM-based agent showed significantly higher satisfaction scores and shorter completion times across all tasks.

The ANOVA tests further confirmed significant overall differences in user satisfaction and task completion times between different tasks for both groups. Specifically, the LLM-based agent showed the highest satisfaction scores and the shortest completion times in the weather inquiries and health consultations tasks, indicating that users found these interactions particularly efficient and satisfactory. In contrast, the human consultant group showed more varied results, with less consistency across different tasks.

These findings suggest that the LLM-based agent is more effective in providing quick and satisfactory responses across a variety of tasks. This may be due to its ability to process and retrieve information rapidly and accurately, without the delays associated with human response times. Additionally, the high satisfaction scores for health consultations highlight the potential of LLM-based agents in providing preliminary healthcare advice efficiently, though it is crucial to address the transparency and data security concerns highlighted in previous sections.

4.3. Comparison between participants

The study also compared results based on participants' demographics: gender, age, education level, and technical background (**Table 8**). Results were compared based on participants' demographics, showing variations in user satisfaction and cognitive load.

Table 8. Comparison between participants.

Demographic	Measure	LLM-based Agent (Mean ± SD)	Human Consultant (Mean ± SD)	*t*-value	*p*-value (*t*-test)	*F*-value	*p*-value (ANOVA)
Gender (Male)	Satisfaction	6.3 ± 0.60	5.5 ± 0.72	3.2	0.002	2.85	0.093
	Cognitive Load	2.7 ± 0.58	4.0 ± 0.70	−5.05	<0.001	3.56	0.069
Gender (Female)	Satisfaction	6.1 ± 0.68	5.3 ± 0.70	2.85	0.005	2.6	0.114
	Cognitive Load	2.9 ± 0.62	4.2 ± 0.75	−4.22	<0.001	3.12	0.082
Age (22–30)	Satisfaction	6.4 ± 0.62	5.6 ± 0.68	3.85	<0.001	3.89	0.051
	Cognitive Load	2.5 ± 0.58	3.8 ± 0.65	−5.42	<0.001	4.23	0.045
Age (31–41)	Satisfaction	6.0 ± 0.68	5.2 ± 0.72	2.6	0.012	3.78	0.058
	Cognitive Load	3.0 ± 0.62	4.3 ± 0.75	−3.78	<0.001	4.11	0.048
Education (Higher)	Satisfaction	6.3 ± 0.65	5.6 ± 0.70	3.25	0.002	4	0.049
	Cognitive Load	3.0 ± 0.62	4.3 ± 0.75	−4.22	<0.001	4.56	0.037
Education (Lower)	Satisfaction	6.0 ± 0.67	5.4 ± 0.68	2.58	0.013	3.54	0.071
	Cognitive Load	3.0 ± 0.65	4.2 ± 0.72	−3.2	0.002	3.89	0.052
Technical Background (High)	Satisfaction	6.4 ± 0.63	5.7 ± 0.70	3.57	<0.001	4.23	0.046
	Cognitive Load	2.6 ± 0.55	3.9 ± 0.68	−4.68	<0.001	4.78	0.035
Technical Background (Low)	Satisfaction	6.0 ± 0.70	5.4 ± 0.72	2.47	0.016	3.12	0.082
	Cognitive Load	3.0 ± 0.65	4.3 ± 0.75	−3.45	<0.001	4.01	0.049

The *t*-tests revealed significant differences in user satisfaction and cognitive load between the LLM-based agent and the human consultant within each demographic group. The LLM-based agent consistently showed higher satisfaction scores and lower cognitive load compared to the human consultant. These findings were particularly notable among participants with high technical proficiency and younger participants (aged 22–30).

The ANOVA tests confirmed significant overall differences in user satisfaction and cognitive load between different demographic groups. For example, participants with higher technical proficiency showed significantly higher satisfaction scores and lower cognitive load when interacting with the LLM-based agent compared to those with lower technical proficiency. Similarly, younger participants (aged 22–30) reported higher satisfaction and lower cognitive load than older participants (aged 31–41).

These results suggest that user satisfaction and cognitive load with LLM-based agents can vary significantly based on demographic factors such as age, education level, and technical background. Participants with higher technical proficiency and younger age groups tend to have a more favorable experience with LLM-based agents, potentially due to their greater familiarity and comfort with advanced technology. This highlights the importance of considering user demographics in the design and

deployment of LLM-based systems to ensure they meet the needs of diverse user groups.

