# Situation Awareness of Conversational Assistants in the Age of LLMs

Shih-Hong Huang[1,*], Chieh-Yang Huang[2], Hua Shen[3], Yuxin Deng[4] and Ting-Hao 'Kenneth' Huang[1]

[1]*The Pennsylvania State University, University Park, PA 16802*
[2]*MetaMetrics Inc., Durham, NC 27701*
[3]*University of Washington, Seattle, WA 98195*
[4]*Carnegie Mellon University, Pittsburgh, PA 15213*

**Abstract**

Large language models (LLMs) like ChatGPT enable near-human interaction, yet meaningful, lengthy dialogues need more than just delivering information. This paper argues that future conversational assistants should be aware of users' situations and alternate the conversation format based on the real-world situations. Through a Patrol Study, we demonstrate that users modify their communication approaches depending on their situations. Participants engaging in information-seeking conversations via WhatsApp while patrolling a building preferred voice messaging over text. This paper lays the groundwork for situation awareness in conversational assistants. The enhanced AI capabilities of LLMs make addressing HCI challenges essential to enable human-like, meaningful conversations beyond just providing information and generating fluent responses.

**Keywords**

Conversational Assistant, Situation Awareness, Large Language Models

## 1. Introduction and Background

With tools powered by large language models (LLMs) becoming increasingly accessible, more users are turning to them for assistance with various tasks. One of the most common ways users interact with LLMs is through agents or assistants that understand natural language. These agents allow users to either outsource tasks entirely or request partial assistance. However, most existing LLM assistants rely heavily on text interaction and require users to type their requests. Such reliance on text-based interactions often assumes that users are fully focused on a single task, seated at a computer, or have easy access to the necessary resources. Popular systems like ChatGPT, Claude, Gemini, DeepSeek, and Grok allow users to navigate their information needs, but still primarily operate in a text-based format as mentioned.

Users can covnerse with advanced LLMs almost as if with another human. However, many challenges emerge when incorporating LLMs into deployed conversational systems. A study showed that over 30% of conversations were erroneous, and nearly 30% of those erroneous conversations resulted in breakdowns when users tried to talk to GPT-3.5 via an Echo device [1]. Holding human-like, lengthy conversations requires more than just delivering information and producing fluent responses. In this paper, we argue that future conversational assistants, designed to assist users through text or voice in a turn-taking fashion, should learn to **be aware of users' situations and alternate the conversation format based on the situation**. We define "format" as the general attributes of a conversation– such as input modality and conversation length– separate from the primary content the conversation aims to convey.[1] For example, if the assistant detects that a user is walking outdoors, it should deliver shorter sentences, speak louder, and expect voice input rather than text messages from the user. When the

*AutomationXP25: Hybrid Automation Experiences, April 27, 2025, Yokohama, Japan. In conjunction with ACM CHI'25*
*Corresponding author.


[1]We deliberately chose to use "situation" instead of "context" in our work. This is because "context" often includes dialogue content, like previous conversation history, in dialog systems and natural language processing literature, which is not our focus in this work.

assistants know the user is near their computer, the system might send an email or a Slack message rather than a notification via their Echo device. Furthermore, if the topic of conversation is perceived as less urgent or unengaging, the assistant should try to keep the exchange brief. We call this capability **situation awareness** in conversational agents. This capability differs from the personalization in recommender or dialogue systems: In the cases of personalization, the same query from different users with varied contexts (such as search history, preferences, location, and age) yields different results [2]. Our focus is not on personalizing the system's responses. Instead, we emphasize that once a system has formulated a response, it should determine the best way to deliver it based on the user's situation. Although changing the conversation format can sometimes modify the content, it does not typically alter the core message and can be handled through simple paraphrasing or minor adjustments.

