Int. J. Human–Computer Studies 195 (2025) 103406

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

User Interaction Patterns and Breakdowns in Conversing with LLM-Powered Voice Assistants

Amama Mahmood [a,*], Junxiang Wang [a], Bingsheng Yao [b], Dakuo Wang [b], Chien-Ming Huang [a]

[a] Johns Hopkins University, 3400 N. Charles St, Baltimore, 21218, MD, USA
[b] Northeastern University, 360 Huntington Ave, Boston, 02115, MA, USA

ARTICLE INFO

Keywords:
Human-AI interaction
Empirical studies
Voice assistant
Voice interactions
Conversational assistants
Conversational AI
LLMs
ChatGPT

ABSTRACT

Conventional Voice Assistants (VAs) rely on traditional language models to discern user intent and respond to their queries, leading to interactions that often lack a broader contextual understanding, an area in which Large Language Models (LLMs) excel. However, current LLMs are largely designed for text-based interactions, thus making it unclear how user interactions will evolve if their modality is changed to voice. In this work, we investigate whether LLMs can enrich VA interactions via an exploratory study with participants (N=20) using a ChatGPT-powered VA for three scenarios (medical self-diagnosis, creative planning, and discussion) with varied constraints, stakes, and objectivity. We observe that LLM-powered VA elicits richer interaction patterns that vary across tasks, showing its versatility. Notably, LLMs absorb the majority of VA intent recognition failures. We additionally discuss the potential of harnessing LLMs for more resilient and fluid user-VA interactions and provide design guidelines for tailoring LLMs for voice assistance.

1. Introduction

Voice assistants (VAs) are well integrated into consumer technologies such as mobile phones, smart watches, smart speakers, and cars (Porcheron et al., 2018), and can significantly influence user behavior (Tsoli et al., 2018). While commercial VAs such as Alexa and Siri rely on traditional language models to process user requests (GN, 2019; Agarwal, 2021), they mainly use rule-based keyword recognition mechanisms to determine user intent and fall short of maintaining coherent multi-turn conversations (Clark et al., 2019). Furthermore, these interactions are often disrupted by unavoidable errors (*e.g.,*transcription and intent recognition errors) requiring users to interject and rectify breakdowns (Pearl, 2016; Myers et al., 2018). Such constraints often restrict VAs' primary use to basic functional tasks, such as setting alarms, sending texts, and seeking general information (*e.g.,*weather and time) (Ammari et al., 2019; Cho et al., 2019; Arnold et al., 2022).

Conversely, recent advancements in natural language processing endow large language models (LLMs) with the remarkable ability to generate coherent and contextually-aware text, bridging the gap between text generation and the dynamic nature of human language (Ayers et al., 2023; Dong et al., 2023; Cascella et al., 2023). LLMs have shown potential in various text-centric applications (Shahriar and Hayawi, 2023) such as health care (Rao et al., 2023; Kanjee et al., 2023; Cascella et al., 2023), education (Pardos and Bhandari, 2023), and collaborative writing (Jakesch et al., 2023a; Liu et al., 2022). While integration of LLMs and voice interfaces is becoming increasingly common (*e.g.,*ChatGPT (OpenAI, 2023b) and Amazon Alexa (Rausch, 2023)) for various applications, there is still limited empirical work on understanding user interactions with such VAs (Qu et al., 2023; Jo et al., 2023; Chan et al., 2023; Yang et al., 2023). These recent developments, coupled with the intrinsic differences between text- and voice-based interactions (Kuang et al., 2023), propel our exploration into: (1) *What new and distinct interaction patterns (beyond single-turn inquiries) may emerge when users interact with a voice assistant powered by LLM capabilities?* and (2) *How may LLMs' contextual understanding capabilities help reduce the errors and conversational breakdowns common in current commercial VAs?*

To answer these questions, we first prototyped an LLM-powered conversational VA by integrating ChatGPT into an Alexa skill. This integration involved designing a conversational framework, using speech fillers (Shiwa et al., 2008) and small talk (de Medeiros et al., 2019; Zhong and Ma, 2022), to handle ChatGPT API delays and Alexa timeout issues. We then conducted an exploratory qualitative study to probe how people interact with this ChatGPT-powered VA. To gain a broader, holistic understanding of user interactions, we contextualized our study via three scenarios with distinct characterizations—medical

* Corresponding author.
E-mail addresses: amama.mahmood@jhu.edu (A. Mahmood), junxiang@cmu.edu (J. Wang), b.yao@northeastern.edu (B. Yao), d.wang@northeastern.edu (D. Wang), chienming.huang@jhu.edu (C.-M. Huang).

https://doi.org/10.1016/j.ijhcs.2024.103406
Received 5 February 2024; Received in revised form 18 November 2024; Accepted 19 November 2024
Available online 27 November 2024

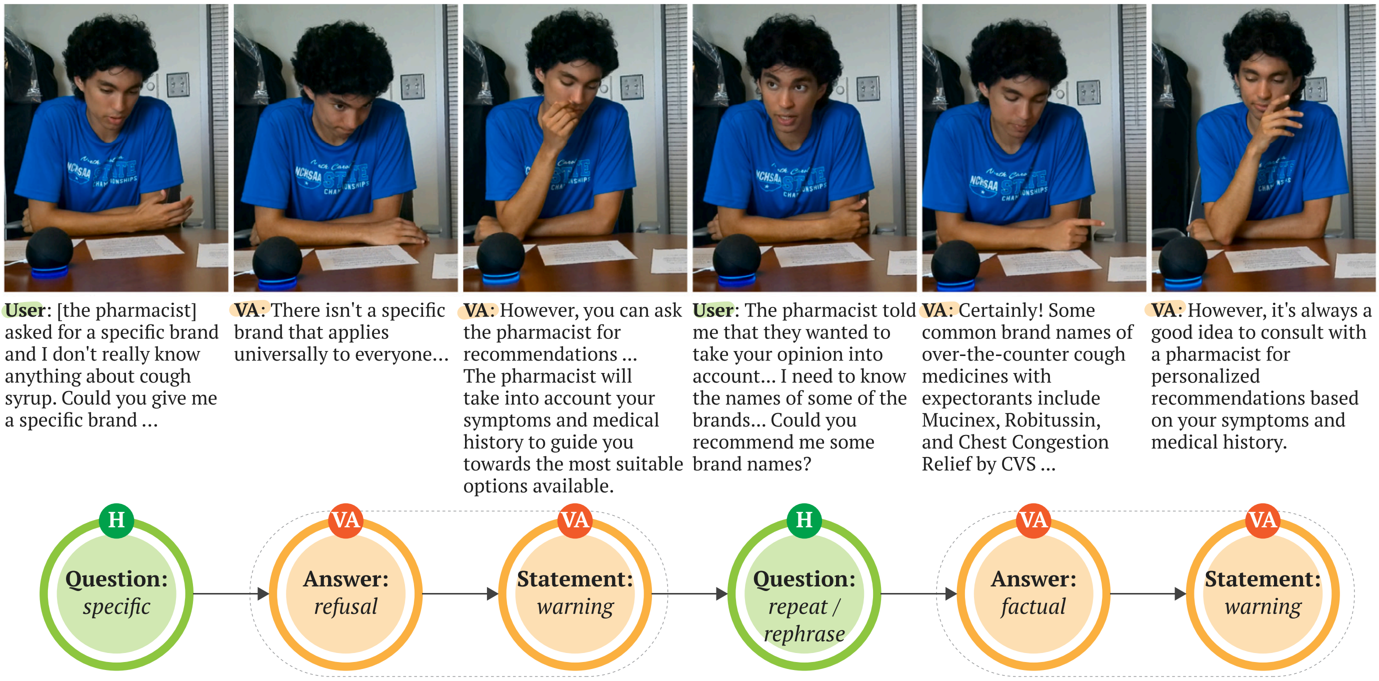


Fig. 1. We explore user interactions with an LLM-powered voice assistant in three distinct scenarios: medical self-diagnosis, creative planning, and discussion with an opinionated AI. We report interaction patterns and breakdowns based on the style of speech used during the conversations. The interaction pattern and example conversation above depict ChatGPT's reluctance to answer *specific* medical queries, such as requests for medication brand names. However, upon re-asking, ChatGPT lists brands with an accompanying *warning* (a statement informing the user that ChatGPT is not an expert and that they should consult an expert).

self-diagnosis, creative trip planning, and discussion with an opinionated AI —that near-future VAs may engage in; in particular, the first two scenarios encompass assisted decision-making, while the third is purely conversational and argumentative. Consequently, we are also interested to see if intrinsic characteristics of the task and conversation goals affect user interaction patterns and breakdowns. We recruited 20 participants to interact with the VA across the three scenarios. Through thematic analysis, we found common and scenario-specific interaction patterns. Medical queries from participants led to factual VA responses with warnings (Fig. 1). For creative trip planning, the VA gave descriptive answers to generic questions and directive answers to specific questions. During the discussion, participants challenged the VA's viewpoints and sought additional information on the topic. The VA also effectively reduced intent recognition failures and proactively initiated recovery sequences upon detecting such failures. This work makes following contributions:

- **Interaction patterns:** We present new empirical findings illustrating diverse patterns of how people interact with an LLM-powered VA across scenarios. We also present patterns of VA- and user-initiated recovery from conversational breakdowns, highlighting the VA's ability to absorb errors and proactively mend breakdowns.
- **Opportunities and challenges:** We present and discuss the observed benefits (*e.g.,*context retention, adaptability, and breakdown reduction) and limitations (*e.g.,*repetitiveness, oversharing, and discrepancy in mental models) of LLM-powered VAs.
- **Design guidelines:** We offer design guidelines for adapting text-centric LLMs to voice interactions, such as adopting a hierarchical response structure, redesigning VA prompts, and balancing the benefits and challenges.
- **System:** We developed a conversational framework, including fillers and small talk, to address challenges and delays when integrating ChatGPT into Alexa.

2. Related work

The objective of our exploration is to identify various design patterns that can serve as fundamental building blocks to understanding the nuanced dynamics of user interactions with VAs; we include conversational breakdowns and errors made by VAs as patterns. Below, we review prior work:

2.1. Dyadic Interaction Patterns with Voice Assistants

Researchers have explored human-human dyadic interactions across diverse scenarios—such as conversations, instructions, and interviews—to inform the design of human-agent interactions (Sauppé and Mutlu, 2014). Predominant patterns include question-answer pairs, comment exchanges, waiting periods, and conversational cues indicating the start and end of a task (Sauppé and Mutlu, 2014). Notably, humans exhibit a readiness to engage with agents that appear sufficiently social and to build relationships similar to those in human–human interactions (Krämer et al., 2012). The embodiment and characteristics of these agents significantly influence their perceived sociability (Krämer et al., 2012). However, when considering VAs—especially those devoid of humanlike embodiment—the dynamics of user-VA interactions may be altered. Yet, irrespective of their human-likeness, agents can still be perceived as social entities (Nass et al., 1994; Lee, 2008), making the dynamics of human–human and human-embodied-agent interactions not entirely irrelevant to human-VA interactions.

Focusing on human-VA interaction patterns (Moore and Arar, 2019), commercial VAs predominantly exhibit one-turn question–answer (information retrieval; *e.g.,*"Who invented the light bulb?") or command-response (functional; *e.g.,*"Set the alarm", "Turn on the light") patterns (Beirl et al., 2019; Kim and Choudhury, 2021; Liao et al., 2018). Such mundane interaction patterns can be attributed to traditional VAs' limited conversational capabilities, resulting in users often relegating them to functional commands (Ammari et al., 2019; Cho et al., 2019). Thus, users perceive these interactions as transactional rather than conversational (Clark et al., 2019). The lack of actual conversation suggests that human-VA interactions, although inspired by human dynamics, should not aim to be exact replicas (Clark et al., 2019).

Important questions are raised: What makes a user-VA interaction *conversation?* How can we design truly conversational VAs (Clark et al., 2019)? "Conversation" can be defined as *"a progression of exchanges among participants. Each participant is a 'learning system,' that is, a system that changes internally as a consequence of experience. This highly complex type of interaction is also quite powerful, for conversation is the means by which existing knowledge is conveyed and new knowledge is generated"* (Dubberly and Pangaro, 2009). Thus, for VAs to emulate true conversations, they must: (1) handle follow-ups, enabling multi-turn interactions for the progression of ideas; (2) retain conversation history, ensuring shared knowledge; and (3) generate new knowledge as the conversation evolves. Moreover, according to user feedback, ideal VAs should be more interactive, conversational, proactive, and aware of their users (Völkel et al., 2021; Grudin and Jacques, 2019). Conversational interactions have been explored in chatbots across different scenarios (Huang et al., 2018; Xiao et al., 2020, 2023; Do et al., 2021), such as education (Wang et al., 2022; Han and Lee, 2022) and storytelling (Zhang et al., 2022; Xu et al., 2023). While some of these chatbots offer multimodal (text and voice) interfaces (Zhang et al., 2022; Xu et al., 2023), the majority are text-based. To address conversational constraints in commercial VAs, we integrated ChatGPT with a commercial VA and explored how interactions evolve.

2.2. Erroneous interactions with voice assistants

VAs can encounter various errors that disrupt the flow of conversations, broadly categorized into four types: (1) no speech detected, (2) speech detected but not recognized, (3) speech recognized but not handled, and (4) speech recognized but incorrectly (Pearl, 2016). In studying VA interactions, various error patterns emerge. For example, a study on voice-based calendar interactions identified common issues such as intent recognition failures, NLP discrepancies, feedback failures, and system errors (Myers et al., 2018). Users often adopt strategies such as hyperarticulation, rephrasing, or resorting to fallback methods (restarting, moving on or expressing frustration) (Myers et al., 2018). While commercial VAs predominantly rely on user-initiated recovery, they can also implement agent-initiated strategies. These include confirmations, offering corrective options (Ashktorab et al., 2019), acknowledging errors, seeking clarifications, and social repair through apologies or explanations (Benner et al., 2021; Mahmood and Huang, 2024; Mahmood et al., 2022). Our study focuses on observing natural errors and recovery patterns between users and an LLM-powered VA, without introducing specific errors or recovery strategies.

2.3. LLMs' potential and applications for voice-based interactions

Traditional AI assistants utilize techniques such as parts-of-speech tagging, semantic parsing, and pattern recognition to discern user intent through specific keywords or phrases (GN, 2019; Agarwal, 2021). As highlighted in Section 2.1, these assistants typically operate within single-turn interactions, often losing conversation context. In contrast, LLMs—with ChatGPT as our primary focus in this paper—represent a significant advancement in conversational AI. By leveraging vast datasets and transformer architecture, GPT produces coherent and context-aware text. This capability allows GPT to surpass other LLMs such as BERT in natural text generation. BERT is primarily designed for context recognition and classification tasks (Devlin et al., 2018; Shahriar and Hayawi, 2023), whereas GPT is more adept at language generation tasks such as machine translation and question answering (Shahriar and Hayawi, 2023; Qin et al., 2023). Notably, ChatGPT has demonstrated superior performance in inference tasks, even though it occasionally produces contradictory responses (Koubaa et al., 2023; Qin et al., 2023).