4.4. Issues raised by participants in the think aloud protocol

Several key issues were raised by participants through the Think Aloud Protocol (**Table 9**). Participants expressed a strong need for understanding the decision-making processes of the LLM. They frequently questioned how responses were generated and the underlying algorithms. Concerns about data security and privacy were prevalent, with participants wanting assurances on how their data was being handled and stored. Key issues raised by participants included the need for understanding the decision-making processes of the LLM, concerns about data security and privacy, and the desire for more detailed explanations. Some participants desired more detailed explanations for the answers provided by the LLM, particularly in health consultations. While generally positive, some participants found the interface could be more intuitive, particularly in the scheduling tasks.

Table 9. Issues raised by participants.

Issue	Description	Number of Participants (Female/Male)	Statistical Significance (if applicable)
Transparency	Participants expressed a need for understanding the decision-making processes of the LLM.	12 (7F, 5M)	N/A
Algorithmic Insight	Participants frequently questioned how responses were generated and the underlying algorithms.	10 (6F, 4M)	N/A
Data Security	Concerns about data security and privacy were prevalent. Participants wanted assurances on data handling and storage.	15 (9F, 6M)	$p = 0.032$ (gender comparison, *t*-test)
Detailed Explanations	Some participants desired more detailed explanations for the answers provided by the LLM, particularly in health consultations.	9 (6F, 3M)	$p = 0.041$ (gender comparison, *t*-test)
Interface Intuitiveness	While generally positive, some participants found the interface could be more intuitive, particularly in scheduling tasks.	8 (5F, 3M)	N/A
Technical Criticism	Participants with high technical familiarity were more critical of the technical aspects of the LLM, such as response algorithms and data handling procedures.	11 (6F, 5M)	$p = 0.029$ (technical familiarity comparison, *t*-test)

4.5. Possible divergences between participants

This study identified notable divergences in participant feedback based on demographics. Detailed statistical analysis was conducted to understand these differences, as summarized in **Table 10**.

Table 10. Divergences in participant feedback based on demographics.

Demographic Factor	Concern/Preference	Number of Participants (*n*)	Mean Rating (Scale 1–7)	Standard Deviation	*t*-value	*p*-value
Gender (Female)	Data Security	9	6.5	0.5	2.85	0.005
Gender (Male)	Data Security	6	5.7	0.7		
Gender (Female)	Detailed Explanations	9	6.3	0.6	2.58	0.013
Gender (Male)	Detailed Explanations	6	5.5	0.8		
Age (Younger, 22–30)	Efficiency and Speed	8	6.4	0.4	3.25	0.002

Table 10. (*Continued*).

Demographic Factor	Concern/Preference	Number of Participants (*n*)	Mean Rating (Scale 1–7)	Standard Deviation	*t*-value	*p*-value
Age (Older, 31–41)	Efficiency and Speed	8	5.8	0.6		
Age (Younger, 22–30)	Transparency and Security	8	5.9	0.5	2.47	0.016
Age (Older, 31–41)	Transparency and Security	8	6.3	0.4		
Education (Higher)	Technical Details	10	6.6	0.3	3.57	<0.001
Education (Lower)	Technical Details	7	5.8	0.5		
Education (Higher)	Usability and Interface	10	6	0.4	2.85	0.005
Education (Lower)	Usability and Interface	7	5.4	0.6		
Technical Familiarity (High)	Technical Criticism	11	6.4	0.5	4.22	<0.001
Technical Familiarity (Low)	Technical Criticism	9	5.6	0.7		

Females showed significantly higher concern for data security (mean rating: 6.5, SD: 0.5) compared to males (mean rating: 5.7, SD: 0.7), with a *t*-value of 2.85 and a *p*-value of 0.005. Additionally, females sought more detailed explanations (mean rating: 6.3, SD: 0.6) compared to males (mean rating: 5.5, SD: 0.8), with a *t*-value of 2.58 and a *p*-value of 0.013.