In the field of conversational assistants, much research has been devoted to understanding the broader context in which users interact with these systems, including their use in everyday household scenarios [3, 4]. Within Ubiquitous Computing, a significant body of work has been done on "context-aware computing" [5]. This research often revolves around enhancing conversational assistants with the ability to sense user context [6, 7]. For example, using ambient acoustic data to detect human activities like typing or walking [8] and utilizing respiration sensor to assist users in managing breathing patterns [9]. Likewise, in Affective Computing, efforts have been made to adjust conversational strategies based on detected user emotions– in which the "context" is the emotion– whose goals were often to resonate with users' emotions and thus enhance user engagement [10, 11, 12, 13]. However, despite intriguing research that took user context into account, the majority of dialogue systems research, including recent LLM-based chatbots, remained largely disconnected from such considerations [14, 15, 16].should we put some dialogue system research and gpt papers here sure In contrast, the natural language processing (NLP) and dialogue system communities focus primarily on producing accurate responses without considering the user's specific situation; most benchmark datasets for dialogue systems lack user context information [17, 18, 19]. The phrase "context-aware" conversational systems in NLP literature often refers to those that consider the domain or user's chat history, emphasizing the content of the conversation over the user's situation [20, 21, 22, 23]. Consequently, the systems considering user context have typically been developed only as ad-hoc projects with specific sensing capabilities, such as emotion-detection or environmental-sensing features [24, 25].

In this paper, we respond to this gap by advocating **a separation of conversation format from content, with situational information mainly influencing format**. This separation offers two-fold advantages: First, it ensures that effective existing dialogue systems like ChatGPT maintain their focus on content. This way, they can continue enhancing their capabilities using current datasets, model frameworks, and infrastructure while potentially benefiting from additional situation awareness. Second, it provides a straightforward path for researchers focused on user context to leverage the advancements of LLM-powered chatbots in their studies. With the impressive capabilities of modern LLMs, tailoring a response to suit a user's situation is now more attainable than ever.

To emphasize our unique approach, we have chosen to use "situation" instead of more commonly used terms like "context" or "scenarios."

We establishes our argument through a study focusing on a specific attribute of conversation format, to highlight the necessity for conversational assistants to adapt their formats according to user situations. Namely, **users modify their communication approaches based on their situations**, so situation awareness in assistants enhances their communicative reciprocity. The Patrol Study, validates this by involving participants in a walking task where they patrolled a building's interior while concurrently engaging in information-seeking conversations with a remote human helper via WhatsApp, using either voice or text messages. Despite the predominant use of text messaging in WhatsApp (over 90%) [26], the majority, when placed in this situation, preferred voice messaging.

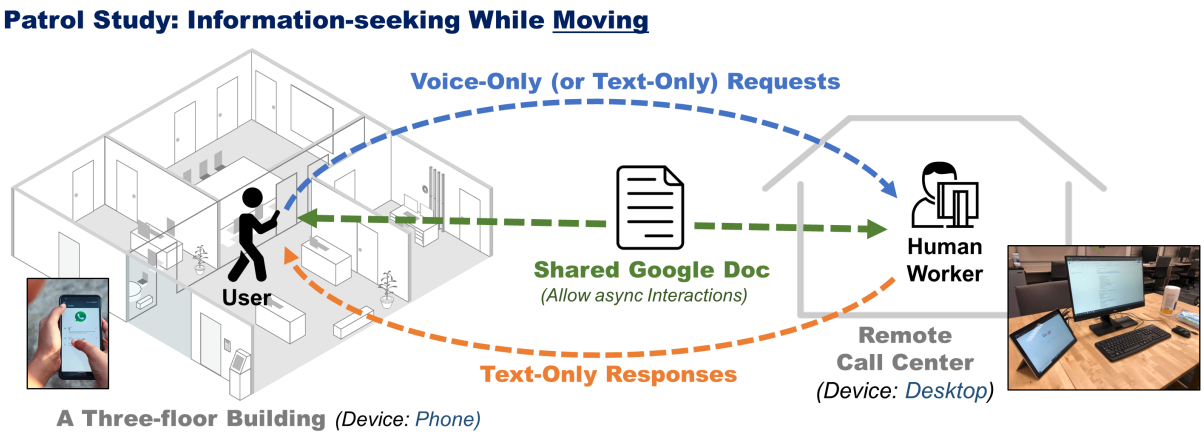


**Figure 1:** Patrol study overview. While performing a room-checking task, the user was required to work on an information-seeking task with a remote helper. The user could reach out to the remote helper through text or voice. All the information was organized on a shared Google Doc.

## 2. Patrol Study: Effects of Using Voice Interfaces to Receive Remote Help

The Patrol Study focused on one attribute of conversation format: the user's **input modality**. The objective of this study is to substantiate the argument that **users adapt their communication behavior according to their situations**, in particular, altering their preferred input modality in specific situations. WhatsApp Messenger, a widely-used instant messaging application, was utilized as the platform for this study. Notably, over 90% of messages were conveyed in text form, with a mere 7 out of every 100 billion messages being voice communications [26]. We hypothesize that users will deviate from their default preferences under specific circumstances.