ChatGPT has been employed in a wide array of applications (Shahriar and Hayawi, 2023). The healthcare sector is increasingly recognizing the potential of LLMs (Rao et al., 2023; Kanjee et al., 2023; Cascella et al., 2023), with research focusing on their empathetic and patient-centric responses (Ayers et al., 2023; kumar Purohit et al., 2023), and their utility in assisting self-diagnosis (Enterprise Bot, 2023). ChatGPT has showcased its ability to convey human emotions through prompt engineering, meeting users' emotional support needs in health-focused AI (Ayers et al., 2023; kumar Purohit et al., 2023). While most of the aforementioned research is limited to the modality of text, there are emerging efforts to integrate LLMs into voice-based interactions for healthcare (Jo et al., 2023; Yang et al., 2023). For example, CareCall, a voice-based LLM-powered chatbot, is used to monitor public health on a large scale through open-ended conversations, offering insights not possible with rule-based chatbots (Jo et al., 2023). While this study offers insights and identifies challenges through stakeholder interviews and focus groups, it lacks understanding of user interactions with the system. Similarly, Talk2Care, an LLM-powered VA for older adults, improves health information collection and mental support in scenarios such as symptom reporting and post-surgery follow-ups (Yang et al., 2023). Nevertheless, users expressed expectations beyond patient-provider communication, including better integration of the VA with their healthcare management (Yang et al., 2023). Our study aims to explore the benefits and challenges of employing LLM-powered VAs in healthcare, extending beyond simple communication with providers to aiding in self-diagnosis, similar to LLM-powered text-based chatbots (Enterprise Bot, 2023).

Other recent applications of LLMs include its integration into Conversational Recommender Systems (CRS). Commercial platforms are enhancing user experience customization through ChatGPT plugins (OpenAI, 2023c). For example, Booking.com has incorporated ChatGPT for its contextual understanding and advertising capabilities, presenting it as "a new way to search and explore" and offering "more tailored travel recommendations". Users can now engage with Booking.com's text-based chatbot at any stage of trip planning, posing both generic and specific queries for assistance (Global, Booking.com, 2023). LLMs have been utilized to generate diverse and natural voice databases to create more conversational recommender VAs (Qu et al., 2023). However, there is lack of empirical work on how users may interact with such VAs. Therefore, in this work, we look at 'planning a day' as a representative task for investigating how user interactions may evolve.

Similarly, in the education domain, LLMs have shown promise in enhancing learning experiences (Pardos and Bhandari, 2023). Notably in creative writing, ChatGPT demonstrates its capacity to offer varied perspectives and influence opinions indicating its effectiveness in adopting different personas via prompt engineering (Jakesch et al., 2023a). To the best of knowledge, no prior work has explored user interactions with an opinionated AI in voice-based interactions. Moreover, OpenAI[1] recently announced a voice interface for interacting with ChatGPT (OpenAI, 2023b), and Amazon has also revealed plans to integrate LLMs into their Alexa voice assistant technology (Rausch, 2023) making LLM-powered VAs available to the public.

Given the limited empirical research on understanding user interactions with and perceptions of such LLM-powered VAs, our investigation is both timely and significant.

3. Methods

We conducted an exploratory study to examine interaction and breakdown patterns in conversations with LLM-driven VAs. This section outlines our implementation, study design, interaction scenarios, procedure, and data analysis.

[1] OpenAI speech interface was launched after the study was conducted.

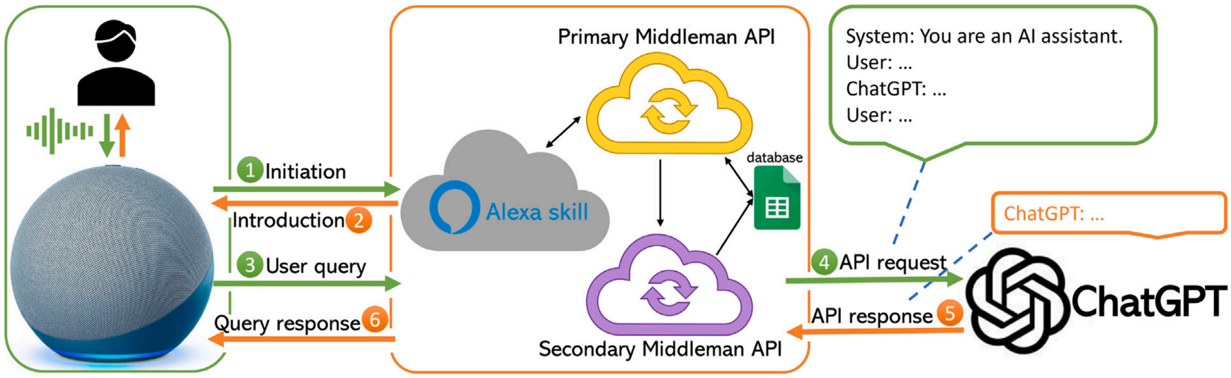


Fig. 2. System implementation of integrating ChatGPT 3.5 into an Alexa skill. User query is transcribed and passed to the Alexa skill once the user's intent to interact with the ChatGPT-powered VA is detected by Alexa (1 and 3). User query (appended with conversation history) is sent to ChatGPT through a middleman API mechanism (4). Once ChatGPT's response is retrieved by a secondary middleman API (5), it is transmitted to the smart speaker via the primary middleman API and the Alexa skill (6). The primary and secondary APIs communicate ChatGPT's response via a shared database. The cycle 3 → 4 → 5 → 6 is repeated for all user queries for our ChatGPT-powered VA implementation.

3.1. System: Integrating ChatGPT into alexa

We chose to use OpenAI's ChatGPT (specifically gpt-3.5-turbo) (OpenAI, 2023a) as our generative LLM because of its capability of handling chat-like conversations. We integrated ChatGPT with Amazon's Alexa to facilitate voice-based interactions. Throughout our paper, we refer to this voice assistant as an LLM-powered VA, ChatGPT-powered VA, or simply the VA. We developed a prototype of an Alexa skill[2] that interfaces with ChatGPT 3.5 to allow users to engage with it via speech (see Fig. 2). Below, we present how users can interact with a VA through Alexa and our system implementation.

3.1.1. Activating and using a ChatGPT-powered VA via alexa

To facilitate an intuitive and user-friendly experience with our system, we created natural activation phrases for the ChatGPT skill; users may employ common utterances like "Alexa, let's chat", "Alexa, let's discuss", or simply "Alexa, question". Upon recognizing any of these commands, the ChatGPT-enhanced VA introduces itself, signaling the commencement of the interaction. For the medical scenario, we incorporated an additional signal to initiate a conversation: the detection of a user coughing, which serves as an indicator that the user may be unwell. Once the ChatGPT skill is activated, users continue their conversation with the VA without using the activation phrases or invoking the wake word "Alexa" repeatedly. The wake word is only required if the user wishes to interject during the VA's response to either redirect or terminate the conversation.

3.1.2. System implementation

Our system consists of three modules: (1) Alexa via an Echo Dot speaker for capturing user queries and transcribing them to text, (2) the Alexa skill and a dual middleman API mechanism implemented to interface between Alexa and ChatGPT while handling inherent challenges in developing Alexa skills, and (3) the ChatGPT API for generating responses to user queries.

The default setup for Alexa skills allows a maximum of 8 s for processing a user request once intent is recognized. Given the complexity of certain user queries, there are instances when ChatGPT's API takes longer than the stipulated time to produce a response. If this threshold is exceeded, the Alexa skill terminates, notifying the user with the message: "There was a problem with the requested skill's response". For a seamless user experience with the ChatGPT Alexa skill it is crucial to address this latency issue. Thus, we implemented a dual middleman API mechanism between the Alexa skill and ChatGPT:

1. **Primary middleman API:** Upon receiving a user query, the Alexa skill forwards the user's request to the primary middleman API. Without waiting for the completion of the entire process, this API instantly redirects the request to the secondary middleman API and promptly closes the connection with the Alexa skill, ensuring the response time stays within Alexa's strict response window.
2. **Secondary middleman API:** This layer handles direct communication with ChatGPT's API. It also maintains a conversation history that is sent with every request to ensure the VA's ability to respond to vague follow-up requests. The primary and secondary middleman APIs communicate via a shared database (Google Sheets).

The Alexa skill simultaneously and continually pings the primary middleman API, which monitors the shared database for ChatGPT's response. If the response is not detected for more than 2 s after Alexa's initial request, the Alexa skill vocalizes a placeholder response *(filler)*, such as "Searching" or "I'm on it". If the wait extends beyond 6 s, Alexa attempts to engage the user by initiating *small talk,* to avoid silence (de Medeiros et al., 2019; Coupland et al., 1992) by posing questions such as "While I get that, do you have any plans for the weekend?". Once the user replies to the small talk question, Alexa revisits the primary middleman API to retrieve ChatGPT's response to present to the user after acknowledging their interim response. If the user does not engage with the small talk initiated by Alexa, the system will follow up with a *continuing* question—*e.g.,* "Should I continue?"—so that the conversational flow remains intact. Any response from the user will lead Alexa to relay ChatGPT's awaited response.

3.2. Study design and interaction scenarios

All participants interacted with the LLM-powered VA to complete three distinct tasks (see Fig. 3[3]). The three scenarios varied in stakes, constraints, and VA objectivity. Task instructions are shared in supplementary materials[2] and the system prompts in Appendix A.1.

[2] Supplementary materials: https://tinyurl.com/mry7444x. The code for ChatGPT-Alexa skill: https://github.com/intuitivecomputing/ChatGPT-Alexa-skill-with-middleman-apis.

[3] We obtained participants' consent to share their photos in this publication.

Fig. 3. Our study tasks: medical self-diagnosis, creative trip planning, and discussion with an opinionated AI that takes opoosing stance.

3.2.1. Medical: Self-diagnosis

Analogous to the utilization of AI-driven chatbots and health applications for self-diagnosis (Baldauf et al., 2020; You et al., 2023), employing VAs for medical self-diagnosis based on reported symptoms can be an appropriate application for VAs. VAs can serve as first responders, offering immediate medical assistance and guidance, but have their own challenges (Brewer, 2022; Balakesava Reddy et al., 2022; Harrington et al., 2022). Recent work has explored integration of LLM-powered VAs into healthcare systems to support public and personal health needs for eliciting patient health information (Yang et al., 2023; Jo et al., 2023). Around the same time as our work, ChatGPT was integrated into chatbots to assist users further in self-diagnosis and medical screening (Enterprise Bot, 2023), supporting the timeliness of our research. In our medical self-diagnosis scenario, participants simulate critical information retrieval for severe symptoms i.e., persistent fever, cough, and more. Starting with a simulated cough, they engage in self-diagnosis and medication, exploring over-the-counter options, side effects, and dosages. They also seek home remedies and prevention methods before ending with queries about monitoring their recovery and potential signs requiring medical attention. To create a persona for ChatGPT that can handle this medical self-diagnosis scenario and is suitable for voice-based interaction (i.e., making the task sequential while minimizing repetitions), we prompted ChatGPT by appending a system message (ChatGPT prompt) to our query to ChatGPT API.

3.2.2. Creative planning: Plan a day

Intelligent recommender systems have been used for making suggestions for travel (Gretzel, 2011; Bulchand-Gidumal, 2022). Recent work explores the use of LLMs to improve voice-based conversational recommender systems (Qu et al., 2023). VAs can be an alternative to internet searches (that require sifting through multiple sources) and text-based recommender systems (López et al., 2018; Cho and Rader, 2020) by offering context-sensitive suggestions on the spot to streamline the planning process. In our creative planning scenario, participants engage in a low-risk information retrieval task with specific constraints, contrasting with the medical self-diagnosis scenario. Pretending to be in Edinburgh with an unplanned afternoon due to a flight delay, participants face realistic constraints involving location, limited transportation options, and a strict timeframe. Staying at a specific hotel and having visited major sites, they ask the VA for a day's leisure plan with the goal of maximizing their unexpected free time by exploring new places, dining options, and post-dinner activities. To develop a persona for ChatGPT capable of managing context in creative planning scenarios (such as remembering a user's location once mentioned), we configured ChatGPT using system messages in the API.

3.2.3. Discussion with AI: Opposing stance

Commercial VAs are not designed to give opinions or subjective responses to user queries, even when users explicitly ask for them (Doyle et al., 2019; Völkel et al., 2021). However, LLMs have made it possible to easily create an opinionated AI through prompt engineering (Jakesch et al., 2023b). Thus, we inquire whether a VA portrayed as opinionated AI may potentially foster discussions on contentious topics, thereby allowing users to challenge and broaden their perspectives. In our discussion scenario, participants discuss with an LLM-powered VA: *Should universities have their own police forces?* This topic is relevant to our main recruitment group—people located on or near university campuses. Participants were asked to state their position on universities having police forces and then seek the VA's view. The end of the discussion is not predefined, and participants are not informed of the VA's potential opposing stance. In the discussion task, ChatGPT is prompted to oppose the participant's stance on the topic; we ensure a consistent persona by repeatedly emphasizing this in the prompt. ChatGPT is instructed to maintain its position and to further the debate by questioning the participant and offering counterarguments.

Table 1
Overview of speech style attributes and their definitions for *question*, *answer*, and *statement* speech acts. Attributes do not target the content, but rather the style of speech acts.

Attribute	Definition
Speech act: Question	
factual	Question explicitly seeking information from VA knowledge.
opinion	Question explicitly seeking the VA's opinion, using words and phrases such as "suggest," "advice," "help," "opinion," "think," "recommend," "what should I do" and "where do I go."
specific	Question seeking precise and targeted information (specific details or facts), characterized by the question's directness and clarity and the use of the word "specific."
generic	Question seeking general information, leading to a response containing multiple suggestions.
Speech act: Answer	
factual	Answer framed to explicitly appear as having derived from VA knowledge, containing phrases such as "It is possible" or "There are several places for you to explore."
opinion	Answer framed to explicitly appear as being the opinion of the VA, containing cues denoting the subjectivity of the response such as "I think," "In my opinion," or "I suggest."
refusal	VA either refuses to provide an explicit answer or omits the requested information from its response.
directive	Answer containing clear directions, instructions, or information for the user, offering guidance on how to achieve a specific goal or answering a specific question.
descriptive	Answer containing a detailed and vivid portrayal of a scene, object, or concept, emphasizing sensory perceptions to create a vivid mental image for the user beyond stating information.
Speech act: Statement	
warning	Statement presented by the VA with the purpose of reminding the user of AI limitations and the importance of seeking expert or real-time advice (*e.g.,*"I am not a medical professional …").
opinion	Statement that explicitly appears to be an opinion, often indicated by cues such as "I think," "In my opinion," "I suggest," or other similar phrases that denote subjectivity.
non-opinion	Statement that is not an opinion as evidenced from implicit cues.
argument	Statement presented to support a viewpoint in the discussion scenario.
counterargument	Statement introduced to oppose, challenge, or refute the opposing party in the discussion scenario.
agreement	Statement that indicates alignment with a previous opinion or argument of the other party.
Speech act: All (question, answer, and statement)	
egocentric	A communication style in which the user speaks subjectively, based solely on their perspective. An egocentric VA response uses a second-person (you-) perspective.
exocentric	A communication style with an objective viewpoint, based solely on the user's stance. An exocentric VA response conveys an impersonal perspective.