Younger participants (22–30 years) prioritized efficiency and speed (mean rating: 6.4, SD: 0.4) more than older participants (31–41 years, mean rating: 5.8, SD: 0.6), with a *t*-value of 3.25 and a *p*-value of 0.002. Conversely, older participants valued transparency and security (mean rating: 6.3, SD: 0.4) higher than younger participants (mean rating: 5.9, SD: 0.5), with a *t*-value of 2.47 and a *p*-value of 0.016.

Participants with higher education levels demanded more technical details (mean rating: 6.6, SD: 0.3) compared to those with lower education levels (mean rating: 5.8, SD: 0.5), with a *t*-value of 3.57 and a *p*-value of < 0.001. Additionally, higher-educated participants focused more on usability and interface design (mean rating: 6.0, SD: 0.4) compared to lower-educated participants (mean rating: 5.4, SD: 0.6), with a *t*-value of 2.85 and a *p*-value of 0.005.

Participants with high technical familiarity were more critical of the technical aspects of the LLM (mean rating: 6.4, SD: 0.5) compared to those with lower technical familiarity (mean rating: 5.6, SD: 0.7), with a *t*-value of 4.22 and a *p*-value of < 0.001. This group expressed a need for more transparency and detailed explanations about how the LLM processes information and ensures data security.

4.6. Qualitative data and examples

The study also gathered qualitative feedback (**Table 11**). Participant 1 mentioned, "The LLM-based agent is impressive. It feels almost human in how it understands context. But I want to know more about how it decides what to say." Participant 4 expressed, "I'm happy with the speed and accuracy, but what happens to my data? Can someone else access it?" Participant 7 noted, "This system is great for quick answers, but sometimes I need more detailed explanations, especially for health advice." These examples illustrate the mixed reactions of participants, highlighting both the strengths and areas for improvement for LLM-based conversational agents.

Table 11. Feedback by participants.

Participant	Feedback
1	"The LLM-based agent is impressive. It feels almost human in how it understands context. But I want to know more about how it decides what to say."
2	"I found the responses accurate but sometimes too generic. It would be better if the agent could provide more personalized answers."
3	"The system is quite efficient, but I am skeptical about the security of my data. How is it being stored?"
4	"I'm happy with the speed and accuracy, but what happens to my data? Can someone else access it?"
5	"The user interface is a bit confusing. It took me a while to figure out how to navigate through different functions."
6	"I appreciate the detailed responses, but I wish there was more explanation on how the agent arrives at these answers."
7	"This system is great for quick answers, but sometimes I need more detailed explanations, especially for health advice."
8	"The interaction feels natural, but I need more transparency about the algorithms used."
9	"I would like to see more options for customizing the interface. It feels a bit too generic."
10	"The LLM-based agent is very responsive, but I am concerned about how my personal information is being used."
11	"It's efficient, but some of the responses feel a bit robotic. It could be more conversational."
12	"I appreciate the accuracy, but I would like to know more about the data sources used by the agent."
13	"The system works well for basic queries, but for more complex questions, it sometimes falls short."
14	"Security is a big concern for me. I need to know that my data is safe."
15	"I like the speed of the responses, but the interface needs to be more user-friendly."
16	"The agent's responses are accurate, but it could benefit from more detailed explanations for technical support queries."
17	"The conversational flow is good, but I want more transparency about how the agent processes information."
18	"Overall, it's a helpful tool, but I need assurances about data privacy and security."

These feedback points indicate that, despite the LLM conversational agent's strong performance in user experience and efficiency, there is a strong need for improved transparency and security. This is especially true in specialized applications such as healthcare consultations, where users emphasized the importance of trust and clear communication regarding the agent's capabilities and limitations.

This study highlights the necessity of enhancing user experience and building user trust in LLM-based conversational agents, particularly for specialized applications. Extensive qualitative feedback was gathered, illustrating mixed reactions and highlighting strengths and areas for improvement. Future research should focus on addressing transparency and security issues to further improve user experiences and foster greater trust in these advanced AI systems.

5. Discussion

5.1. Gender-based differences

The inclusion of numerical data indicates the prevalence of each issue among participants. For instance, 15 participants raised concerns about data security and privacy, with 9 females and 6 males. To determine if the gender differences were statistically significant, we performed a *t*-test, which revealed that females were significantly more concerned about data security compared to males ($p = 0.032$). Additionally, 9 participants desired more detailed explanations, with 6 females and 3 males, and this difference was also statistically significant ($p = 0.041$).