**Study Design.** Figure 1 shows the overview of the Patrol Study. The two main components of the study were *(i)* **information-seeking** and *(ii)* **room-checking**. The information-seeking task required users to interact with a remote helper to ask for help on certain tasks and questions. Room-checking required users to walk around a three-story university building looking for certain rooms or research labs to check the availability of the rooms. We used the room-checking task to create a realistic, daily scenario– like conversing while navigating a building– and to potentially make participants prefer voice over text. For each session, we asked participants to simultaneously reach out to a remote helper for help on the information-seeking task while they were doing the room-checking task.[2] A shared Google Doc was updated by the remote helper and checked by the user in order to give further requests or ask clarifying questions. Each session took a total of 25 minutes. Users spent the first 15 minutes performing the room-checking and information-seeking tasks simultaneously. One research team member, dubbed the in-person helper, accompanied the users while they walked around the building to provide help if necessary, waiting at the end of the hallway. The in-person helper kept track of the time and told users when the 15 minutes were up. The in-person helper provided only logistics-related help outside the scope of the study, such as unexpectedly locked entrances. After the 15-minute patrol, the participant returned to the research lab and prepared for a short oral presentation on the information they obtained with the help of the remote helper. Users then had up to seven minutes for preparation and three minutes to present their findings. The details of information-seeking (Appendix A.2) and room-checking (Appendix A.1) tasks can be found in the Appendix.

[2]We intentionally informed participants that they were conversing with a human rather than conducting a Wizard-of-Oz study. This decision stemmed from our pilot studies, which indicated that, in such open-ended conversation settings, concealing the fact that participants were interacting with a human proved to be notably challenging.

**Study Procedure.** The study consisted of five sessions: a pre-study session, two interactive sessions, one multitasking session, and one evaluation session (see Figure 3 in Appendix A).

1. The **pre-study session** introduced the study and included a short tutorial.
2. **Two Interactive Sessions (Text Condition / Voice Condition):** At each interactive session, users were asked to perform the information-seeking and room-checking tasks and to reach out to the remote helper via WhatsApp to help solve the information-seeking task. They used texting in one session and voice messaging in the other. They were able to check the progress of the remote helper through a shared Google Doc. Additional questions regarding the remote helper's progress were communicated through WhatsApp. At the end of each interactive session, users were asked to verbally summarize what they learned from the information-seeking task.
3. **Multitasking Session (No-Help Condition):** The multitasking session required users to perform both the information-seeking and room-checking tasks without the help of the remote helper and to verbally summarize the information they gathered.
4. Users filled out questionnaires about the study during the **evaluation session**.

The order of the interactive session with text input (Text Condition) and interactive session with voice input (Voice Condition) was randomized for each participant. Conversation logs between the user and the remote helper were collected for both interactive sessions. Verbal summarization given by users during each session was recorded in audio form for further analysis. It took each participant 1.5-2 hours to complete the entire study, and participants were compensated with $30.00. This study was approved by the IRB office of the institute of the authors.

**Participants.** The participants for the Patrol Study were recruited through personal networks and university mailing lists. A total of 16 individuals were recruited as users of the study (participants were ID coded P1-P16 in the following paper): seven males and nine females. Fourteen of the participants were between 18 and 35; two participants were above the age of 36. The majority of the participants were undergraduate and graduate students at the university. Among all participants, only two had no prior experience using virtual assistants of any kind (*e.g.*, Siri, Google Assistant, or Alexa). Participants were informed that they would be interacting with a human assistant during the study instead of an automated agent. Participants were not aware of whether our study focused on voice or text assistance.

## 3. Results

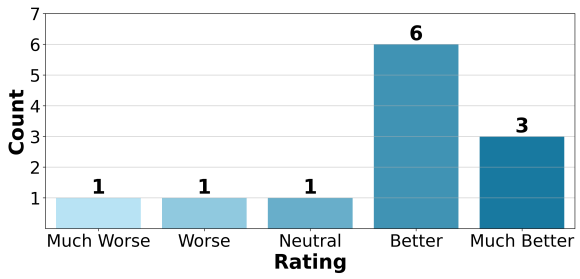


Figure 2: User's experience of using voice compared to using text in the context of the study (N=12). Participants preferred the voice interface over the text interface.