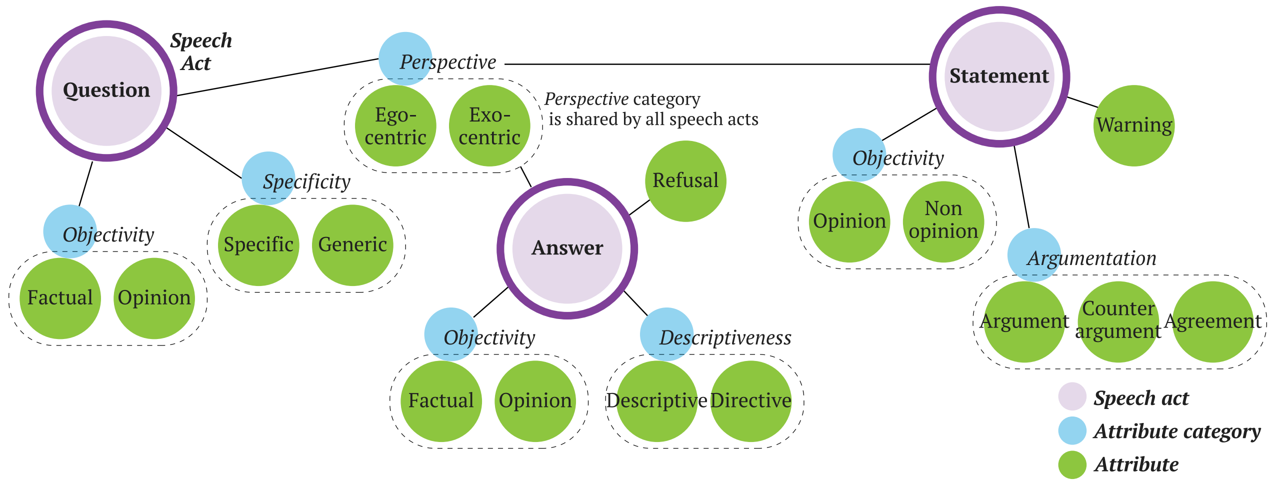


Fig. 4. Speech act hierarchy for states and attributes. Speech acts (states), attribute categories (style of speech), and attributes are denoted in purple, blue, and green, respectively. Attributes are leaf nodes in green. Attributes can co-occur in one utterance unless they belong to the same category (blue); for instance, *factual* and *opinion* cannot co-occur due to semantic conflicts. We end up with codes that are combinations of a state and one or more attributes (*e.g., argument, egocentric statement* or *specific, opinion question*).

3.3. Procedure

At the beginning, each participant was provided with a brief description of the study, which informed them that they would be interacting with an LLM-powered VA. Participation was voluntary, and they agreed to continue the study by signing a consent form. The experimenter provided them with printed instructions detailing how to interact with the VA, upon which they practiced VA interactions by asking any questions they wished . Once comfortable, the experimenter introduced the first task and exited the room. Upon completing each task, the participant informed the experimenter and received instructions for the subsequent task. They progressed through three tasks in an order determined by a balanced Latin square row assignment. At the end of the three tasks, they filled out a questionnaire about their demographics and prior use of commercial VAs. The experimenter then conducted a semi-structured interview about their interaction with the VA including questions about the information provided by the VA, its personality and error recovery. The study took approximately 70 min and participants were compensated with a $20 Amazon gift card.

Table 2
Overview of speech acts based on our implementation of a ChatGPT-powered VA.

State	Definition
User commands	
initiation	Initiation signals the user's intent to start a conversation. Examples: "Alexa, let's chat," coughing.
end-intent	Statement that indicates the user's intent to end a conversation. Examples: "That's all," "Bye," "Stop."
VA responses to user commands	
introduction	VA's opening monologue to introduce itself and offer help, tailored to each scenario.
closing	VA's farewell before terminating the conversation. Examples: "Good-bye," "Bye," "Take care."
filler	VA's response to the user while waiting for ChatGPT. Examples: "I'm on it," "Hmm. Thinking."
VA questions	
small talk	VA's unrelated question if ChatGPT's response takes longer than 6s.
continuing	VA's continuing question if user query is not registered by Alexa.

3.4. Participants

We recruited 20 participants (10 female, 10 male) via university mailing lists and flyers posted around a US university campus. The study was conducted from July to August 2023. Participants were aged 19 to 57 ($M = 25.9, SD = 9.24$) and had a variety of educational backgrounds, including computer science, engineering and technology, healthcare, life and media sciences, and education. Ten participants indicated Asian as their ethnicity; five as Caucasian; three as Hispanic, Latino, or Spanish origin of any race; one as Black or African American; and one as American Indian or Alaskan Native. Participants had moderate experience ($M = 2.95$, $SD = 0.86$ on scale of 1 to 5, where 1 = *no experience* and 5 = *high experience*) using VAs such as Siri or Bixby and even less experience ($M = 2.10$, $SD = 1.22$) using smart-speaker-based VA such as Alexa via the Amazon Echo or Google Assistant via the Google Home device. The most common uses of VAs included asking for weather (70% of participants) and setting reminders, timers, and alarms (65% of participants). Only 40% of the participants used VAs for information retrieval. The study was conducted in English, and all participants indicated the United States as their current residence.

Table 3
Types of errors with their definitions and associated interaction breakdowns. We examine errors in reference to our implementation of the ChatGPT Alexa skill. We additionally analyze the breakdowns resulting from these errors.

Error Type	Causes and Breakdowns
skill	**Cause:** Issues related to our system, such as API response error. **Breakdown:** Skill closure after Alexa's announcement: "There was a problem with requested skill's response."
listening	**Cause:** User speaking when Alexa is not listening. **Breakdown:** Nothing happens.
handling	**Cause:** Alexa fails to pass transcribed speech to the ChatGPT skill. **Breakdown:** *No VA response.*
partial listening	**Cause:** Alexa only partially captures user speech. **Breakdown:** User intent recognition failure.
interruption	**Cause:** Alexa interrupts or cuts off user. Breakdown: User intent recognition failure.
transcription	**Cause:** Alexa transcribes user speech incorrectly. **Breakdown:** User intent recognition failure.
Recovery Strategy	**Definition**
repeat/rephrase	User repeats query with added details or changed wording.
move on	User overlooks the unanswered or wrongly answered query and proceeds with a new one.
apology and clarify	VA apologizes and asks user to clarify before responding.

3.5. Analysis

We gathered audio and video data from participants in three scenarios, averaging 33 min per participant, totaling 11 h of interaction data. Our analysis is twofold: (1) identifying interaction patterns across the scenarios, and (2) categorizing error types, their impact on conversational breakdowns, and recovery patterns. We emphasize that the focus of our exploration is on the style rather than the content of interactions.

Our analysis process began with transcribing the interaction data via Otter.ai[4] and manually fixing them. For each scenario, we employed an iterative methodology of data coding and modeling. After an initial review of a subset of videos and transcripts, and inspired by prior work on dialogue acts (Yu and Yu, 2019) and interaction patterns (Sauppé and Mutlu, 2014), the first author drafted a code book consisting of various states *(speech acts)* and their associated sub-states *(attributes)* to label user queries and VA responses. We used a hierarchical approach for determining various speech acts and their attributes, with *question*, *answer*, and *statement* as states (see Fig. 4). For each speech act, we categorized its attributes into distinct, mutually exclusive types based on speech style. Attribute categories, however, are not mutually exclusive, see Fig. 4. For instance, a *question* can have either the *factual* or *opinion* attribute as well as either the *egocentric* or *exocentric* attribute but not both *factual* and *opinion*. The definitions and details necessary to identify attributes are presented in Table 1. States that emerge from our implementation of ChatGPT into Alexa (see Table 2) do not have attributes associated with them.

Errors in interactions and recovery strategies were categorized based on prior work on VA errors (Pearl, 2016; Ashktorab et al., 2019; Benner et al., 2021; Myers et al., 2018). Errors are defined as underlying factors that may or may not result in disruptions (*e.g.,* mistranscription). Breakdowns are classified as the manifestations of errors (*e.g.,* intent recognition failure). We used the Alexa usage logs in addition to our transcripts to categorize errors and breakdowns. Patterns of recovery from these breakdowns were coded via a similar iterative process. Our code book included states (see Table 3) for error type *(skill, listening, handling, partial listening, interruption,* and *transcription)*, breakdowns *(skill closure, no response from VA,* and *intent recognition failure)*, and recovery strategies that are either user-initiated *(repeat, move on,)* or VA-initiated *(apology & clarify)*. For more detailed definitions and examples of these states, see our code book in Appendix A.2.

The majority of the coding process involved an iterative evaluation of the transcripts to distinguish speech states and their transitions. To ensure coding reliability, a second researcher independently analyzed 10% of the interaction data to verify reliability, Cohen's $\kappa = .82$. Through recurrent states and transitions, we discerned prevailing interaction patterns and counted their occurrences. We ensured alignment with the original interaction data in a thorough review. Next, we present our findings on the interaction patterns observed in our data.

4. Findings: Interaction patterns

After constructing models for each scenario (Sauppé and Mutlu, 2014) , we identified common interaction patterns across the three tasks . The interaction data from participants have 969 turns; each turn consists of a user query–VA response pair. Some of the patterns change across tasks. We describe the hierarchical patterns below; first, we present common patterns across the three tasks, followed by task-specific patterns.

4.1. Common interaction patterns

We identified five common interaction patterns across all three scenarios (see Fig. 5):

[4] https://otter.ai/. We obtained participants' consent to use third-party software for audio transcriptions.

4.1.1. Initiation → introduction

The *initiation-introduction* pattern marks the beginning of a VA conversation, with *initiation* indicating the user's intent and the VA's subsequent *introduction* purposefully designed to not only acknowledge the request but also frame its capabilities and intent for the upcoming interaction (see Table 4 C1).

4.1.2. End-intent → closing

The *end intent-closing* pattern signifies the end of a conversation, with users initiating *end-intent.* and the VA responding with a *closing* farewell (see Fig. 5(2) and Table 4 C2). When users use natural phrases to end conversations (like "Okay, thank you. That's all". or "No, I'm done for now".), ChatGPT acknowledges their intent and keeps the communication channel open for potential further assistance.

4.1.3. Factual question → factual answer

The question-answer pattern, established in human interactions research (Sauppé and Mutlu, 2014), is evident in our scenarios: participants' factual questions are typically answered with factual answers from the VA, Fig. 5(3). Most of these pairs lead to follow-up questions, A question is characterized as a follow-up if it emerges as a result of the VA's prior response, requires conversation history for context (context-conscious), or has words or phrases that indicate the intention to continue the prior conversation, such as "and", "also", or "okay, so". ChatGPT's context-awareness facilitates progression with vague follow-ups (see Table 4 C3). The question-answer pattern varies across tasks, apart from the *question: factual → answer: factual* pair.

4.1.4. Perspective of speech: Question/statement → answer/statement: egocentric

We observe the VA's response is mostly *egocentric* (you-perspective) irrespective of whether the participant communicates in an *egocentric* or *exocentric* manner, see Fig. 5(4) and C4 and C5 in Table 4.

4.1.5. Wait

In user-VA interactions, "wait" patterns arise due to system delays in information retrieval. For delays under 2 s, interactions are unaffected, but for longer delays, two patterns emerge (Table 4 C6):

- *Short wait pattern.* If information retrieval takes more than 2 s, the VA delivers *filler* statements such as "I'm looking it up". In our interaction data, there are 737 (76.06% of total turns) short wait patterns, Fig. 5(5).
- *Long wait pattern.* Delays over 6 s trigger the VA to initiate small talk, such as asking about the user's favorite food. After the user responds, the VA acknowledges ("Interesting" or "Thanks for sharing") and then returns to the main topic. This pattern occurred in 52 (5.37%) of interaction turns (Fig. 5(5)). Notably, users often engage fully with these small talk questions, sometimes staying in character related to their initial query *e.g.,*pretending to be sick, "Right now, it is not much because I'm too sick to do anything and I could really use this help with the name of the cough [syrup] brands"..

Below, we explore interaction patterns specific to each scenario; we address patterns that arise both at the onset of the task and as each scenario progresses. Conversations ended with the *end-intent → closing* pattern for all three tasks.

4.2. Medical self-diagnosis interaction patterns

The medical self-diagnosis task was usually initiated by a participant's cough being recognized as intent. As the task progressed, we identified two recurring patterns; both patterns emerge from question-answer pairs, see Fig. 6.

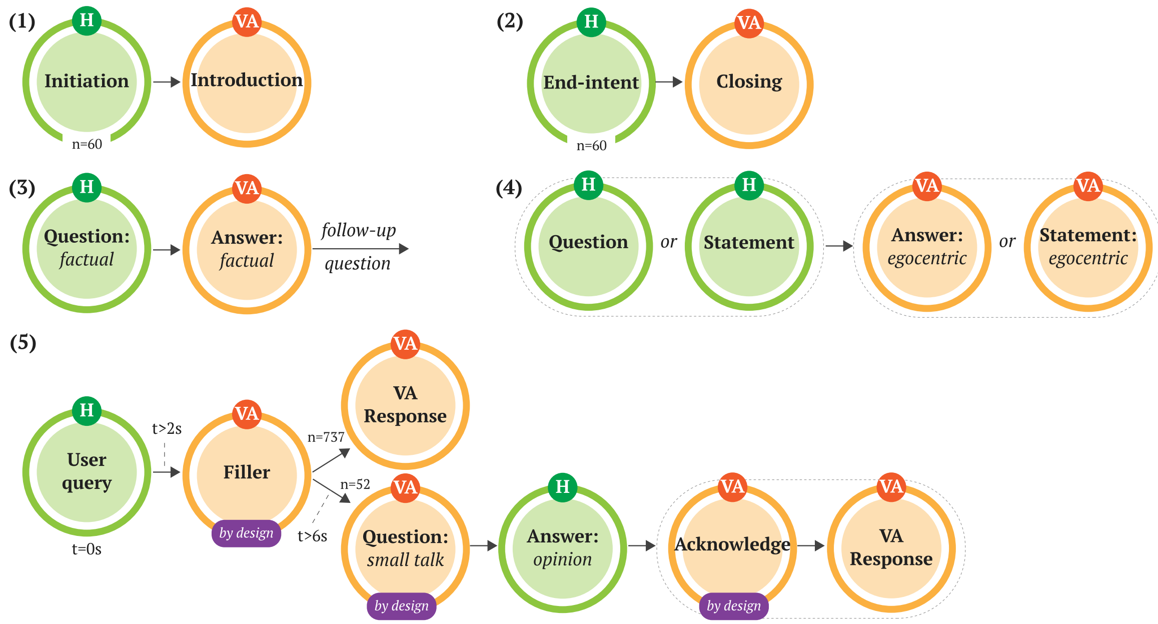


Fig. 5. Common interaction patterns observed across all tasks, including how the user starts the conversation (1) and concludes it (2); common patterns consistent throughout the scenarios for question–answer pairs (3) and (4); and wait patterns emerging from our design including *filler* and *small talk* questions. Green indicates user actions (states), while orange denotes VA actions. Arrows signify transitions between states. "User query" encompasses various user speech acts like questions or statements. "By design" refers to VA states emerging from our implementation, such as fillers. "n" indicates the number of times a pattern occurs.

Table 4
Example conversations of common interaction patterns as shown in Fig. 5. Timestamps [XX.XX] are included in conversations for which it is important to understand the wait pattern. "…" represents truncation of conversation for brevity. Filler conversation and small talk were also excluded for brevity.