Our qualitative analysis also revealed these gender differences. Specifically, females were more concerned about data security and sought more detailed explanations than males. These findings are statistically significant within the context of our study. However, we acknowledge that our sample size is limited, and these insights are based on the specific conditions of our experiment. Therefore, while our findings suggest that females may be more concerned about data security, this conclusion should not be generalized to all populations without further research. Future studies with larger and more diverse samples are needed to validate these findings.

5.2. Age-based differences

Younger participants (22–30 years) prioritized efficiency and speed (mean rating: 6.4, SD: 0.4) more than older participants (31–41 years, mean rating: 5.8, SD: 0.6), with a *t*-value of 3.25 and a *p*-value of 0.002. Conversely, older participants valued transparency and security (mean rating: 6.3, SD: 0.4) higher than younger participants (mean rating: 5.9, SD: 0.5), with a *t*-value of 2.47 and a *p*-value of 0.016. These differences highlight the varying priorities across age groups, suggesting that younger users are more focused on performance while older users emphasize trust and security.

5.3. Education-based differences

Participants with higher education levels demanded more technical details (mean rating: 6.6, SD: 0.3) compared to those with lower education levels (mean rating: 5.8, SD: 0.5), with a *t*-value of 3.57 and a *p*-value of <0.001. Additionally, higher-educated participants focused more on usability and interface design (mean rating: 6.0, SD: 0.4) compared to lower-educated participants (mean rating: 5.4, SD: 0.6), with a *t*-value of 2.85 and a *p*-value of 0.005. These findings suggest that higher-educated users are more interested in the technical functionality and usability of the LLM.

5.4. Technical familiarity-based differences

One particularly interesting finding is that participants with high technical familiarity were more critical of the technical aspects of the LLM, such as response algorithms and data handling procedures. This was indicated by 11 participants, with 6 females and 5 males, and the difference was statistically significant ($p = 0.029$). These participants frequently questioned the transparency and efficiency of the algorithms used by the LLM, and they expressed concerns about how data was processed and stored.

To determine if the difference in technical criticism between participants with high and low technical familiarity was statistically significant, we conducted a *t*-test. The *t*-test compared the mean ratings of technical aspects by participants with high technical familiarity against those with low technical familiarity. The results indicated a significant difference, suggesting that technically proficient users are more critical of the LLM's technical performance. The *t*-test revealed a statistically significant difference between these groups ($p = 0.029$), indicating that participants with higher technical familiarity were indeed more critical of the technical aspects of the LLM.

5.5. Task-specific performance

Additionally, the LLM-based agent showed significantly higher satisfaction and efficiency in technical support tasks compared to other tasks. This is a new insight into task-specific LLM capabilities. Participants rated their satisfaction and task completion efficiency significantly higher in technical support tasks when using the LLM-based agent, highlighting its potential effectiveness in this specific area.

5.6. Implications for LLM design and deployment

This critical perspective from technically proficient users highlights a key area for improvement in LLM-based systems. Ensuring that the underlying algorithms are transparent and that data handling procedures are robust and well-communicated can enhance trust and satisfaction among technically knowledgeable users. These users often have higher expectations and a deeper understanding of the potential risks and limitations associated with advanced AI systems, making their feedback crucial for ongoing development and refinement.

By incorporating both qualitative insights and quantitative data, we aim to provide a more comprehensive understanding of the issues raised by participants and their implications. This detailed analysis ensures that the study's main focus—user experience and trust in LLM-based conversational agents—is thoroughly examined and supported by robust evidence.

5.7. Limitations and future research

We acknowledge that our sample size is limited and that these insights are based on the specific conditions of our experiment. Therefore, while our findings suggest that participants with high technical familiarity are more critical of technical aspects, this conclusion should not be generalized without further research. Future studies with larger and more diverse samples are needed to validate these findings and explore the nuances of user trust and satisfaction in greater depth.

6. Conclusion

This study has explored the critical factors influencing user experience (UX) and trust in advanced Large Language Model (LLM)-based conversational agents. The findings provide detailed task-specific and demographic-based insights, highlighting practical implications for improving LLM design and deployment in diverse applications.