Table 1: User's satisfaction level on text and voice. Participants were satisfied with both of the conditions.

| ↑, N=12 | User's Satisfaction | | T-Test |
|---|---|---|---|
| | Mean | 95% CI | Voice |
| **Test** | 4.083 | [3.395, 4.772] | 0.389 |
| **Voice** | 4.167 | [3.458, 4.875] | - |

**Table 2**
Length statistic for conversation data (N=12). Participants spoke more words when using the voice interface.

| N=12 | Text | | Voice | | T-test |
|---|---|---|---|---|---|
| | Mean Rating | 95% CI | Mean Rating | 95% CI | |
| **#Sentences/Session** | **6.833** | [4.903, 8.764] | 5.833 | [2.959, 8.708] | 0.123 |
| **#Words/Sentence** | 7.379 | [6.451, 8.307] | **10.728** | [9.162, 12.293] | $<0.001^{***}$ |
| **#Words/Session** | 45.333 | [33.174, 57.493] | **62.250** | [25.426, 99.074] | 0.136 |

### 3.1. The voice interface was preferred over the text interface.

In the post-study survey, we asked participants to rate how satisfied they were when using text and voice on a five-point Likert scale from Very Dissatisfied (1) to Very Satisfied (5). Results shown in Table 1 indicate that the participants were satisfied with both conditions. We then asked participants to directly compare using text and voice to interact with the assistant in the context of the study on a five-point Likert scale from Much Worse (1) to Much Better (5). The average score was 3.857. Figure 2 shows the histogram of the responses and indicates that although users were satisfied with both communication interfaces, they preferred the voice interface over the text interface. These results validate our hypothesis: **Despite WhatsApp's predominant use of text messaging, participants, when placed in certain specific situations, diverge from their default behaviors to favor voice messaging.**

We asked participants to elaborate on their ratings for the comparison between text and voice interaction in the survey. P6 commented, *"I needed to take care of typos to deliver my message clearly, which is time-consuming. Also, I felt difficult to text in walking, while not hard to send a voice message in walking."* Others shared similar experiences when comparing using text or voice to communicate with the remote helper. P7 said, *"It's hard to type and walk at the same time! I prefer voice because I can talk and walk much easier."* P9 also said, *"When texting, I needed to concentrate on what I typed, when using voice, it was easier to speak the query I wanted and needed less effort."*

### 3.2. The participants tended to speak more when using voice.

To understand user behavior when using different communication interfaces, we calculated the number of sentences and words used in the conversation. Table 2 shows conversations statistics. We found that when using voice, participants tended to say more words in each sentence (number of words per sentence: 7.379 for text vs. 10.728 for voice) but use slightly fewer sentences in each conversation (number of sentences per session: 6.833 for text vs. 5.833 for voice). Overall, participants said more words in each session (number of words per session: 45.333 for text vs. 62.250 for voice). We include a few conversations in Appendix B. These findings suggest that **users modify their behavior in different situations**.

### 3.3. Positive preferences about using voice for other similar situations.

We also asked, "Think about day-to-day scenarios like running errands and walking between buildings, but you need to complete some other tasks at the same time. How likely would you choose to use voice instead of text to interact with a remote assistant to seek for help?" with a five-point Likert scale from Very Unlikely (1) to Very Likely (5). The average score was 4.357, suggesting that most of the participants felt positive about using voice in other situations. Participants in general agreed that the voice interface is easier to use when doing other tasks. For example, P8 said, *"Because voice is easier to communicate and lesser efforts over text;"* P7 reported, *"When using voice I can concentrate on my surroundings better. Using text requires me to look at the screen and type;"* P9 said, *"Using a voice assistant reduces the amount of work you need to do in terms of typing;"* and P12 commented, *"It is much easier*

*to use voice than text while performing other tasks."* However, participants also pointed out that the voice interface may be inappropriate in some situations. For example, P4 said, *"Voice may be difficult to use in public;"* and P3 said *"I think it would still depend on how comfortable I am speaking given my surroundings."* P15 adopted a neutral stance: *"In my opinion, texting and using voice is identical. It doesn't have much difference between them. Because, with current technology, the voice technology is not that good enough with accuracy. So it might be the same as well."*

# 4. Discussion

This paper introduces a study to advocate integrating situation awareness in conversational assistants. We demonstrate that users adjust their communication methods according to their situations, validating that assistants can communicate more effectively by recognizing and adapting to these situations.