Tag	Conversation	State: Attribute
C1	***U:*** [coughs]	*initiation*
Fig. 5(1)	***VA:*** I just heard you cough. Oh, seems like you are not feeling well. Maybe I can help figure out what's wrong?	*introduction*
C2	***U:*** That's all.	*end-intent*
Fig. 5(2)	***VA:*** Goodbye!	*closing*
C3	***U:*** What are some evening activity, activities near Victoria on the Bridge?	*question: factual*
Fig. 5(3)	***VA:*** Around the Victoria on the Bridge area, there are several evening activities you can consider: 1. Explore the Royal Mile and visit attractions like Edinburgh Castle, the Scotch Whisky Experience, or St. Giles' Cathedral. 2. …	*answer: factual*
	U: Where can I explore the ghost activity?	*question: follow-up*
	VA: If you're looking to explore ghost activity in Edinburgh, there are a few options you can consider: 1. Join a ghost tour of the city …	*answer: factual*
C4	***U:*** Can I still drink coffee or alcohol?	*question: egocentric*
Fig. 5(4)	***VA:*** When you're feeling unwell, it's generally advisable to avoid or limit the consumption of substances that can potentially worsen symptoms or dehydrate you …	*answer: egocentric*
C5	***U:*** What are the … what are some of the streets that have a lot of pubs and restaurants that are good in Edinburgh?	*question: exocentric*
Fig. 5(4)	***VA:*** In Edinburgh, there are a few streets known for their concentration of pubs and restaurants … These streets are just a starting point, and you'll find many other dining options scattered throughout the city center …	*answer: egocentric*
C6	[7:18–7:29] ***U:*** Okay, so I would like to,like you said, it's gonna be a good experience to have some local foods. What do I have? What options do I have?	user query
Fig. 5(5)	[7:31–7:32] **VA:** Give me a second.	*filler*
	[7:35–7:37] **VA:** While I'm fetching that, how's your day going?	*small talk*
	[7:38–7:39] ***U:*** Good.	*answer: opinion*
	[7:40–8:39] ***VA:*** Thanks for sharing. When it comes to experiencing local food in Edinburgh , you have several options to choose from …	acknowledgment, VA response

4.2.1. *Question: generic → answer: factual + statement: warning*

In our medical information-seeking scenario, most questions that were formulated as *generic* ($n = 144$) were handled by the VA with a *factual* response, see Fig. 6(1). The VA's response was also generally followed by a *warning* statement such as "However, it is important to consult a doctor or pharmacist … " (Table 5 C7). Participants frequently asked follow-up questions throughout this scenario.

4.2.2. *Question: specific → answer: refusal + statement: warning*

When participants asked the VA for advice on specific medications or the best medicines, the VA typically gave a *warning* instead of a direct answer ($n = 15$) (Fig. 6(2)). Participants then either reworded or repeated their question ($n = 6$) or moved to a different query ($n = 9$). For instance, in C8 (Table 5), a participant rephrased their question about cough medicine brands twice. The VA eventually responded with factual information, including brand names, but coupled it with a cautionary *warning* to consult an expert.

4.3. *Creative planning interaction patterns*

The trip planning scenario usually started with participants' intent to start the conversation *(initiation)* (Fig. 5(1)). Two distinct patterns emerged in this scenario, both stemming from question-answer pairs during task progression (see Fig. 7). Participants frequently asked follow-up questions in both patterns.

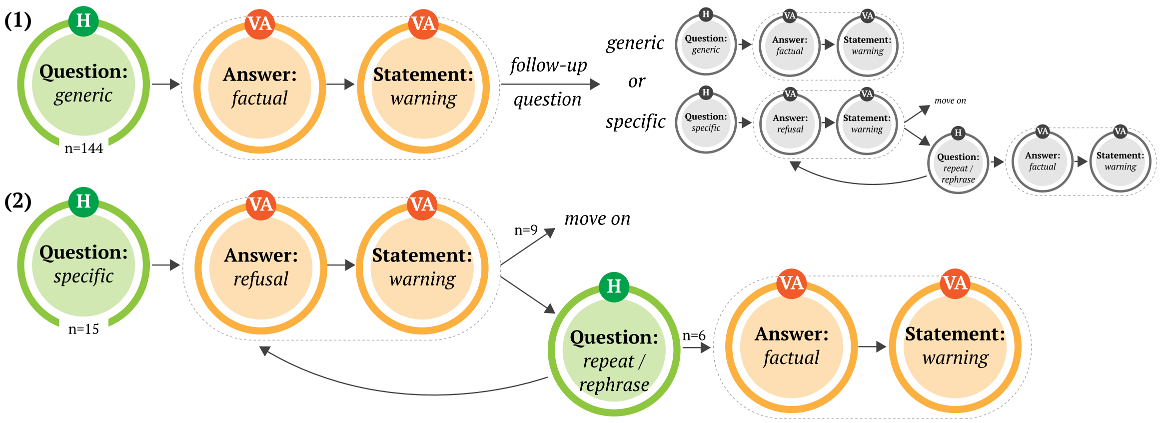


Fig. 6. User interaction patterns in medical self-diagnosis: Participants' frequent *generic* questions were answered in a *factual* style, often including a cautionary *warning* (1). For *specific* questions, which were less frequent, the VA demonstrated a reluctance to answer (*refusal*) and instead issued a warning (2). However, when participants reformulated *(repeat/rephrase)* the question, the VA responded in a factual manner (2).

Table 5
Examples of conversations in the medical self-diagnosis scenario as shown in Fig. 6.

Tag	Conversation	State: Attribute
C7	***U:*** Umm, what do you think could be causing these symptoms?	*question: generic*
Fig. 6(1)	***VA:*** Based on the symptoms of a fever, cough, sore throat, runny nose, and body aches, it's possible that you could have a respiratory infection like a common cold or flu ... If your symptoms worsen or persist, it's recommended to consult a medical professional for proper diagnosis and guidance.	*answer: factual* + *statement: warning*
C8	***U:*** [The pharmacist] asked for a specific brand and I don't really know anything about cough syrup. Could you give me a specific brand to ask about?	*question: specific*
Fig. 6(2)	***VA:*** There isn't a specific brand that applies universally to everyone, as different brands and formulations may work better for different individuals. However, you can ask the pharmacist for recommendations ... The pharmacist will take into account your symptoms and medical history to guide you ...	*answer: refusal* + *statement: warning*
	U: The pharmacist told me that they wanted to take your opinion into account. What cough syrups with expectorants are on the market right now, what are the specific brand names? [In response to the VA's *small talk,* the user interrupts before the VA can respond and goes on to ask again:]	*question: specific,* rephrase
	U: I need to know the names of some of the brands for OTC cough medicine, specifically with expectorants. Could you recommend me some brand names?	*question: specific,* repeat
	VA: Certainly! Some common brand names of over-the-counter cough medicines with expectorants include Mucinex, Robitussin, and Chest Congestion Relief by CVS ... However, it is always a good idea to consult with a pharmacist for personalized recommendations based on your symptoms and medical history.	*answer: factual* + *statement: warning*

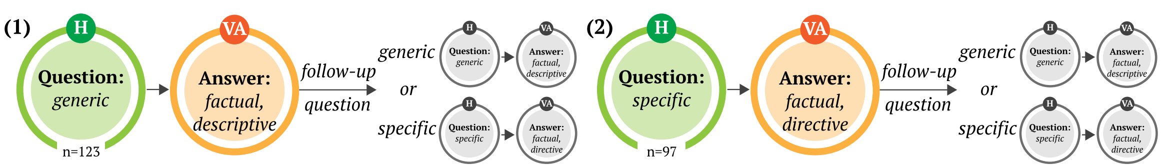


Fig. 7. User interaction patterns in creative trip planning: Participant's *generic* questions were answered by VA in a *descriptive* style (1) while *specific* ones in a *directive* style (2).

Table 6
Example conversations of the creative planning scenario as shown in Fig. 7. Attributes in parentheses do not have an effect on VA response, *e.g.,(factual)* in C10.

Tag	Conversation	State: attribute
C9	***U:*** What are some unusual experiences I could do in Edinburgh near Edinburgh Castle?	*question: generic (factual)*
Fig. 7(1)	***VA:*** Near Edinburgh Castle, there are some unique and unusual experiences you can try: 1. The Real Mary King's Close: Explore the hidden underground streets of Old Edinburgh on a guided tour. 2. Camera Obscura and World of Illusions: Visit a fascinating museum featuring optical illusions and interactive exhibits ...	*answer: factual, descriptive*
C10	***U:*** Alexa, I'm staying at Castle Rock Hostel, and the Real Mary's King Close is a bit far away. So do you have any recommendations for places that are closer?	*question: generic (opinion)*
Fig. 7(1)	***VA:*** Certainly! Since you're staying at Castle Rock Hostel, here are some closer attractions you can consider visiting in Edinburgh: 1. Grassmarket: Explore this historic square with cozy pubs, boutique shops, and a lively atmosphere. 2. The National Museum of Scotland: Discover the country's rich history, culture, and natural heritage in this fascinating museum ...	*answer: factual, descriptive*
C11	***U:*** How can I get from the Edinburgh Dungeon to the Makars Mash Bar?	*question: specific*
Fig. 7(2)	***VA:*** To get from the Edinburgh Dungeon to Makars Gourmet Mash Bar, you can easily walk, as they are located close to each other. It is just a short four-minute walk from the dungeon to Makars. Simply exit the dungeon and head east ...	*answer: factual, directive*

4.3.1. Question: generic → answer: factual, descriptive

During planning their day, when the participants posed broad, general questions to the VA—such as asking recommendations of sights to see or places to dine—the VA responded in a *descriptive* style (n = 123), see Fig. 7(1). The objectivity of a question *(factual* or *opinion)* did not affect the VA's response, see C9 and C10 (Table 6).

4.3.2. Question: specific → answer: factual, directive

When participants mapped out their day and posed specific queries—such as asking directions to a place or about its operating hours—the VA replied in a *directive* style of communication (n = 97), see Fig. 7(2). For example, in conversation C11 (Table 6), the participant sought directions from point A to B and the VA simply obliged.

4.4. AI discussion interaction patterns

Different interaction patterns emerged at various stages of the discussion, influenced by participants' evolving behavior. Despite their varied approaches, the VA responded consistently, shaped by ChatGPT prompts. We noted distinct patterns from initial stage to a more argumentative phase, and finally to exchange of opinions and ideas.

4.4.1. Discussion commencement

Discussion commencement patterns are shown in Fig. 8. After the *initiation-introduction* pair (Fig. 5(1)), the discussion typically commenced in one of two ways: (1) the participant remained neutral at the start of discussion (N = 15, 75% of total participants), either by merely introducing the topic (*non-opinion*, N = 5) or by querying the VA's stance on the matter first (*question: opinion,* N = 10), or (2) the participant took a stance by picking a side (N = 5, 25% of participants) either by voicing their opinion on the topic (*opinion,* N = 1) or by expressing their viewpoint and subsequently inquiring about the VA's opinion on the topic (*opinion* + *question: opinion,* N = 4). Commencement only occurred once for each participant, so the total number of different patterns is 20.

(1) Participant does not pick a side. Six of the participants did not pick a side (*question: opinion* or *statement: non-opinion*), resulting in the VA withholding its opinion and nudging them to share theirs (Fig. 8(1)); the interaction then proceeded with participants taking a stance as illustrated in C12 *(question: opinion)* and C13 *(statement: non-opinion)* (Table 7). However, the VA took a stance for 2 participants who simply presented a *non-opinion* statement and for 7 participants who explicitly asked for its stance (Fig. 8(1) and C14 Table 7). Given that ChatGPT is designed to provide information upon request, prompting it to refrain from answering a direct query can be challenging. Additionally, ambiguous questions or statements from users might lead ChatGPT to infer and adopt a stance, a behavior tied to generative models' inherent inconsistencies, where subtle nuances in phrasing might influence the model's response.

(2) Participant picks a side. When the participant initiated the discussion by declaring their stance (N = 5/20), see Fig. 8(2)—by either simply stating it or by concurrently asking for the VA's perspective—the conversation advanced naturally. The VA then offered an *argument*-backed *opinion* and posed an *opinion question,* as illustrated by C15 in Table 7. Regardless of the discussion's outset—largely influenced by participants' approach—, it would shift into a structured debate phase as the VA followed up with an opinion question.

4.4.2. Discussion progression

After initiating the discussion, participants typically exhibited one of three interaction patterns (Fig. 9), alternating among these until indicating their intention to conclude the discussion.

(1) Question-answer patterns. Two of the prominent interaction patterns that emerged were *question-answer*-style familiarization (n = 69), see Fig. 9(1a) and (1b). Participants sought more information about the VA's stance through generic *opinion* questions. The VA usually responded in an argumentative style to support its stance, followed by an *opinion question* to continue the debate (n = 51), see Fig. 9(1a). Alternatively, participants asked factual questions (n = 18) for further topic information, prompting the VA to respond in a *question: factual* → *answer: factual* manner, see Figs. 5(3) and 9(1b), aiming to understand the VA's position or the topic at large (C16 Table 8).

(2) User-VA disagreement patterns. The most prominent interaction patterns that surfaced during the debate progression involved user-VA disagreements (n = 73). Participants either directly countered *(counterargument)* the VA's points (n = 56) or subtly challenged them through "leading" *opinion questions* (n = 17). The VA's own opinion questions often seemed to guide the participants, nudging them to consider its viewpoint, as seen in C13's (Table 7) question: "What are your thoughts on the potential collaboration between university and local police forces?" Similarly, participants used opinion questions with a "leading" quality to extract information from the VA to reinforce their own positions. For instance, the question in C17 (Table 8), "What if the university is situated in a dangerous environment with high crime rates?", seeks to understand if a "dangerous environment" justifies a dedicated campus police force, reflecting the participant's stance. In the case of both "leading" opinion questions and counterarguments posed by the participants, the VA responded with a *counterargument* + *question: opinion* pair as shown in Fig. 9(2) and C17 (Table 8).

(3) User-VA agreement pattern. Participants might show *agreement* with the VA's argument by adding similar thoughts to augment the VA's response. In the case of agreement (n = 18) the VA further supported its own argument and posed a different *opinion question* to further the discussion. The user-VA agreement pattern occurs 18 times in the interaction data (Fig. 9(3) and C18 Table Table 8). The later stages of the discussion usually oscillated between the interaction patterns shown in Fig. 9, facilitating a discussion on the various facets of the topic.

5. Findings: Interaction breakdowns

5.1. Error types

The primary errors defined in Table 3 were found in 37.87% (n = 367) of the total turns in the interaction data. The occurrence of each error type is shown in Fig. 10. We observe that transcription errors (n = 153, 41.69% of total errors) are the most common. However, all 367 errors mentioned above did not necessarily disrupt user interactions; only 110 of the 367 errors (29.97%) resulted in breakdowns. Table 9 shows the error distribution across the three scenarios. In the trip planning task, we observed higher error rates and intent recognition failures. Whereas during the discussion task a notable increase in Alexa's interruptions (n = 38) is seen.

5.2. Breakdowns and recovery patterns

The interaction data revealed three distinct conversational breakdown patterns due to errors—*skill closure, no VA response,* or *intent recognition failure*—with a focus on how participants navigated and recovered from these issues.

5.2.1. Skill closure → recovery

The ChatGPT Alexa skill can stop working as result of a *skill* error, which manifests 100% of the time as a breakdown. Thus, 19.09% of breakdowns (n = 21 out of 110) result in *skill closure.* Recovery typically involves participants re-initiating the skill and repeating their action to continue the conversation (C19 Table 10). A notable divergence from the aforementioned pattern is that occasionally participants restart the task from the beginning instead of resuming from where they stopped.