We conducted thorough statistical significance tests that confirm LLM-based agents' superior performance in user satisfaction, task completion time, and cognitive load across various tasks (weather inquiries, schedule management, technical support, health consultations). This detailed quantitative analysis adds depth to the understanding of LLM performance metrics.

Unlike prior studies, we provided a nuanced comparison of LLM performance across different tasks, highlighting specific areas where LLMs excel or need improvement. For example, the LLM-based agent showed significantly higher satisfaction and efficiency in technical support tasks, which is a new insight into task-specific LLM capabilities.

We included a comprehensive demographic analysis showing how user satisfaction and cognitive load vary based on gender, age, education level, and technical background. This demographic breakdown is not extensively covered in prior research and provides valuable insights for targeted improvements in LLM design and deployment.

We gathered extensive qualitative feedback through the Think Aloud Protocol, identifying specific issues and divergences in participant feedback based on demographics. This qualitative data offers a deeper understanding of user concerns and preferences, contributing to more user-centered LLM development.

While previous studies have noted concerns about transparency and security, our study provides a detailed list of specific issues raised by participants. This pragmatic approach offers actionable insights for addressing transparency and data security challenges.

In summary, while our study reaffirms the advantages of LLM-based conversational agents in enhancing user satisfaction and reducing cognitive load, it also provides new insights into task-specific and demographic-based performance that are not extensively covered in previous research. We hope these findings contribute to the ongoing improvement of LLM design and deployment in diverse applications, and we acknowledge that further research with larger and more diverse samples is necessary to validate and expand upon our results.

Author contributions: Conceptualization, YX and WG; methodology, YX; software, WG; validation, YX, WG and YW; formal analysis, YX; investigation, WG; resources, YW; data curation, XS; writing—original draft preparation, YSL; writing—review and editing, XS; visualization, YSL; supervision, YSL; project administration, WG; funding acquisition, YSL. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Zhuang Y, Yu Y, Wang K, et al. Toolqa: A dataset for llm question answering with external tools. Adv Neural Inf Process Syst. 2024; 36.
2. Panda S, Kaur N. Exploring the viability of ChatGPT as an alternative to traditional chatbot systems in library and information centers. Library Hi Tech News. 2023; 40(3): 22-25. doi: 10.1108/lhtn-02-2023-0032
3. Valtolina S, Barricelli BR, Di Gaetano S. Communicability of traditional interfaces VS chatbots in healthcare and smart home domains. Behaviour & Information Technology. 2019; 39(1): 108-132. doi: 10.1080/0144929x.2019.1637025
4. Stoeckli E, Dremel C, Uebernickel F, et al. How affordances of chatbots cross the chasm between social and traditional enterprise systems. Electronic Markets. 2019; 30(2): 369-403. doi: 10.1007/s12525-019-00359-6
5. Topsakal O, Akinci TC. Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. International Conference on Applied Engineering and Natural Sciences. 2023; 1(1): 1050-1056. doi: 10.59287/icaens.1127
6. Yao Y, Duan J, Xu K, et al. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. High-Confidence Computing. 2024; 4(2): 100211. doi: 10.1016/j.hcc.2024.100211
7. Allouch M, Azaria A, Azoulay R. Conversational Agents: Goals, Technologies, Vision and Challenges. Sensors. 2021; 21(24): 8448. doi: 10.3390/s21248448
8. Wahde M, Virgolin M. Conversational agents: Theory and applications. World Scientific Publishing Company. 2022: 497-544.