The concept of "situation" warrants clarification. While "situation" and "context" can have varied meanings across different domains, like ubiquitous computing, affective computing, and NLP/dialogue systems, this work intentionally separates the content and format of conversation. We focus on adjusting the conversation format using factors external to content, countering the mainstream dialogue system research's emphasis on content, and underscoring the importance of meta aspects. Enabling extensive, human-like conversations involves more than merely disseminating information and crafting fluent responses.

## 4.1. Limitations

**Generalizability.** Although we attempt to create a scenario that emulates real life, there are still intricacies that are difficult to replicate. For example, the level of urgency while the task is being completed. Another significant factor is personal preference, as some users are just inclined to use one modality over others, which was reflected in the post-study questionnaire. While we chose questions that are more complex than typical voice commands, they were still generic and did not relate to the users on a personal level. We are aware that people care about highly social and personal questions [27], which can lead to reduced engagement with the questions we provided. Furthermore, privacy is a major concern if we are to ask really personal questions.

**Limitations of Human Ratings.** In our study, we had MTurk workers and Toloka workers rate the quality of transcriptions of the conversations as opposed to the audio. This setup did not capture the extra contextual information passed to remote helpers via voice. Furthermore, the third-party ratings of a conversation only reflected the perceived quality rather than the speaker's experience. We attempted to mitigate this gap by averaging the ratings collected from ten distinct workers, but it is still possible that the owner of the message disagrees with aggregated social perception.

# 5. Conclusion

This paper introduces a study that advocates for the situation awareness of conversational assistants. The Patrol Study illustrates that users adapt their communication strategies based on their respective situations, suggesting that assistants, aware of these situations, could better fulfill user needs. We argue for developing future conversational assistants that can recognize user situations and accordingly adjust conversational formats. Looking ahead, we aim to explore the potential for conversational assistants to automatically detect situations and make strategic communication decisions. Just as we do not explicitly instruct our friends about conversational preferences, we should not need to configure conversational assistants during each interaction. Leveraging the enhanced AI capabilities of LLMs to address HCI challenges is crucial for facilitating human-like, meaningful conversations that transcend mere information provision and fluent response generation.

## Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check.

## References


[1] A. Mahmood, J. Wang, B. Yao, D. Wang, C.-M. Huang, Llm-powered conversational voice assistants: Interaction patterns, opportunities, challenges, and design guidelines, 2023. arXiv:2309.13879.

[2] Y. Sun, Y. Zhang, Conversational recommender system, in: The 41st international acm sigir conference on research & development in information retrieval, 2018, pp. 235–244.

[3] M. Porcheron, J. E. Fischer, S. Reeves, S. Sharples, Voice interfaces in everyday life, in: proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–12.

[4] A. Sciuto, A. Saini, J. Forlizzi, J. I. Hong, " hey alexa, what's up?" a mixed-methods studies of in-home conversational agent usage, in: Proceedings of the 2018 designing interactive systems conference, 2018, pp. 857–868.

[5] A. K. Dey, Context-aware computing, in: Ubiquitous computing fundamentals, Chapman and Hall/CRC, 2018, pp. 335–366.

[6] U. G. Acer, M. v. d. Broeck, C. Min, M. Dasari, F. Kawsar, The city as a personal assistant: turning urban landmarks into conversational agents for serving hyper local information, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6 (2022) 1–31.

[7] S. W. Chan, S. Sapkota, R. Mathews, H. Zhang, S. Nanayakkara, Prompto: Investigating receptivity to prompts based on cognitive load from memory training conversational agent, Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 4 (2020) 1–23.

[8] C. Park, C. Min, S. Bhattacharya, F. Kawsar, Augmenting conversational agents with ambient acoustic contexts, in: 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services, 2020, pp. 1–9.

[9] A. Shamekhi, T. Bickmore, Breathe deep: A breath-sensitive interactive meditation coach, in: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare, 2018, pp. 108–117.

[10] A. Ghandeharioun, D. McDuff, M. Czerwinski, K. Rowan, Emma: An emotion-aware wellbeing chatbot, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2019, pp. 1–7.

[11] J. Casas, T. Spring, K. Daher, E. Mugellini, O. A. Khaled, P. Cudré-Mauroux, Enhancing conversational agents with empathic abilities, in: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, 2021, pp. 41–47.

[12] S. Samrose, K. Anbarasu, A. Joshi, T. Mishra, Mitigating boredom using an empathetic conversational agent, in: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, 2020, pp. 1–8.

[13] X. Yang, M. Aurisicchio, W. Baxter, Understanding affective experiences with conversational agents, in: proceedings of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–12.