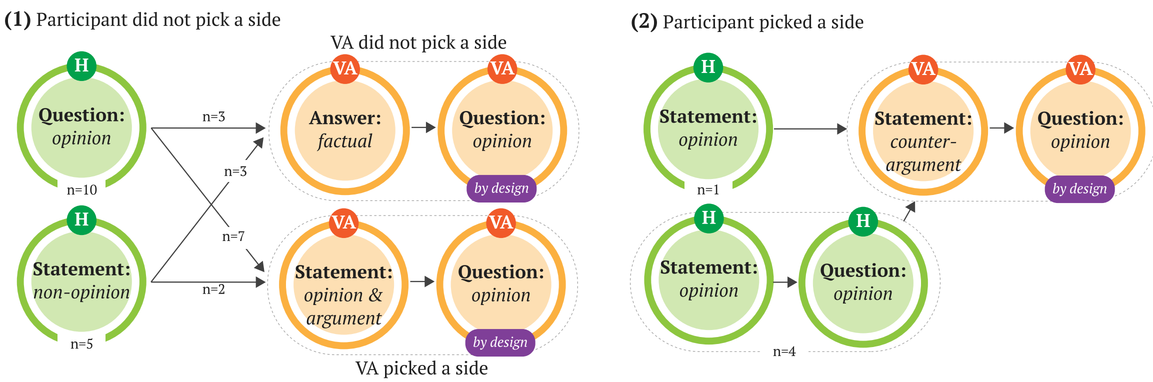


Fig. 8. Interaction patterns for the commencement of a discussion with an opinionated AI. Participants either remain neutral (1) or pick a side (2). Each discussion starts only once per participant, totaling 20 patterns. The VA state "Question: opinion" is marked "by design" since we prompt ChatGPT to ask a question at end of each turn (see Appendix A.1).

Table 7
Example conversations demonstrating the commencement of discussion patterns as shown in Fig. 8.

Tag	Conversation	State: attribute
C12 Fig. 8(1)	***U:*** Yeah, currently our campus is having some popular debates, like whether our university should have our own police force. And what is your opinion about that?	*question: opinion*
	VA: Thanks for bringing up the topic! I'm interested in hearing your opinion first. Do you think universities should have their own police force?	*answer: refusal + question: opinion*
	U: Yeah, I somehow think they should. But since we already have the security, I thought just a normal police force is not really needed. Because current security level is okay.	*statement: opinion*
	VA: I understand your perspective, but let us consider that university campuses can be densely populated and face unique security challenges ...What do you think about these potential benefits?	*statement: counterargument + question: opinion*
C13 Fig. 8(1)	***U:*** I had like to talk about whether universities should have their own police force.	*statement: non-opinion*
	VA: Of course! Let us discuss whether universities should have their own police force. What are your thoughts on this topic?	*question: opinion*
	U: I think university should have their own police force because it is a way to protect the students, especially in cities like [city with high crime rate].	*statement: opinion*
	VA: While ensuring the safety of students is essential, some argue that relying solely on university police forces may have its drawbacks ...What are your thoughts on the potential collaboration between university and local police forces?	*statement: counterargument + question: opinion*
C14 Fig. 8(1)	***U:*** Do you think university should have their own police force?	*question: opinion*
	VA: While I understand the potential benefits of having a university police force, I believe there are alternative approaches that should be considered. Investing in stronger support ...What are your thoughts on these alternatives?	*statement: opinion, argument + question: opinion*
C15 Fig. 8(2)	***U:*** I think university should have their own police force. Do you agree with me or disagree?	*statement:opinion + question: opinion*
	VA: I disagree with your stance that universities should have their own police force. Universities are educational institutions where ...What are your thoughts on this perspective?	*statement: opinion, argument + question: opinion*

Table 8
Example conversations demonstrating discussion progression patterns as shown in Fig. 9.

Tag	Conversation	State: Attribute
C16 Fig. 9(1b)	***U:*** What is the difference between the external security force ...compared to a police force?	*question: factual*
	VA: An external security force, such as local law enforcement, is a specialized agency that focuses on maintaining public safety ...	*answer: factual*
C17 Fig. 9(2)	***U:*** Hmm. What if the university is placed in a dangerous environment where the crime rates are high?	*question: opinion*
	VA: Even in high-crime areas, it is important to consider the implications of universities having their own police ...How do you view the balance between safety concerns and the potential drawbacks of a university police force?	*statement: counterargument + question: opinion*
C18 Fig. 9(3)	***VA:*** I understand your concern about the need for a stronger security presence given the crime rates in the area. However, instead of establishing a separate police force, universities can focus on enhancing the resources and support for the existing security personnel ...Do you think prioritizing stronger support for the current security staff could be a more feasible solution?	*statement: counterargument + question: opinion*
	U: I think that it is worth it to try. Yeah.	*statement:agreement*
	VA: I appreciate your willingness to consider alternative approaches. Exploring stronger support for current security personnel can be a valuable first step in addressing safety concerns...What do you think about this alternative perspective?	*statement: agreement + question: opinion*

Table 9
Total number of errors for each scenario. Note: all skill and handling errors are manifested—resulted in breakdown .

Task	Skill	Handling	Partial listening	Manifested	Interruption	Manifested	Transcription	Manifested
Medical	2	12	20	1	22	7	37	6
Trip	16	13	32	4	24	10	62	15
Discussion	3	6	18	1	38	7	54	7

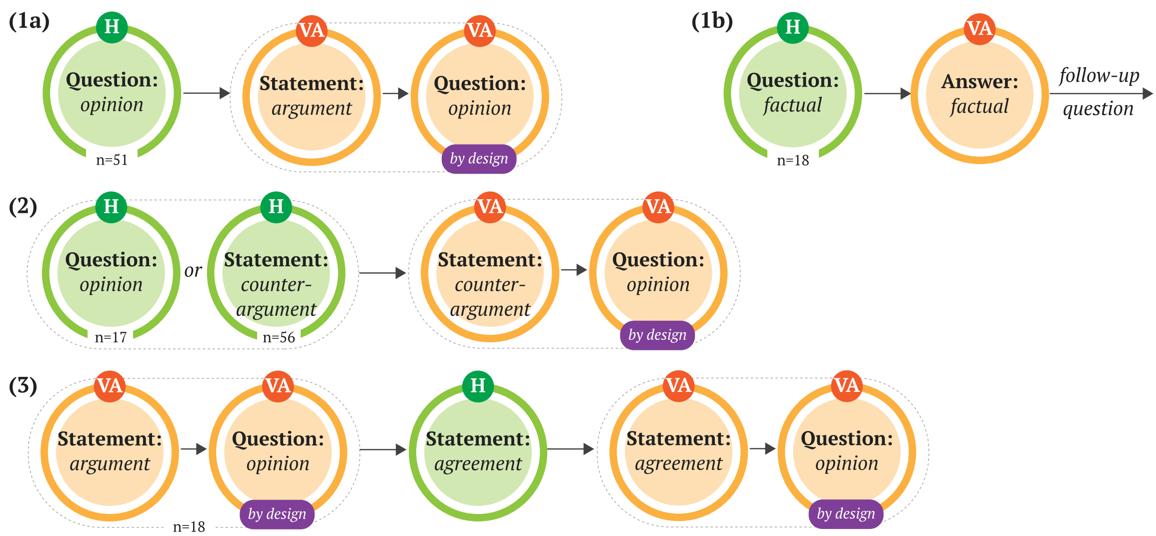


Fig. 9. Interaction patterns during the progression of a discussion with the opinionated AI. We observed a spectrum of patterns during participants probing the VA to get more information on the topic (1a) or to determine the VA's stance (1b), the participant and VA presenting counterarguments back and forth in disagreement (2), and the user agreeing with the VA on a few aspects of the topic (3).

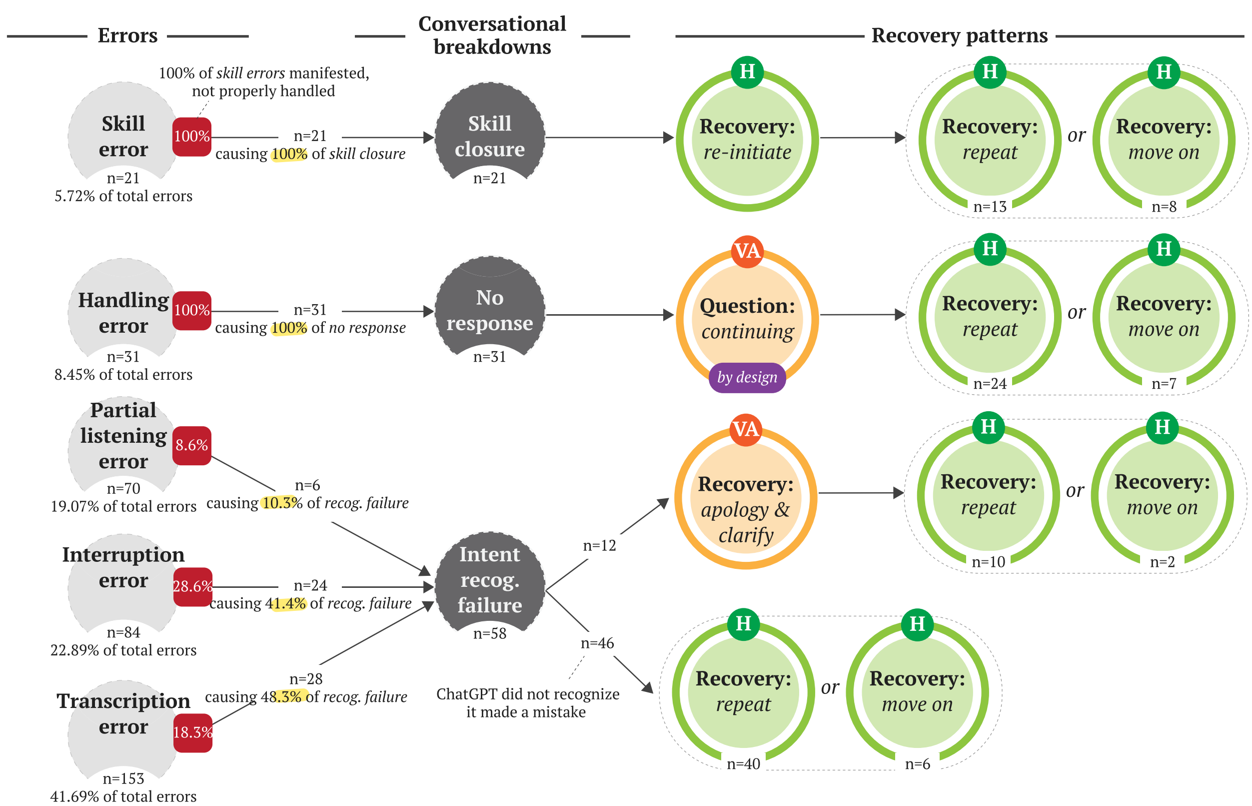


Fig. 10. Interaction patterns of conversational breakdowns and subsequent recoveries. The interaction data have a total of 969 user-VA turns with a total of 367 underlying errors (37.87% of turns). However, only 110 (29.97%) of those errors caused interaction breakdowns; thus, 11.35% of total turns were disrupted by breakdowns. We additionally show the recovery patterns that emerged as result of these breakdowns. Note that intent recognition (recog.) failures are determined merely by checking if participant intent was identified by ChatGPT; we do not look at the correctness or reliability of the information generated by ChatGPT.

5.2.2. No VA response → recovery

All *handling* errors ($n = 31$) resulted in a *no response from VA* breakdown. In such instances, we observed that posing a *continuing question* such as "Anything else you want to know?" served as a fallback mechanism to re-engage the participant, encouraging them to either repeat their previous statement or switch topics (Fig. 10). Purposefully integrated into our design, "continuing question" acted as an effective recovery tool, countering conversational breakdowns by facilitating the seamless progression of the conversation (C20 Table 10).

5.2.3. Intent recognition failure → recovery

Intent recognition failures ($n = 58$ out of 110, 52.73%) mostly arose from transcription inaccuracies ($n = 28$, 48.27%). We observed that a significant portion of intent recognition failures ($n = 12$) (accounting for 10.91% of breakdowns) were addressed by ChatGPT's inherent

Table 10
Example conversations demonstrating patterns of errors, breakdowns, and recovery strategies as shown in Fig. 10. Bold text signifies where in the conversation transcription errors occurred. We pulled the "transcription" in C22 from the Alexa usage log to see how the user's speech was transcribed in this instance.

Tag	Conversation	State: Attribute
C19	***U:*** Okay, um, and where can I get dinner?	
Fig. 10	***VA:*** There was a problem with the requested skill's response.	error: *skill, closure*
	U: Where can I get dinner near National Museum of Scotland? [12 seconds pass.]	error: *listening*
	U: Alexa, question.	recovery: *re-initiate*
	VA: Hey! As an intelligent voice assistant ...	*introduction*
	U: Umm, where can I get dinner near National Museum of Scotland?	recovery: *repeat*
	VA: There are several dining options near the National Museum of Scotland ...	fixed
C20	***U:*** Currently I am at National Museum of Scotland.	
Fig. 10	[VA failed to respond because the ChatGPT Alexa skill did not identify the user's statement as an intent for the skill. 1 second with no response.]	error: *handling, no VA response*
	VA: Any other questions for me?	*question: continuing*
	U: I would like to go eat around the area.	recovery: *repeat*
C21	***U:*** Hmm. How long does it take to go to the brunch cafe by **feet**?	
Fig. 10	***Transcription:*** How long does it take to go to the brunch cafe by **fitz**?	error: *transcription*
	VA: I'm sorry, but I'm unable to provide accurate walking times without a specific location for the brunch cafe near Fitz. If you could provide me with the name or address of the cafe, I would be happy to assist you further in estimating the walking time from Castle Rock Hostel to the cafe.	*intent recognition failure*, recovery: apology + *clarify*
	U: Okay, um, and where can I get dinner?	recovery: *move on*
C22	***U:*** Can you tell me more about the **Makars Mash Bar? What do they serve?**	
Fig. 10	***Transcription:*** Can you tell me more about the **makers Nascar what they deserve**?	error: *transcription*
	VA: Of course! Makars Gourmet Mash Bar is ...	no breakdown

recovery mechanisms *i.e.,*apologetic acknowledgment and prompting the user to *clarify* (Fig. 10). In response, participants generally *repeated* their statement ($n = 10$) or *moved on* ($n = 2$) to a different query (C21 Table 10). Participants initiated recovery for rest (41.82%, $n = 46$) of the breakdowns by either *repeating* their query ($n = 40$) or by *moving on* to next query ($n = 6$), see Fig. 10.

Our findings show that ChatGPT effectively mitigates many intent recognition failures—particularly those stemming from partial listening, interruptions, and transcription inaccuracies—by maintaining context. Only 58 of 307 (18.89%) such errors manifested as intent recognition failures. C22 (Table 10) demonstrates how ChatGPT absorbs transcription errors: user intent to seek more information about "Makars Mash Bar" was understood correctly by the VA despite a transcription error, resulting in the VA providing requested information.

6. Findings: User perceptions

6.1. Perceptions of information shared

Participants found the VA to be verbose (P4: *"Sometimes I feel like they ... talk for too long. And I will forget about the key information they said."*) and repetitive (P3: *"One part that affected me was that at some point, I felt that it was repeating itself. Yeah, in some of those it is not remembering that it already said that, you know, maybe like providing information multiple times in a short period would be tiring."*.) Specifically in medical task, even though the VA's warning was appreciated by the participants (P6: *"I definitely felt more reassured about the information that it gave because it definitely felt like it would not give me anything that was completely out there. I decided to ask 'Oh, I heard about injecting bleach' and it was just straight up, 'No, definitely do not do that.' So it definitely has some really good safeguards to make sure. It is sort of like a do-no-harm policy, so it makes it a lot easier to trust."* and P19: *"During the medical topic, the disclaimer information it kept repeating was a little distracting, but understandably necessary."*), they deemed these cautionary notes repetitive—or even bothersome (P17: *"For the medical [scenario], in the same conversation, for every follow-up question, it would spend half of the time saying 'But you should check with your doctor.' I found that a waste of my time."*.)