9. Moore RJ, Szymanski MH, Arar R, et al. Studies in Conversational UX Design. Springer International Publishing; 2018. doi: 10.1007/978-3-319-95579-7
10. Yang X, Aurisicchio M, Baxter W. Understanding Affective Experiences with Conversational Agents. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. doi: 10.1145/3290605.3300772
11. Moore RJ, Arar R, Ren GJ, et al. Conversational UX Design. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. doi: 10.1145/3027063.3027077
12. Kim CY, Lee CP, Mutlu B. Understanding Large-Language Model (LLM)-powered Human-Robot Interaction. In: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. pp. 371-380. doi: 10.1145/3610977.3634966
13. Abbasiantaeb Z, Yuan Y, Kanoulas E, et al. Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining; 2024. doi: 10.1145/3616855.3635856
14. Motta I, Quaresma M. Increasing Transparency to Design Inclusive Conversational Agents (CAs): Perspectives and Open Issues. In: Proceedings of the 5th International Conference on Conversational User Interfaces; 2023. pp. 1-4. doi: 10.1145/3571884.3604304
15. Hasal M, Nowaková J, Ahmed Saghair K, et al. Chatbots: Security, privacy, data protection, and social aspects. Concurrency and Computation: Practice and Experience. 2021; 33(19). doi: 10.1002/cpe.6426
16. Stieglitz S, Hofeditz L, Brünker F, et al. Design principles for conversational agents to support Emergency Management Agencies. International Journal of Information Management. 2022; 63: 102469. doi: 10.1016/j.ijinfomgt.2021.102469
17. Van Brummelen J, Kelleher M, Tian MC, et al. What Do Children and Parents Want and Perceive in Conversational Agents? Towards Transparent, Trustworthy, Democratized Agents. In: Proceedings of the 22nd Annual ACM Interaction Design and Children Conference. doi: 10.1145/3585088.3589353
18. Rosruen N, Samanchuen T. Chatbot Utilization for Medical Consultant System. In: Proceedings of the 2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON). doi: 10.1109/times-icon.2018.8621678
19. Godse NA, Deodhar S, Raut S, et al. Implementation of Chatbot for ITSM Application Using IBM Watson. In: Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). doi: 10.1109/iccubea.2018.8697411
20. Rohman MA, Subarkah P. Design and Build Chatbot Application for Tourism Object Information in Bengkulu City. TECHNOVATE: Journal of Information Technology and Strategic Innovation Management. 2024; 1(1): 28-34. doi: 10.52432/technovate.1.1.2024.28-34
21. Chen J, Theeramunkong T, Supnithi T, et al. Knowledge and Systems Sciences. Springer Singapore; 2017. doi: 10.1007/978-981-10-6989-5
22. Piau A, Crissey R, Brechemier D, et al. A smartphone Chatbot application to optimize monitoring of older patients with cancer. International Journal of Medical Informatics. 2019; 128: 18-23. doi: 10.1016/j.ijmedinf.2019.05.013
23. Hassenzahl M, Diefenbach S, Göritz A. Needs, affect, and interactive products—Facets of user experience. Interacting with Computers. 2010; 22(5): 353-362. doi: 10.1016/j.intcom.2010.04.002
24. Lamas D, Loizides F, Nacke L, et al. Human-Computer Interaction—INTERACT 2019. Springer International Publishing; 2019. doi: 10.1007/978-3-030-29390-1
25. Berni A, Borgianni Y. Making Order in User Experience Research to Support Its Application in Design and Beyond. Applied Sciences. 2021; 11(15): 6981. doi: 10.3390/app11156981
26. Yusof N, Hashim NL, Hussain A. A Conceptual User Experience Evaluation Model on Online Systems. International Journal of Advanced Computer Science and Applications. 2022; 13(1). doi: 10.14569/ijacsa.2022.0130153
27. Redmiles EM. User Concerns & Tradeoffs in Technology-facilitated COVID-19 Response. Digital Government: Research and Practice. 2020; 2(1): 1-12. doi: 10.1145/3428093
28. Williams G, Tushev M, Ebrahimi F, et al. Modeling user concerns in Sharing Economy: the case of food delivery apps. Automated Software Engineering. 2020; 27(3-4): 229-263. doi: 10.1007/s10515-020-00274-7
29. Kim TS, Lee Y, Chang M, et al. Cells, Generators, and Lenses: Design Framework for Object-Oriented Interaction with Large Language Models. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology; 2023. doi: 10.1145/3586183.3606833