[14] D. Ham, J.-G. Lee, Y. Jang, K.-E. Kim, End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 583–592. URL: https://aclanthology.org/2020.acl-main.54. doi:10.18653/v1/2020.acl-main.54.

[15] OpenAI, Gpt-4 technical report, ArXiv abs/2303.08774 (2023). URL: https://arxiv.org/abs/2303.08774.

[16] OpenAI, Introducing chatgpt, ???? URL: https://openai.com/blog/chatgpt.

[17] J. Zhou, J. Deng, F. Mi, Y. Li, Y. Wang, M. Huang, X. Jiang, Q. Liu, H. Meng, Towards identifying social bias in dialog systems: Framework, dataset, and benchmark, in: Findings of the Association for

Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 3576–3591. URL: https://aclanthology.org/2022.findings-emnlp.262. doi:10.18653/v1/2022.findings-emnlp.262.

[18] N. Dziri, H. Rashkin, T. Linzen, D. Reitter, Evaluating attribution in dialogue systems: The BEGIN benchmark, Transactions of the Association for Computational Linguistics 10 (2022) 1066–1083. URL: https://aclanthology.org/2022.tacl-1.62. doi:10.1162/tacl_a_00506.

[19] R. Lowe, N. Pow, I. Serban, J. Pineau, The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems, in: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, Prague, Czech Republic, 2015, pp. 285–294. URL: https://aclanthology.org/W15-4640. doi:10.18653/v1/W15-4640.

[20] O. Dušek, F. Jurčíček, A context-aware natural language generator for dialogue systems, in: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, Los Angeles, 2016, pp. 185–190. URL: https://aclanthology.org/W16-3622. doi:10.18653/v1/W16-3622.

[21] H. Zhou, M. Huang, X. Zhu, Context-aware natural language generation for spoken dialogue systems, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 2032–2041. URL: https://aclanthology.org/C16-1191.

[22] H. Zhang, M. Liu, Z. Gao, X. Lei, Y. Wang, L. Nie, Multimodal dialog system: Relational graph-based context-aware question understanding, in: Proceedings of the 29th ACM International Conference on Multimedia, MM '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 695–703. URL: https://doi.org/10.1145/3474085.3475234. doi:10.1145/3474085.3475234.

[23] C. Snell, S. Yang, J. Fu, Y. Su, S. Levine, Context-aware language modeling for goal-oriented dialogue systems, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 2351–2366. URL: https://aclanthology.org/2022.findings-naacl.181. doi:10.18653/v1/2022.findings-naacl.181.

[24] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, Dialoguernn: An attentive rnn for emotion detection in conversations, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 6818–6825. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4657. doi:10.1609/aaai.v33i01.33016818.

[25] S. Lee, Q. Zhu, R. Takanobu, Z. Zhang, Y. Zhang, X. Li, J. Li, B. Peng, X. Li, M. Huang, J. Gao, ConvLab: Multi-domain end-to-end dialog system platform, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Florence, Italy, 2019, pp. 64–69. URL: https://aclanthology.org/P19-3011. doi:10.18653/v1/P19-3011.

[26] M. Singh, Whatsapp tops 7 billion daily voice messages, 2022. URL: https://techcrunch.com/2022/03/30/people-are-sending-7-billion-voice-messages-on-whatsapp-every-da.

[27] S.-H. Huang, C.-Y. Huang, Y.-F. Lin, H. Ting-Hao Kenneth, What types of questions require conversation to answer? a case study of askreddit questions, in: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, CHI EA '23, Association for Computing Machinery, 2023. To appear.

[28] Y. Zhang, S. K. Jauhar, J. Kiseleva, R. White, D. Roth, Learning to decompose and organize complex tasks, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021.

## A. Details of PATROL STUDY

Figure 3 shows the procedure of the Patrol Study.

## A.1. Room-checking Task

In the room-checking task, we asked users to conduct a 15-minute patrol task inside a university building. How we created the room list, provided introductions to users, and conducted the patrolling process is described below.