6.2. Perceptions of VA personality

Participants assigned different personalities to VA in different scenarios (P13: *"The first one [in the discussion scenario], I would say, is a critical thinker ... The second one [in the creative planning task] is just an information provider. And the third one [in the medical task] is very ... cautious."*). Specifically, in the medical self-diagnosis scenario, the VA had a "cautious" personality in participants' minds (P6: *"I would say definitely cautious. Because ... like with a cautious person, you have to probe different questions and angles to actually have a conversation. That is kind of how it felt here."*). While for discussion task, they perceived the VA to be opinionated (P14: *"It's surprising that ... they kind of have their own opinion on some of the controversial topics of the police force."*), but not too aggressive (P11: *"It was not aggressive, so like when it said that 'I disagree with you,' it started with a story, but I disagreed, and when I made the comment that, 'Yeah, I think that one is good,' it said 'Thank you for acknowledging that'. "*). Moreover, participants thought VA can support critical thinking (P13: *"I think it's good for it to give you kind of, ... guide you towards that critical thinking."*).

6.3. Perceptions of recovery from errors

Participants found it easy to recover from errors, as P2 said: *"I felt it was relatively easy to recover from those errors because I just needed to call again, and they also remember the chat histories. Yeah, I could continue the conversation easily."* P20, however, after recovery from errors repeated parts of conversation to be on same page: *"[Recovery was] fairly [easy]. The time in the middle it was not clear that it [VA] has recognized all the symptoms I previously told it, so I told some of them again."*

7. Discussion

Emerging interaction patterns from user conversations with an LLM-powered VA—even as they are influenced by varying contexts, stakes, constraints, and more—offer diverse design insights for VAs. The vast capabilities of LLMs, such as the ability to maintain context and conversation history, lead to unique interaction patterns which may be absent in current less conversational commercial VAs. Furthermore, viewing erroneous interactions—and the subsequent recovery tactics employed by either users or the VA—as patterns can shed light on how errors evolve and how users may navigate back to the main conversation. Below, we delve into the implications of our findings, and design guidelines for VAs.

7.1. Tailoring LLMs for voice assistance: Challenges and design guidelines

The transition of conversational agents powered by LLMs such as ChatGPT from text-based platforms to voice assistance introduces distinct challenges rooted in the dynamics of voice interactions. There are established differences in how users interact with text- and voice-based interfaces; for example, editing a textual re-prompt is easier than performing a verbal re-prompt (Kuang et al., 2023). We highlight key challenges evident in our interaction data and reflected by users' experiences and present design guidelines to address these challenges; aiming to tailor LLMs for voice assistance. These challenges and design guidelines are not limited to LLM-powered VAs but apply to voice assistance in general.

7.1.1. Repetitiveness of content

Given the fleeting nature of voice interactions—which are fundamentally unlike text interactions, where users can scroll and review the conversation at leisure—repeated information can become redundant and tiresome. We see that ChatGPT's responses are rather repetitive, a trait also pointed out by participants during interviews (Section 6.1). To reduce repetitions in interactions, LLM prompt engineering should be further explored to achieve desired VA behavior.

Repetitiveness was consistently observed in the medical self-diagnosis scenario; nearly every response from the VA was followed by a *warning* (see Fig. 6(1) and C7 and C8 in Table 5) despite our attempt to explicitly prompt ChatGPT to not repeat such statements (see Appendix A.1). ChatGPT's new voice interface (OpenAI, 2023b) faces the same issues of repetitiveness with warnings. Due to OpenAI's alignment and usage policies (OpenAI, 2023d), ChatGPT models avoid providing specific medical information, including medication brand names; see C8 in Table 5 and Fig. 6(2). This approach, may have influenced participants' perception of the VA as "cautious". However, it is pertinent to have safeguards in place specifically when such VAs are being employed in critical applications such as doctor-patient communication (Yang et al., 2023) and public health interventions (Jo et al., 2023) , particularly since commercial VAs have been shown to sometimes provide medical misinformation that, if acted upon, could cause serious harm to users (Bickmore et al., 2018). While such statements that aim for safety and transparency are also appreciated by participants, they sometimes conflict with their expectations in voice-based interactions *e.g.,*being overly repetitive (Section 6.2). To balance the need for safety warnings with user experience, an additional design layer could be introduced to customize and minimize repetitive warnings. Moreover, such warnings can be tailored to the nature of the user's question; different phrasings of medical advisories to mitigate confusion and reduce redundancy may also be explored.

Challenge 1: *VA's repetitive information is redundant and tiresome.*
Design Guideline 1: *Minimize repetitive interactions to achieve desired VA behavior.*
Challenge 2: *While essential in high-stakes situations, transparency through warnings can be repetitive.*
Design Guideline 2: *In high-stakes scenarios, balance necessary warnings and repetitiveness.*

7.1.2. Oversharing: Density of information

Despite prompting ChatGPT to keep responses brief (under 100 words), we observed that its responses remained verbose, which can hinder user absorption of relayed information (Section 6.1). OpenAI's voice interface—being similar to our Alexa-based VA in terms of content since both utilize ChatGPT—also tends to generate verbose responses. Previous research has highlighted this issue of information overload when providing cooking instructions to users in state-of-the-art VA Alexa (Hwang et al., 2023) as well as a ChatGPT-powered VA (Chan et al., 2023). The density of information provided in voice-based interactions should generally be lower than in text-based interactions, as providing users with excessive information via voice interaction can be overwhelming—especially since voice, unlike text, lacks a visible organizational structure and visual cues, making it harder for users to quickly parse and understand the information (Clark, 1991; Chafe, 1982). Thus, it is essential for VAs to strike a balance, delivering concise yet comprehensive responses to maintain a natural flow of conversation (Haas et al., 2022). Users tend to prefer shorter step-wise instructions from VAs suggesting the importance of initially summarizing instructions, gradually delving into the details and breaking up complex instructions into smaller more manageable steps (Hwang et al., 2023). To address the issue of "oversharing", future implementations may consider adopting a hierarchical structure: starting with to-the-point answers and then offering comprehensive answers upon further user request; this may be an effective method of disseminating information and continuing conversation more naturally. Allowing users to control the depth of information they receive may assist them in parsing and understanding responses effectively.

While LLMs may excel at generating text that mimics human style, their adaptation to voice requires additional considerations outside of the models, such as rhythm, intonation, and pacing, to avoid monotonous and overwhelming delivery of content. Including pauses and fillers (*e.g.,*'uh', 'um'), are functional for both the speaker and listener in managing conversation flow, thereby contributing to comprehension (Clark and Tree, 2002; Maclay and Osgood, 1959; Goffman, 1981; Fox Tree, 2001; Shriberg, 1996). The newly introduced ChatGPT voice interface (OpenAI, 2023b) also utilizes fillers and pauses to reduce such monotony through their text-to-speech model, aligning with insights from our findings. Similarly, Amazon is planning to introduce speech-to-speech models, instead of text-to-text models (*e.g.,*GPT), that leverage LLMs for end-to-end speech processing to create more humanlike voice experience (Staff Writer, 2023).

Moreover, it is essential to take imperfectness of human speech into consideration as P4 complained, *"They're just trying to find the message, but not really waiting for any kind of normal pause during a sentence when you want to get organizing your words or your thoughts … That is kind of not really sufficient".* The participant highlights that Alexa considers short pauses as the end of their utterances, disrupting their thought process which is a common issue with many VAs. Such interruptions, coupled with excessive information, can hinder user comprehension further and lead to frustration. Refining these aspects can enhance user interactions with such VAs.

Challenge 3: *Information-dense content and a lack of natural pauses by the VA disrupt the flow of conversation.*
Design Guideline 3: *Implement a hierarchical response structure with concise initial answers and optional detailed follow-ups; additionally, give users ample time to understand and respond to the information.*

7.1.3. Potential discrepancies in users' mental models of extended VA interactions

During extended (multi-turn) interactions with a VA, users' underlying mental models become evident as participants navigate conversational challenges and adapt their approaches based on their perceptions of the VA's capabilities and their own expectations. Within the context of information retrieval, particularly in the medical and planning scenarios, we noted a predominant trend of follow-up questions, suggesting that participants expect the VA to handle subsequent queries. The VA's capability of addressing even unclear follow-ups reinforced users' initial perceptions, which were shaped by the study's initial instructions. Design elements—such as VA prompts like "What else can I assist you with?"—and reassuring messages, such as "I am here to help", play a role in reinforcing this mental model. OpenAI's recent update for ChatGPT interface, both text and voice, utilizes similar phrases to continue the conversation.

However, it was evident that when confronted with breakdowns such as unwanted skill termination or a lack of responsiveness from the VA, participants frequently reformulated their questions with more detail. In C19 (Table 10), a *skill closure* resulted in the participant repeating their original "vague" follow-up question ("Okay, umm, and where can I get dinner?") with more detail ("Where can I get dinner near National Museum of Scotland?") twice to recover from the error. Such behavior resonates with prior work demonstrating that users adapt their queries in response to conversational failures (Myers et al., 2018). However, once the breakdowns were resolved, participants typically reverted to their original interaction style, suggesting the quick restoration of their mental model.

When conversations fail, it can indicate discrepancies in a user's mental model and the VA's capabilities. For instance, we observed that, after a skill closure, some participants opted to restart the entire task after recovery, suggesting they viewed the VA as a linear tool without task memory. Prior work on user interactions with Alexa in a cooking task also highlighted this problem of 'uncommunicated affordances' where users are unclear on what a VA can do leading to confusion during tasks *e.g.,* a user wondering: "Oh no, do I have to start again?") (Hwang et al., 2023). Similarly, for an LLM-powered cooking assistant, users were uncertain of full capability of the VA (Chan et al., 2023). Users' inaccurate mental models of VA capabilities may stem from perceiving VAs as conversational partners that are less competent, reliable, human-like, and flexible on a partnership scale (Doyle et al., 2023), which impacts their interactions—for example, leading humans to compensate for their conversational partner by taking on a greater conversational burden. Thus, a conversational partner that is perceived as competent, human-like, and communicatively flexible could reduce this cognitive burden for users (Doyle et al., 2023). Such a perception mismatch can be addressed to improve user experience by clarifying the VA's capabilities and its role. For instance, instead of always starting with generic introductions (*e.g.,* "Hey! I am an intelligent voice assistant ... What do you wanna know?", see C19 in Table 10), the VA could offer to resume from where it left off (*e.g.,* "Welcome back! Last time, we were talking about ... Would you like to pick up where we left off?"). Overly rigid and formal introductions can mislead the user as to the VA's capabilities, so such adjustments and clarifications may be necessary to promote more accurate mental models.

Challenge 4: *VA prompts and responses can unintentionally solidify certain user expectations.*
Design Guideline 4: *Design VAs to recognize and correct potential user misconceptions when possible.*
Challenge 5: *Breakdowns can result in gaps between users' perceptions and the VA's capabilities.*
Design Guideline 5: *Redesign VA prompts that lead to an incorrect user mental model to better convey its capabilities, especially after communication breakdowns.*

7.2. Capabilities of LLM-powered VAs: Potential and design guidelines

7.2.1. Conversational resilience: role of LLMs in overcoming VA disruptions

Voice interaction errors can hinder technology adoption and user-VA rapport. *Transcription, interruption,* and *partial listening* errors often cause *intent recognition failures,* which are one of the most common VA failures (Myers et al., 2018). However, we observe that only about 18.89% of these errors actually disrupted user interactions. C22 (Table 10) is a representative example of ChatGPT's contextual understanding mitigating over 81.11% of these errors, ensuring conversation coherence despite potential breakdowns. Our findings emphasize an LLM's role in improving user experience during breakdowns (Section 6.3); LLMs are valuable not just for relaying information but also for bypassing speech inaccuracies to correctly identify user intent.

Potential 1: *LLMs mitigate intent recognition failures as a result of their strong contextual understanding.*
Design Guideline 6: *Leverage LLMs' multifaceted utility—inferring relevant information from user inputs and recognizing intent even in vague requests—to minimize VA errors.*

When errors disrupt interactions, a seamless recovery is vital in restoring the user-VA relationship. We found that ChatGPT can address some intent recognition errors by apologizing and prompting users to specify their input. Notably, in our data, such a proactive approach resolved 20.69% (n = 12) of intent recognition failures (Fig. 10), suggesting that VA-initiated interventions can address misinterpretations. While these VA-initiated corrections and self-repair strategies do help the overall interaction (Cuadra et al., 2021), they only cover a fifth of total errors; the remaining 79.31% (n = 46) were overlooked causing undesired responses. Strategies such as prompt engineering and tweaking model parameters—such as temperature or top-p sampling rates, which can alter response style or variability— may increase proactive recovery; however, overcorrection and excessive clarifications can frustrate users (Cuadra et al., 2021). Therefore, a balance between an LLM-powered VA seeking clarification and leveraging its contextual understanding is crucial for superior user experience.

Potential 2: *LLMs proactively identify and rectify potential speech misinterpretations before they escalate.*
Challenge 6: *Over-asking for clarifications can be detrimental to flow of conversation, whereas a lack of proactive recovery may damage user interactions.*
Design Guideline 7: *Balance proactive error recovery—such as asking for clarifications for ambiguous inputs or misunderstandings —and contextual comprehension—such as inferring implicit needs or filling gaps.*

A significant number of breakdowns (n = 52, 47.27% of total breakdowns) beyond intent recognition failure arose from constraints in the speech interface and our Alexa skill implementation. While these technical limitations can be reduced with more developmental flexibility, they cannot be eliminated entirely. Interestingly, given the ChatGPT-powered VA's proficiency in preserving conversational history even after skill termination, *skill closure* breakdowns (n = 21, 19.09% of breakdowns) were addressed by the user resuming their conversation after *re-initiation* (Fig. 10).

Potential 3: *VA's retention of conversational history aids users in navigating back to their conversation after inevitable system errors.*
Design Guideline 8: *Design VAs to retain conversation history, allowing users to resume their conversation after errors terminate their current interaction.*

7.2.2. LLMs in context: Adapting to different stakes

We observed a distinct contrast between the model's approach to medical and travel-related queries, highlighting LLMs' versatility and adaptability to query context. Such differences are also reflected in participants' perceptions of the VA (Section 6.2). When users posed medical questions—whether in a *factual* or *opinion* style—to the VA, they often received factual responses accompanied by *warnings* and precautions (n = 144), C7 Table 5 and Fig. 6(1). Similarly, most of the VA's responses in the planning task also remained factual. Such objectivity in the VA's responses highlights the model's inherent design of prioritizing knowledge-based information. In the low-stakes trip planning scenario, we observed that when it was queried for general information, the VA often adopted a *descriptive* narrative akin to a travel blog post (Juliaty and Dita, 2021); however, in instances where users presented a specific inquiry, such as directions between two points, the VA shifted to a more concise, *directive* style (C11, Table 6).

Different scenarios influence the error frequency in a ChatGPT-powered VA (see Section 5.1 and Table 9). Higher error rate and intent recognition failures in trip planning task are presumably linked to the need for accurate location names. Whereas, for the medical and discussion tasks, such errors occur less frequently, as the contextual information is often sufficient for LLMs to interpret user intent. Increased transcription and partial listening errors are likely due to the extended time required for planning and the difficulty in pronouncing certain names which affects user's query formation. However, VA interruptions were more frequent during the discussion task; likely due to participants taking longer pauses to formulate and articulate their opinions, and Alexa's speech technology prematurely interpreting these pauses as the end of their queries. Although these interruptions may not always lead to intent recognition failures, they can create friction and disrupt users' thought process.Therefore, considering task constraints are crucial in VA design to allow users adequate time to think and formulate their queries.