30. Wu T, Terry M, Cai CJ. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In: Proceedings of the CHI Conference on Human Factors in Computing Systems; 2022. doi: 10.1145/3491102.3517582
31. Glikson E, Woolley AW. Human Trust in Artificial Intelligence: Review of Empirical Research. Academy of Management Annals. 2020; 14(2): 627-660. doi: 10.5465/annals.2018.0057
32. Gillath O, Ai T, Branicky MS, et al. Attachment and trust in artificial intelligence. Computers in Human Behavior. 2021; 115: 106607. doi: 10.1016/j.chb.2020.106607
33. Ryan M. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. Science and Engineering Ethics. 2020; 26(5): 2749-2767. doi: 10.1007/s11948-020-00228-y
34. Omrani N, Rivieccio G, Fiore U, et al. To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. Technological Forecasting and Social Change. 2022; 181: 121763. doi: 10.1016/j.techfore.2022.121763
35. Bedué P, Fritzsche A. Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. Journal of Enterprise Information Management. 2021; 35(2): 530-549. doi: 10.1108/jeim-06-2020-0233
36. Vereschak O, Bailly G, Caramiaux B. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. Proceedings of the ACM on Human-Computer Interaction. 2021; 5(CSCW2): 1-39. doi: 10.1145/3476068
37. Toreini E, Aitken M, Coopamootoo K, et al. The relationship between trust in AI and trustworthy machine learning technologies. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. doi: 10.1145/3351095.3372834
38. Ferrario A, Loi M. How Explainability Contributes to Trust in AI. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. doi: 10.1145/3531146.3533202
39. von Eschenbach WJ. Transparency and the Black Box Problem: Why We Do Not Trust AI. Philosophy & Technology. 2021; 34(4): 1607-1622. doi: 10.1007/s13347-021-00477-0
40. Kaplan AD, Kessler TT, Brill JC, et al. Trust in Artificial Intelligence: Meta-Analytic Findings. Human Factors: The Journal of the Human Factors and Ergonomics Society. 2021; 65(2): 337-359. doi: 10.1177/00187208211013988
41. Emaminejad N, Maria North A, Akhavian R. Trust in AI and Implications for AEC Research: A Literature Analysis. In: Proceedings of the Computing in Civil Engineering 2021. doi: 10.1061/9780784483893.037
42. Luo B, Lau RYK, Li C, et al. A critical review of state-of-the-art chatbot designs and applications. WIREs Data Mining and Knowledge Discovery. 2021; 12(1). doi: 10.1002/widm.1434
43. Chaves AP, Gerosa MA. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. International Journal of Human—Computer Interaction. 2020; 37(8): 729-758. doi: 10.1080/10447318.2020.1841438
44. Zhou L, Gao J, Li D, et al. The design and implementation of xiaoice, an empathetic social chatbot. Comput Linguist. 2020; 46(1): 53-93.
45. Rahman AM, Mamun AA, Islam A. Programming challenges of chatbot: Current and future prospective. In: Proceedings of the 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). doi: 10.1109/r10-htc.2017.8288910
46. Skjuve M, Følstad A, Fostervold KI, et al. My Chatbot Companion - a Study of Human-Chatbot Relationships. International Journal of Human-Computer Studies. 2021; 149: 102601. doi: 10.1016/j.ijhcs.2021.102601
47. Følstad A, Araujo T, Law ELC, et al. Future directions for chatbot research: an interdisciplinary research agenda. Computing. 2021; 103(12): 2915-2942. doi: 10.1007/s00607-021-01016-7
48. Thorat SA, Jadhav V. A Review on Implementation Issues of Rule-based Chatbot Systems. SSRN Electronic Journal. 2020. doi: 10.2139/ssrn.3567047
49. Kumar R, Ali MM. A review on chatbot design and implementation techniques. Int J Eng Technol. 2020; 7(11): 2791-2800.
50. Nagarhalli TP, Vaze V, Rana NK. A Review of Current Trends in the Development of Chatbot Systems. In: Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). doi: 10.1109/icaccs48705.2020.9074420
51. Shingte K, Chaudhari A, Patil A, et al. Chatbot Development for Educational Institute. SSRN Electronic Journal. 2021. doi: 10.2139/ssrn.3861241
52. Casas J, Tricot MO, Abou Khaled O, et al. Trends & Methods in Chatbot Evaluation. In: Proceedings of the 2020 International Conference on Multimodal Interaction. doi: 10.1145/3395035.3425319

53. Santos GA, de Andrade GG, Silva GRS, et al. A Conversation-Driven Approach for Chatbot Management. IEEE Access. 2022; 10: 8474-8486. doi: 10.1109/access.2022.3143323
54. Abdellatif A, Costa D, Badran K, et al. Challenges in Chatbot Development. In: Proceedings of the 17th International Conference on Mining Software Repositories; 2020. doi: 10.1145/3379597.3387472
55. Ericsson KA, Simon HA. How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. Mind, Culture, and Activity. 1998; 5(3): 178-186.