**Creating the Room Lists.** Firstly, we created a list of room numbers of the building. The building has 200,000 square feet of floor area and three floors. Users could access two elevators and multiple staircases at all times. A total of 30 rooms were selected for users to navigate. The selected rooms were classrooms, research labs, and meeting rooms spread across all three floors of the building; all had windows with a view from the hallway, as the study required participants to look in the window to take notes on how many people were in the room. The 30 rooms were randomly distributed into three lists containing similar numbers of rooms on each floor so that each user was required to travel approximately the same distance. The order of rooms within each list was randomized. The goal was to have all the lists require the same level of physical and cognitive effort from the participants. The study consisted of three room-checking sessions; therefore, it was ideal to keep the variation of difficulties between lists as minimal as possible. We also asked users to take their time and check as many rooms on the list as they possibly can without rushing, even though we did not intend for the users to finish the entire list within the 15-minute time frame. Empirical experience showed that five to seven rooms can be checked in one session under such room arrangements.

**User Instruction Details.** We instructed the users to find the listed rooms. Particularly, we asked them to record the number of people inside the room and record the time they checked. Additionally, we suggested but did not require that they follow the order of rooms on the list. Users were informed that the results of the room-checking and information-seeking would be treated equally, and encouraged to complete both tasks as much as possible.

**Patrolling Process.** After the instruction, the users started to navigate the building based on their room list. We constrained the building navigation task to last 15 minutes. We specifically asked one author of the paper to provide optional in-person assistance during the user's patrolling process. Before the navigation began, the in-person helper briefly introduced the building layout if the users were not familiar with it. Thereafter, the in-person helper remained in a convenient position (*e.g.*, around the elevator) to respond if the users asked for help. When the navigation had lasted 15 minutes, the in-person helper found the users and told them their time was up.

## A.2. Information-seeking Task

While users performed the room-checking task, they were asked to reach out to the remote helper via WhatsApp on their smart phone in order to answer the question they were given for this task. We chose the WhatsApp platform because it can take both text and voice input and can export the conversation log (both text and voice messages) for further analysis. The remote helper updated their progress to a

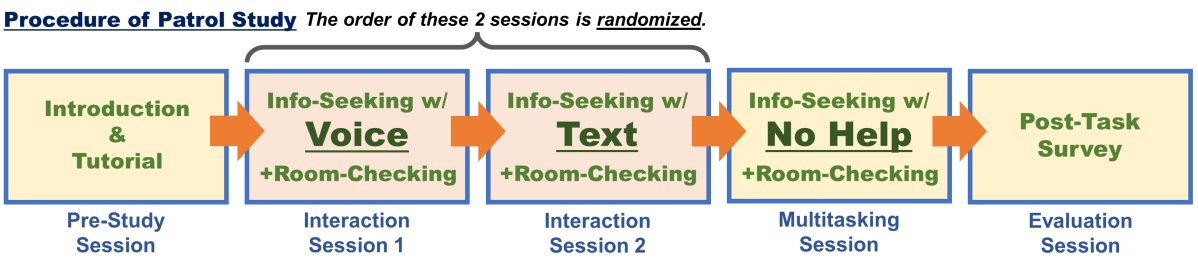


**Figure 3:** Patrol study contains a pre-study session, two interaction sessions (the order of the text and voice session is randomized), a multitasking session, and an evaluation session.

shared Google Doc accessible by users via their smart phone, so the user could monitor the progress of the remote helper as they gathered data at the user's instruction.

We asked users to give a short oral debriefing after they finish the 15 minutes of room-checking and information-seeking.. The purpose of the presentation was to assess the amount and accuracy of information obtained through an information-seeking task. We hypothesized that using a better communication channel could potentially enhance the effectiveness of information retrieval and improve the accuracy of the information gathered. We considered the scenario in real life when people reach out to others for help. It is likely that they will go through the received information, do further preparation, and validation before actually using the information. Users were asked to compose an oral presentation on how to carry out the task based on the information in the Google Doc. They were allowed to use the web links provided by the remote helper on the shared document for details and search for additional information if necessary. Users were not prohibited from searching for additional information on the internet. We believe that it is more realistic to acknowledge that people may want to verify the information they have been provided with or might have additional thoughts after taking the time to gather their thoughts. However, due to the time constraints, participants had to prioritize their efforts between searching for new and organizing existing information accordingly. Users had up to seven minutes to prepare for their presentation and were encouraged to speak for three minutes, or as long as they needed to convey the message they wanted to deliver. The debriefing was audio-recorded for further analysis.

In the following, we describe how we prepared the questions and the workflow of the remote helper.