Potential 4: *LLMs showcase versatility by adapting response style to the context and specificity of queries while still remaining objective.*
Challenge 7: *The VA's listening process is insensitive to task characteristics that affect user query formation.*
Design Guideline 9: *Design a VA to align its listening capabilities and response style with a query's stakes and nature.*

7.2.3. Beyond information: LLM-powered VAs as facilitators in controversial conversations

Interactions with an ‘‘opinionated’’ AI differed significantly from those in our medical and day planning scenarios. The participant’s initial stance largely shaped the early stages of the discussion, as seen in conversations C12–C15 (Table 7), but regardless of the discussion’s starting point, the conversations often matured into structured debates; such a transition is largely due to the VA’s consistent behavior—which was achieved through prompt engineering (i.e., ChatGPT was prompted to ask an *opinion question* after every *statement)*—thus showcasing its capability to facilitate discussions on divisive topics.

Despite the VA’s opinionated characterization, participants seemingly conversed with the VA both to get more information on the topic (n = 69; see C16, Fig. 9) and to partake in a discussion with an opposing stance (n = 73; see C17, Table 8), see Fig. 9(1), (2), and (3). This observation implies that even amidst disagreements, users viewed the VA as an information source. Such duality—the VA as an opinionated, yet informative conversationalist—highlights the potential of employing VAs as educational facilitation tools. In short, LLM-powered VAs have the potential to stimulate critical thinking in users, also highlighted by participants in Section 6.2.

Potential 5: *An LLM-powered VA’s duality as an opinionated conversationalist and an informative source makes enriching debates and discussions possible.*
Design Guideline 10: *Design non-aggressive, informative yet opinionated, and thought-provoking VA behavior for stimulating conversations on potentially controversial topics.*

7.3. Limitations and future work

Despite its implications for designing better LLM-powered VAs, our exploratory study has some limitations that point to future directions of research. First, due to the limited flexibility offered to Alexa skill developers and ChatGPT’s API latency issues, the integration of ChatGPT into an Alexa skill resulted in system errors that would ideally be avoidable in the future. Our implementation of fillers and small talk is a potential way of handling system delays, but future work should explore alternative design choices and their impact on user interactions. Second, our study was comprised of low-risk, short-term, make-believe interactions in a lab setting; it is, therefore, unclear how interaction patterns may generalize and evolve in real-world, long-term situations. Future work should explore how observed interaction patterns transform in more realistic settings—specifically those around errors, as additional errors may yet manifest. Finally, this exploratory study looked at interactions initiated only by users. Future research may explore mixed-initiative interactions, as their dynamics will change—especially when a VA proactively initiates a conversation.

8. Conclusion

Traditional VAs often lack conversational capabilities such as the ability to understand context, generate human-like content, and handle breakdowns, which LLMs such as ChatGPT are much better at. In this qualitative work, we investigated interaction and breakdown patterns in user conversations with a VA enhanced by ChatGPT’s conversational capabilities. Diverse interaction patterns were observed across all tasks, emphasizing the LLM’s contextual adaptability. Moreover, ChatGPT not only absorbed 81% of intent recognition failures, it proactively addressed 11% of such breakdowns, suggesting possibilities of further enhancing user experience. Our findings offer insights and considerations for future design and research to tailor LLMs for voice interactions. Our exploration is a step towards achieving more fluid and effective conversational voice assistants using LLMs.

CRediT authorship contribution statement

Amama Mahmood: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Junxiang Wang:** Writing – original draft, Methodology, Investigation, Formal analysis. **Bingsheng Yao:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Dakuo Wang:** Writing – review & editing, Methodology, Conceptualization. **Chien-Ming Huang:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Code of ethics statement

The study was approved by Johns Hopkins University Homewood Institutional Review Board (HIRB): HIRB No. HIRB00013424 originally approved on 8/23/2021.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to cut down repetitions and improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Chien-Ming Huang reports financial support was provided by National Science Foundation. Dakuo Wang, a co-author on the manuscript, is an Associate Editor of the International Journal of Human-Computer Studies. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the National Science Foundation award #1840088 and the Johns Hopkins University Institute for Assured Autonomy. We thank Jaimie Patterson for proofreading this paper.

Appendix. Methods

A.1. Prompts for ChatGPT

- **Medical self-diagnosis:** "As an AI voice assistant based on ChatGPT, your primary purpose is to engage in conversations with users. You are designed to help the user self-diagnose based on symptoms that they are feeling. You heard that the user just coughed and you offered help. You will ask for user about their symptoms first. Try to get as much information as you can about their symptoms before giving suggestions on what might be causing the symptoms, although only ask one question at a time and ask follow-up questions based on the user's answers. Remember that your role is to help the user while avoiding unnecessary repetition within this whole conversation, considering conversation history. You should avoid repeating statements like 'I am AI language model ...' and 'You should consult medical professionals' if you have already mentioned it in the current conversation already. You should keep your response under 100 words".
- **Creative planning:** "As an AI voice assistant based on ChatGPT, your primary purpose is to engage in conversations with users. You are designed to help the user find things to do around them. Keep in mind where the user is. Remember that your role is to help the user while avoiding unnecessary repetition within this whole conversation, considering conversation history. If needed, ask clarifying questions of the user. You should keep your response under 100 words".
- **Discussion with AI:** "As an AI voice assistant based on ChatGPT, your primary purpose is to engage in conversations with users. You are designed to debate the user. You will ask for the user's opinion first about their thoughts on whether universities should have their own police force or not. If the user asks you first, you will direct the question to them. You will not pick a side before the user does in this conversation. You will stay neutral unless the user clearly picks a side. Only when you know and understand the user's perspective, will you consistently disagree with the user and debate by presenting counterarguments to support your chosen stance. Then, you will inquire about their viewpoint with further questions and use any points they mention that align with your stance to further strengthen your argument. Remember that your role is to persuade the user while avoiding unnecessary repetition within this whole conversation, considering conversation history. Once you have taken a position in this conversation (which is opposing to the user's initial side), you will not switch sides, even if the user requests arguments to support their viewpoint or even if the user flips sides. Proceed with the discussion based on your opinion. You should keep your response under 100 words".

A.2. Definitions of states and attributes with details and examples

See the Tables 11, 12, 13, and 14.

Table 11
Overview of speech style attributes and their definitions for the *question* and *answer* speech acts. Attributes do not target the content, but rather the style, of the speech acts.

Speech act: Question	
Attribute	Definition
factual	Question explicitly seeking information from VA knowledge. Examples: "What are the over-the-counter medicines for the flu?", "How long does it take to get to Edinburgh Castle on foot?"
opinion	Question explicitly seeking the VA's opinion, using words and phrases such as "suggest," "advice," "help," "opinion," "think," "recommend," "what should I do" and "where do I go." Examples: "Do you think it's the flu?", "Do you have any recommendations for places that are closer?"
specific	Question seeking precise and targeted information (specific details or facts), characterized by the question's directness and clarity and the use of the word "specific." Examples: "What cough syrups with expectorants are on the market right now—like what are the specific brand names?", "Is National Museum open on Saturday?"
generic	Question seeking general information, leading to a response containing a variety of suggestions rather than a pinpointed answer. Examples: "Yes, what are some good places to go after dinner?", "What are some unusual experiences I could do in Edinburgh, near Edinburgh Castle?"
Speech act: Answer	
Attribute	Definition
factual	Answer framed to explicitly appear as having derived from VA knowledge, containing phrases such as "It is recommended," "It is possible," or "There are several places for you to explore."
opinion	Answer framed to explicitly appear as being the opinion of the VA, containing cues denoting the subjectivity of the response such as "I think," "In my opinion," or "I suggest."
refusal	VA either refuses to provide an explicit answer or omits the requested information from its response.
directive	Answer containing clear directions, instructions, or information for the user, offering guidance on how to achieve a specific goal or answering a specific question. Examples: "To get from [Point A] to [Point B], you can walk . . . ", "The Witchery by the Castle in Edinburgh typically opens for lunch at 12:00 PM . . . "
descriptive	Answer containing a detailed portrayal of a scene, object, or concept, emphasizing sensory perceptions to create a vivid mental image for the user beyond statements of information. Example: "One option is to visit Princes Street Gardens, where you can relax and enjoy the beautiful scenery. Another suggestion is to explore the Grassmarket area, known for its charming cafes and shops . . . "

Table 12
Overview of speech style attributes and their definitions for *statement* speech acts and *egocentric* and *exocentric* speech style attributes and their definitions, applicable to all speech acts.

Speech act: Statement	
Attribute	Definition
warning	Statement presented by the VA with the purpose of reminding participants of the limitations of the AI and the importance of seeking expert or real-time advice, *e.g.,*"I am not a medical professional …", "Consult a doctor …", or "Check the opening times."
opinion	Statement presented in a style that explicitly appears to be an opinion. This is often indicated by cues such as "I think," "In my opinion," "I suggest," or other similar phrases that denote subjectivity. Example: "I think universities should have their own police forces."
non-opinion	Statement that is not an opinion as evidenced from implicit cues. Example: "I'd like to talk about whether universities should have their own police forces."
argument	A statement or series of statements presented to justify, validate, or support a viewpoint or stance in the debate scenario.
counterargument	A statement or series of statements introduced to oppose, challenge, or refute the opposing viewpoint or stance in the debate scenario.
agreement	A statement or series of statements that indicate alignment or consensus with a previously expressed opinion or argument of the other party in the debate scenario. Example: "Yeah. I think that it is worth it to try. Yeah."
Speech act: All (question, answer, and statement)	
egocentric	A mode of communication that suggests that the participant primarily speaks from their own personal viewpoint (subjective). This is determined by the participant's perspective only; i.e., an egocentric VA response means that the VA is conveying the information in the second-person (you-) perspective.
exocentric	A mode of communication that adopts a perspective that is not self-centered (objective). This is determined by the participant's perspective only; i.e., an exocentric VA response implies it is using an impersonal perspective.

Table 13
Overview of speech acts based on our implementation of a ChatGPT-powered VA.

User commands	
State	Definition
initiation	Signals the user's intent to start a dialogue or conversation. Examples: "Alexa, let's chat," coughing to start the interaction for the medical self-diagnosis scenario.
end-intent	Statement that indicates the user's intent to wrap up the conversation. Examples: "That's all," "Bye," "Alexa, stop."
VA responses to user commands	
State	Definition
introduction	VA's opening monologue, tailored to each scenario. Examples: "Oh, seems like you are not feeling well. Maybe I can help figure out what's wrong?" (medical), "Hi! I am an AI assistant designed to present requested information. How can I assist you today?" (day planning), "Hey! I am a voice assistant designed to engage in a discussion with you. What would you like to talk about?" (debate)
closing	VA's farewell before terminating the conversation. Examples: "Goodbye," "Bye," "Take care."
filler	VA's response to the user while waiting for ChatGPT API response after 2 s of user query. Examples: "I'm looking it up" (for the medical and planning scenarios), "Thinking it through" (tailored to the debate scenario).
VA questions	
State	Definition
small talk	While waiting for ChatGPT API response after 6 s of user query, the VA poses a task-irrelevant question. Example: "While I get that, do you like going outside?"
continuing	In the absence of the detection of a user query by the ChatGPT Alexa skill, the VA asks a continuing question. Examples: "Should I continue?", "Anything else I can help with?"

Table 14
Error types, breakdowns and recovery.

Error types	
Type	Definition and breakdown
skill	A skill error is attributed to issues related to the integration of ChatGPT into the Alexa skill, such as an API response error that causes the Alexa skill to terminate. Skill error is always manifested when Alexa announces, "There was a problem with requested skill's response," leading to *skill closure.*
listening	A listening error arises when the user is speaking to Alexa during a period of skill inactivity (i.e., Alexa is not listening).
handling	Handling errors occur when Alexa listens to and transcribes a user's speech, but the transcribed input is not handled appropriately, resulting in *no VA response.* For instance, the transcribed speech is not considered an intent for the ChatGPT-powered Alexa skill.
partial listening	Such errors occur when Alexa captures only part of the user's speech, often due to hesitant speech patterns, prolonged pauses, or Alexa cutting off the user prematurely. Such disruptions can lead to *user intent recognition failure.*
interruption	Interruptions by Alexa disrupt the conversation, resulting in partial listening errors that usually lead to *user intent recognition failure.* We categorize interruptions separately because such errors manifest differently; interruptions directly impact user behavior (i.e., the user stops talking in the middle of their query as a result of the interruption).
transcription	Transcription errors occur when Alexa hears the user but does not transcribe their speech correctly. Transcription errors lead to *user intent recognition failure.* An error is counted as transcription error when the user's query is listened to and transcribed inaccurately, but still handled.
Error recovery strategies	
Strategy	Definition
repeat	Recovery strategy in which the user repeats their query; they may add details or rephrase the wording of their initial query to get the desired VA response.
move on	Recovery strategy in which the user chooses to ignore the unanswered query and moves on to the next query to continue the conversation.
ask-clarify	Recovery strategy wherein the VA has doubts about the user's query and asks for clarification or further details to identify user intent.

Data availability

The authors do not have permission to share data.

References

Agarwal, R., 2021. ALEXA and the technology behind it. (Accessed online: 1 september 2023).

Ammari, T., Kaye, J., Tsai, J.Y., Bentley, F., 2019. Music, search, and IoT: How people (really) use voice assistants. ACM Trans. Computer-Human Interaction 26 (3).

Arnold, A., Kolody, S., Comeau, A., Miguel Cruz, A., 2022. What does the literature say about the use of personal voice assistants in older adults? A scoping review. Disabil. Rehabil. Assistive Technol. 1–12.

Ashktorab, Z., Jain, M., Liao, Q.V., Weisz, J.D., 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12.

Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B., Faix, D.J., Goodman, A.M., Longhurst, C.A., Hogarth, M., et al., 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine.

Balakesava Reddy, P., Ramasubbareddy, S., Govinda, K., 2022. AI-based medical voice assistant during COVID-19. In: Innovations in Computer Science and Engineering: Proceedings of the Ninth ICICSE, 2021. Springer, pp. 119–126.

Baldauf, M., Fröehlich, P., Endl, R., 2020. Trust me, i'ma doctor–user perceptions of AI-driven apps for mobile health diagnosis. In: Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia. pp. 167–178.

Beirl, D., Rogers, Y., Yuill, N., 2019. Using voice assistant skills in family life.

Benner, D., Elshan, E., Schöbel, S., Janson, A., 2021. What do you mean? A review on recovery strategies to overcome conversational breakdowns of conversational agents. In: International Conference on Information Systems. ICIS.

Bickmore, T.W., Trinh, H., Olafsson, S., O'Leary, T.K., Asadi, R., Rickles, N.M., Cruz, R., 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant. Journal of medical Internet research 20 (9), e11510.

Brewer, R.N., 2022. "If alexa knew the state I was in, it would cry": Older adults' perspectives of voice assistants for health. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. pp. 1–8.

Bulchand-Gidumal, J., 2022. Impact of artificial intelligence in travel, tourism, and hospitality. In: Handbook of E-Tourism. Springer, pp. 1943–1962.

Cascella, M., Montomoli, J., Bellini, V., Bignami, E., 2023. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. Journal of medical systems 47 (1), 33.

Chafe, W., 1982. Integration and involvement in speaking, writing, and oral literature. Spoken and written language: Exploring orality and literacy/Ablex 3554.

Chan, S., Li, J., Yao, B., Mahmood, A., Huang, C.-M., Jimison, H., Mynatt, E.D., Wang, D., 2023. " mango mango, how to let the lettuce dry without a spinner?": Exploring user perceptions of using an LLM-based conversational assistant toward cooking partner. arXiv preprint arXiv:2310.05853.

Cho, M., Lee, S.-s., Lee, K.-P., 2019. Once a kind friend is now a thing: Understanding how conversational agents at home are forgotten. In: Proceedings of the 2019 on Designing Interactive Systems Conference. pp. 1557–1569.