**Selecting the Questions and Preparing the Answers.** The questions for the information-seeking tasks were selected from MSComplexTasks dataset [28], where a list of complex tasks is broken down into subtasks. Complex tasks were defined as one *task* requiring two or more individual steps for its completion. The individual steps needed to complete a complex task were considered the *subtasks*. The three topics were (*i*) how to write a business report, (*ii*) how to write a nonfiction book, and (*iii*) how to start a baking business. The number of subtasks required for each task was 14, 12, and 15 respectively. According to the subtasks and dependencies of the subtasks provided in MSComplexTasks [28], the subtasks for each selected task were pre-arranged into a list in ideal order. The remote helper did not collect additional information aside from pre-arranging the listed subtasks and listing the corresponding source web links provided in the dataset. As the users reached out to the remote helper for information on a certain task, the remote helper first updated the shared document with the pre-arranged subtasks and then asked if the user need any additional information. All requests from users outside the scope of the pre-arranged subtasks were researched by the remote helper and discussed with the users in real time. The helper also had access to the results of prior searches and could reuse the information.

**The Workflow of the Remote Helper.** Upon receiving a request from a user, the remote helper first listed the subtasks required to finish the requested task according to MSComplexTasks [28]. The remote helper then followed the user's instructions to look for further information on the internet or answer follow-up questions. While the remote helper performed the search, the user was in charge of directing their efforts. For the conversation in WhatsApp, the remote helper always responded in the form of text. Users, on the other hand, communicated to the remote helper using texting in one session and voice messaging in the other. The remote helper used the desktop version of WhatsApp to communicate with users and was able to hear users' voice messages.

## A.3. Pilot Study

Before the formal study, we conducted a small set of pilot studies with four participants to test the procedure. The first three participants were asked to complete the two interaction sessions (Text and Voice conditions) but not the multitasking session. Informal discussions with the participants inspired us to add the No-Help condition. We also adjusted the room list to avoid some rooms that were too hard to find or required special access. In response to the feedback of users, slight changes to the room

arrangement were made in order to balance the room distribution across different lists. The session outline and room list were finalized after considering the feedback provided by the fourth participant.

## B. Example Conversations for Study 1

We show four complete conversations including both voice and text. As we can see, the remote helper's responses were mostly short and also tried to confirm what information was being searched.

### Voice conversation on topic "How to start a baking business." (P7)

**User: Where is the best bank to try to get a loan from? Like which one has the best interest rate?**
Helper: I will also look for loan options
**User: What licenses do I need in order to start my business?**
**User: Also, where can I find information on how to apply for these licenses?**
Helper: checking licenses needed for baking business
**User: Will I need to have anything notarized**
**User: Also I'm curious how much money I should have saved up in order to start my business. Like my own personal money just in case**
Helper: okay, also looking for where to apply licenses
Helper: Will look up on that part also

### Text conversation on topic "How to start a baking business." (P12)

**User: how to start a baking business**
Helper: let me do the search, I will update in the google doc
**User: what permits and licenses are necessary for a baking business?**
Helper: looking into the permits and licenses required
**User: what equipment is necessary?**
Helper: I will look into that
**User: good advertising strategies for a first time bakery owner?**
Helper: okay, looking for advertising strategies for first time owner

### Voice conversation on topic "How to write a nonfiction book." (P10)

**User: What are the components of a nonfiction book?**
Helper: let me update the information in the shared document
**User: What are the tips to write a good nonfiction?**
Helper: Updated some topics of nonfiction books and the definition of it
Helper: looking for tips to write a nonfiction book
**User: Give me some tips on how to write nonfiction books.**
Helper: listing some steps to write nonfiction books
**User: Who are some of the famous nonfiction writers in English language?**
Helper: looking for famous nonfiction writer in English language
**User: Who are the famous nonfiction writers in English language?**
**User: And how to publish a nonfiction book?**
Helper: updated the top selling nonfiction books and writers
Helper: listing publishing methods for nonfiction books

**Text conversation on topic "How to write a nonfiction book." (P8)**

**User: What are a fee of rhe most successful non fiction titles**

**User: What kind of audiences read non fiction**

Helper: are you thinking about the price of them?

**User: What are a few of the most**

**User: Looking for topics and audience**

Helper: I see, let me look them up

**User: Could you look up as well, common non fiction topics**

Helper: Will do

Helper: [Reply to "Looking for topics and audience"] I do not think there are specific target groups for nonfiction books

**User: could u find average length and intensity of a non fiction book?**

Helper: okay

Helper: can you elaborate on the intensity aspect you are looking for?

**User: yes. if theyre mostly narrative and if so what kind of narrative, or mostly informative**

Helper: I see

Helper: let me check