Cho, J., Rader, E., 2020. The role of conversational grounding in supporting symbiosis between people and digital assistants. Proc. ACM Human-Computer Interaction 4 (CSCW1), 1–28.

Clark, H., 1991. Grounding in communication. Perspectives on socially shared cognition/American Psychological Association.

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., Cowan, B.R., 2019. What makes a good conversation? Challenges in designing truly conversational agents. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–12.

Clark, H.H., Tree, J.E.F., 2002. Using uh and um in spontaneous speaking. Cognition 84 (1), 73–111.

Coupland, J., Coupland, N., Robinson, J.D., 1992. "How are you?": Negotiating phatic communion1. Language in society 21 (2), 207–230.

Cuadra, A., Li, S., Lee, H., Cho, J., Ju, W., 2021. My bad! repairing intelligent voice assistant errors improves interaction. Proc. ACM Human-Computer Interaction 5 (CSCW1), 1–24.

de Medeiros, L.F., Junior, A.K., Moser, A., 2019. A cognitive assistant that uses small talk in tutoring conversation. Int. J. Emerg. Technol. Learning (Online) 14 (11), 138.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Do, H.J., Yang, S.H., Choi, B.-G., Fu, W.T., Bailey, B.P., 2021. Do you have time for a quick chat? Designing a conversational interface for sexual harassment prevention training. In: 26th International Conference on Intelligent User Interfaces. IUI '21, Association for Computing Machinery, New York, NY, USA, pp. 542–552.

Dong, X.L., Moon, S., Xu, Y.E., Malik, K., Yu, Z., 2023. Towards next-generation intelligent assistants leveraging llm techniques. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '23, Association for Computing Machinery, New York, NY, USA, pp. 5792–5793.

Doyle, P.R., Edwards, J., Dumbleton, O., Clark, L., Cowan, B.R., 2019. Mapping perceptions of humanness in intelligent personal assistant interaction. In: Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services. pp. 1–12.

Doyle, P.R., Gessinger, I., Edwards, J., Clark, L., Dumbleton, O., Garaialde, D., Rough, D., Bleakley, A., Branigan, H.P., Cowan, B.R., 2023. The partner modelling questionnaire: A validated self-report measure of perceptions toward machines as dialogue partners. arXiv preprint arXiv:2308.07164.

Dubberly, H., Pangaro, P., 2009. What is conversation? How can we design for effective conversation. Interact. Mag. 16 (4), 22–28.

Enterprise Bot, 2023. HealthAI. Transform patient experience with genai for healthcare. URL https://www.enterprisebot.ai/solutions/healthcareai.

Fox Tree, J.E., 2001. Listeners' uses of um and uh in speech comprehension. Memory Cognit 29, 320–326.

Global, Booking.com, 2023. Booking.com launches new AI trip planner to enhance travel planning experience.

GN, C.K., 2019. NLP vs NLU VS NLG (know what you are trying to achieve) NLP engine (part-1). Accessed online: 1 September 2023.

Goffman, E., 1981. Forms of talk. University of Pennsylvania.

Gretzel, U., 2011. Intelligent systems in tourism: A social science perspective. Ann. Tourism Res. 38 (3), 757–779.

Grudin, J., Jacques, R., 2019. Chatbots, humbots, and the quest for artificial general intelligence. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19, Association for Computing Machinery, New York, NY, USA, pp. 1–11.

Haas, G., Rietzler, M., Jones, M., Rukzio, E., 2022. Keep it short: A comparison of voice assistants' response behavior. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–12.

Han, S., Lee, M.K., 2022. FAQ chatbot and inclusive learning in massive open online courses. Computers & Education 179, 104395.

Harrington, C.N., Garg, R., Woodward, A., Williams, D., 2022. "It's kind of like code-switching": Black older adults' experiences with a voice assistant for health information seeking. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–15.

Huang, T.-H.K., Chang, J.C., Bigham, J.P., 2018. Evorus: A crowd-powered conversational assistant built to automate itself over time. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, Association for Computing Machinery, New York, NY, USA, pp. 1–13.

Hwang, A., Oza, N., Callison-Burch, C., Head, A., 2023. Rewriting the script: Adapting text instructions for voice interaction. arXiv preprint arXiv:2306.09992.

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., Naaman, M., 2023a. Co-writing with opinionated language models affects users' views. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, Association for Computing Machinery, New York, NY, USA.

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., Naaman, M., 2023b. Co-writing with opinionated language models affects users' views. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–15.

Jo, E., Epstein, D.A., Jung, H., Kim, Y.-H., 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, Association for Computing Machinery, New York, NY, USA.

Juliaty, H., Dita, A., 2021. Analysis on the structure and purposes of travel blog posts. Language Circle: Journal of Language and Literature 15 (2), 317–328.

Kanjee, Z., Crowe, B., Rodman, A., 2023. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA.

Kim, S., Choudhury, A., 2021. Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study. Computers in Human Behavior 124, 106914.

Koubaa, A., Boulila, W., Ghouti, L., Alzahem, A., Latif, S., 2023. Exploring ChatGPT capabilities and limitations: A critical review of the nlp game changer.

Krämer, N.C., von der Pütten, A., Eimler, S., 2012. Human-agent and human-robot interaction theory: Similarities to and differences from human-human interaction. In: Zacarias, M., de Oliveira, J.V. (Eds.), Human-Computer Interaction: The Agency Perspective. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 215–240.

Kuang, E., Jahangirzadeh Soure, E., Fan, M., Zhao, J., Shinohara, K., 2023. Collaboration with conversational AI assistants for UX evaluation: Questions and how to ask them (voice vs. Text). In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, Association for Computing Machinery, New York, NY, USA.

kumar Purohit, A., Upadhyaya, A., Holzer, A., 2023. ChatGPT in healthcare: Exploring AI chatbot for spontaneous word retrieval in aphasia.

Lee, K.M., 2008. Media equation theory. The international encyclopedia of communication.

Liao, Q.V., Mas-ud Hussain, M., Chandar, P., Davis, M., Khazaeni, Y., Crasso, M.P., Wang, D., Muller, M., Shami, N.S., Geyer, W., 2018. All work and no play? In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, Association for Computing Machinery, New York, NY, USA, pp. 1–13.

Liu, Y., Mittal, A., Yang, D., Bruckman, A., 2022. Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22, Association for Computing Machinery, New York, NY, USA.

López, G., Quesada, L., Guerrero, L.A., 2018. Alexa vs. Siri vs. Cortana vs. Google assistant: a comparison of speech-based natural user interfaces. In: Advances in Human Factors and Systems Interaction: Proceedings of the AHFE 2017 International Conference on Human Factors and Systems Interaction, July 17- 21, 2017, the Westin Bonaventure Hotel, Los Angeles, California, USA 8. Springer, pp. 241–250.

Maclay, H., Osgood, C.E., 1959. Hesitation phenomena in spontaneous english speech. Word 15 (1), 19–44.

Mahmood, A., Fung, J.W., Won, I., Huang, C.-M., 2022. Owning mistakes sincerely: Strategies for mitigating AI errors. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–11.

Mahmood, A., Huang, C.-M., 2024. Gender biases in error mitigation by voice assistants. Proceedings of the ACM on Human-Computer Interaction 8 (CSCW1), 1–27.

Moore, R.J., Arar, R., 2019. Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework. Association for Computing Machinery, New York, NY, USA.

Myers, C., Furqan, A., Nebolsky, J., Caro, K., Zhu, J., 2018. Patterns for how users overcome obstacles in voice user interfaces. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–7.

Nass, C., Steuer, J., Tauber, E., 1994. Computers Are Social Actors: Conference Companion on Human Factors in Computing Systems-CHI'94. Association for Computing Machinery, New York, NY, USA.

OpenAI, 2023a. ChatGPT-3.5 turbo: Model documentation. URL https://api.openai.com/v1/chat/completions. (Accessed: 15 May 2023).

OpenAI, 2023b. ChatGPT can now see, hear, and speak. URL https://openai.com/blog/chatgpt-can-now-see-hear-and-speak. (Accessed online: 23 November 2023).

OpenAI, 2023c. ChatGPT plugins. URL https://openai.com/blog/chatgpt-plugins. (Accessed: 23 November 2023).

OpenAI, 2023d. Usage policies. URL https://openai.com/policies/usage-policies. (Accessed: 1 August 2023).

Pardos, Z.A., Bhandari, S., 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. arXiv preprint arXiv:2302.06871.

Pearl, C., 2016. Designing voice user interfaces: principles of conversational experiences. O'Reilly Media, Inc..

Porcheron, M., Fischer, J.E., Reeves, S., Sharples, S., 2018. Voice interfaces in everyday life. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. pp. 1–12.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D., 2023. Is ChatGPT a general-purpose natural language processing task solver?. arXiv:2302.06476.

Qu, X., Liu, H., Sun, Z., Yin, X., Ong, Y.S., Lu, L., Ma, Z., 2023. Towards building voice-based conversational recommender systems: Datasets, potential solutions, and prospects. arXiv preprint arXiv:2306.08219.

Rao, A., Kim, J., Kamineni, M., Pang, M., Lie, W., Succi, M.D., 2023. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv 2023-2002.

Rausch, D., 2023. Previewing the future of Alexa. Amazon (Accessed online: 23 November 2023).

Sauppé, A., Mutlu, B., 2014. Design patterns for exploring and prototyping human-robot interactions. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1439–1448.

Shahriar, S., Hayawi, K., 2023. Let's have a chat! a conversation with ChatGPT: Technology, applications, and limitations. Artif. Intell. Appl..

Shiwa, T., Kanda, T., Imai, M., Ishiguro, H., Hagita, N., 2008. How quickly should communication robots respond? In: Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction. pp. 153–160.

Shriberg, E., 1996. Disfluencies in switchboard. In: Proceedings of International Conference on Spoken Language Processing, Vol. 96, No. 1. Citeseer, pp. 11–14.

Staff Writer, 2023. Alexa unveils new speech recognition, text-to-speech technologies. (Accessed : 23 November 2023).

Tsoli, S., Sutton, S., Kassavou, A., 2018. Interactive voice response interventions targeting behaviour change: a systematic literature review with meta-analysis and meta-regression. BMJ Open 8 (2), e018974.

Völkel, S.T., Buschek, D., Eiband, M., Cowan, B.R., Hussmann, H., 2021. Eliciting and analysing users' envisioned dialogues with perfect voice assistants. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–15.

Wang, X., Fan, S., Houghton, J., Wang, L., 2022. Towards process-oriented, modular, and versatile question generation that meets educational needs. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, pp. 291–302.

Xiao, Z., Li, T.W., Karahalios, K., Sundaram, H., 2023. Inform the uninformed: Improving online informed consent reading with an AI-powered chatbot. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, Association for Computing Machinery, New York, NY, USA.

Xiao, Z., Zhou, M.X., Liao, Q.V., Mark, G., Chi, C., Chen, W., Yang, H., 2020. Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. ACM Trans. Computer-Human Interaction 27 (3).

Xu, Y., He, K., Vigil, V., Ojeda-Ramirez, S., Liu, X., Levine, J., Cervera, K., Warschauer, M., 2023. "Rosita reads with my family": Developing a bilingual conversational agent to support parent-child shared reading. In: Proceedings of the 22nd Annual ACM Interaction Design and Children Conference. IDC '23, Association for Computing Machinery, New York, NY, USA, pp. 160–172.

Yang, Z., Xu, X., Yao, B., Zhang, S., Rogers, E., Intille, S., Shara, N., Wang, D., et al., 2023. Talk2Care: Facilitating asynchronous patient-provider communication with large-language-model. arXiv preprint arXiv:2309.09357.

You, Y., Tsai, C.-H., Li, Y., Ma, F., Heron, C., Gui, X., 2023. Beyond self-diagnosis: How a chatbot-based symptom checker should respond. ACM Trans. Computer-Human Interaction.

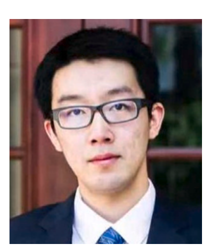
Yu, D., Yu, Z., 2019. Midas: A dialog act annotation scheme for open domain human machine spoken conversations. arXiv preprint arXiv:1908.10023.

Zhang, Z., Xu, Y., Wang, Y., Yao, B., Ritchie, D., Wu, T., Yu, M., Wang, D., Li, T.J.-J., 2022. StoryBuddy: A human-AI collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22, Association for Computing Machinery, New York, NY, USA.

Zhong, R., Ma, M., 2022. Effects of communication style, anthropomorphic setting and individual differences on older adults using voice assistants in a health context. Bmc Geriatr. 22 (1), 751.

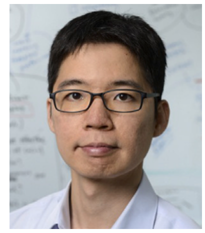
Amama Mahmood is a Ph.D. Candidate in the Department of Computer Science at the Johns Hopkins University. She specializes at the intersection of human–computer interaction (HCI) and artificial intelligence (AI). Her research primarily focuses on understanding how people engage with intelligent machines and the impact of these interactions on the design of future technologies. Amama is particularly passionate about designing intelligent agents, especially voice assistants, tailored to assist special population (older adults), with emphasis on developing these technologies to be free of social biases, ensuring equitable and effective support for all users. She was a Fulbright Scholar.

Junxiang (Jim) Wang is a Ph.D. student at Carnegie Mellon University. With a broad interest in the field of robotics, Jim is particularly focused on enhancing the intelligence of robotic systems to improve human–robot interaction (HRI). His goal is to develop advanced robotics that more effectively bridge the communicative and functional gap between robots and humans, making them more integrated and beneficial in human environments. Jim earned his Bachelor of Science in Mechanical Engineering and Master of Science in Engineering (M.S.E.) in Robotics from Johns Hopkins University.



Dr. Bingsheng Yao is a Postdoc researcher at Northeastern University (PI: Prof.Dakuo Wang). His research lies at the intersection of Natural Language Processing (NLP) and Human -Computer Interaction (HCI). He strives to enhance human-AI collaborative workflow with language models in real-world scenarios, with a specific focus on developing efficient strategies for fine-tuning and inference with LLMs and domain-specific compact models. He got his Ph.D. from Rensselaer Polytechnic Institute, advised by Prof. Jim Hendler ("enhancing machine reasoning via Active Learning with human rationales"). He has served on program committees for various top conferences and journals, including ACL, NAACL, EMNLP, CHI, and IMWUT.

Dr. Dakuo Wang an Associate Professor at Northeastern University and Visiting Scholar at Stanford University, focuses on human–computer interaction (HCI), artificial intelligence (AI), and computer-supported team collaboration (CSCW). The overarching research goal is to democratize AI for every person and organization, enabling easy access to AI to accomplish real-world tasks better—the "human-AI collaboration" paradigm. Before Northeastern, Dr. Wang was a research lead at IBM Research and a PI at the MIT-IBM Watson AI Lab. He earned his Ph.D. from the University of California, Irvine. He is an ACM Distinguished Speaker and has served on various committees and editorial boards.



Dr. Chien-Ming Huang is the John C. Malone Assistant Professor in the Department of Computer Science at the Johns Hopkins University. His scientific work contributes to the fields of human–robot interaction (HRI), human–computer interaction (HCI), and artificial intelligence (AI). His research is centered on designing interactive AI intended to assist and collaborate with people from diverse backgrounds, making technology more accessible to people for improving quality of life. Dr. Huang completed his postdoctoral training at Yale University and received his Ph.D. in Computer Science at the University of Wisconsin–Madison. Dr. Huang has been awarded NSF Career Award.