Automated Administration of Questionnaires during Casual Conversation using Question-Guiding Dialogue System

Yuki Zenimoto[1], Mariko Yoshida[1], Ryo Hori[1], Mayu Urata[1], Aiko Inoue[2], Takahiro Hayashi[3], Ryuichiro Higashinaka[1]

[1]Graduate School of Informatics, Nagoya University
[2]Institute of Innovation for Future Society, Nagoya University
[3]AICHI Medical College of Rehabilitation

zenimoto.yuki.u1@s.mail.nagoya-u.ac.jp, {mariko, horiryo}@nagoya-u.jp, mayu@i.nagoya-u.ac.jp, inoue.aiko.u7@f.mail.nagoya-u.ac.jp, hayashi-taka@yuai.ac.jp, higashinaka@i.nagoya-u.ac.jp

Abstract

In an aging society, regularly assessing the health of older adults is increasingly important. Although questionnaires are commonly used for this purpose, the large number of items and the need for regular administration impose a considerable burden on older individuals. In this study, we propose a question-guiding dialogue system that naturally elicits responses to target questions through casual conversations. Our system maps free-form user responses to predefined questionnaire choices, enabling the collection of data in the same format as standard paper-based questionnaires. We conducted human evaluation experiments, followed by a two-week demonstration experiment in which older adults interacted with our system. The results show that our proposed approach achieves relatively high agreement with paper-based questionnaires.

1 Introduction

In an aging society, regularly assessing the health of older adults is increasingly important (Fried et al., 2001; Dent et al., 2019). Although paper-based questionnaires are commonly used for this purpose, the large number of items and the need for regular administration impose a considerable burden on older individuals.

As a potential solution to these issues, numerous studies have examined dialogue systems that elicit the health status of interlocutors (DeVault et al., 2014; Fadhil, 2018; Liu et al., 2019; Jo et al., 2024). These systems typically ask predefined questions in sequence and pose follow-up questions when needed. However, even when these systems are used to collect questionnaire answers, older adults must still converse with the system for a certain period of time, answering questions, meaning that the burden of completing a questionnaire remains essentially unchanged.

In this study, we propose a question-guiding dialogue system that naturally elicits answers to target questions through casual conversations, enabling health assessment without imposing a burden on older adults (Figure 1). We evaluate our approach in two ways: (1) a human evaluation experiment with text-based dialogues on a crowdsourcing platform and (2) a real-world demonstration experiment in which older adults used our spoken dialogue system for two weeks. We evaluate the degree of agreement between the dialogue-based answers obtained via the dialogue system and the questionnaire-based answers obtained via a paper-based questionnaire. Our results show that our proposed method can naturally elicit answers to questionnaire items while maintaining relatively high agreement with conventional paper-based questionnaires.

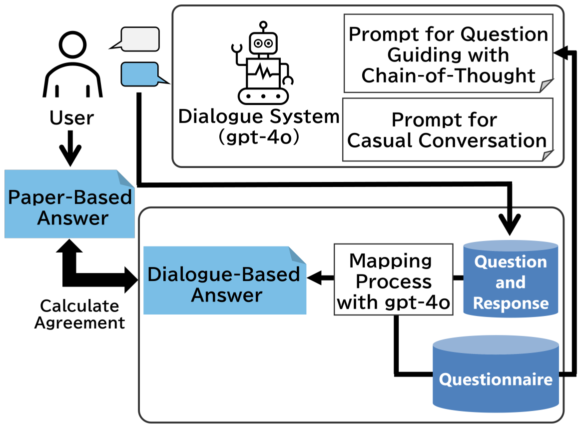

Figure 1: Overview of process for collecting questionnaire answers through question-guiding dialogue system and mapping

2 Related Work

Various studies have been conducted on dialogue systems for eliciting answers to desired questions (Geiecke and Jaravel, 2024; Hashimoto et al., 2025). These studies proposed dialogue systems that conduct interviews using large language models (LLMs) (Brown et al., 2020), aiming to collect

Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue, September 3–5, 2025, Bielefeld, Germany.

the thoughts and experiences of the interlocutor. However, since the goal of these systems is to collect free-form responses, they are not suitable for collecting answers to single-choice questions, in which a respondent selects one option from multiple predefined choices commonly used in existing questionnaires. In addition, these systems do not consider casual conversation and do not explicitly guide a conversation toward a specific context to collect the desired information.

Much research has been conducted to investigate methods for guiding a conversation toward a specific context. Some approaches generate guiding utterances based on keywords that can be transitioned from current contextual keywords (Tang et al., 2019; Xu et al., 2020; Liu et al., 2020), while others explicitly introduce a target topic to guide the conversation directly (Gupta et al., 2022; Kishinami et al., 2022). Although these studies effectively handle topic transitions, we focus on question-guiding, which aims to guide a conversation to ask a specific question naturally.

For asking questions naturally in dialogue, Horiuchi and Higashinaka (2021, 2023) proposed a question-guiding dialogue system, which was constructed by fine-tuning Japanese Dialogue Transformers (Sugiyama et al., 2023) with human-human question-guiding dialogue data. Inspired by their approach, our system leverages an LLM with a chain-of-thought (CoT) prompt (Wei et al., 2022; Kojima et al., 2022) to facilitate smooth transitions toward a target question. Moreover, whereas previous work focuses mainly on general questions (e.g., hobbies, fashion), we specifically focus on questions from medical questionnaires to elicit older adults' health status.

3 Proposed Method

Our goal is to naturally collect answers to specific questionnaire items in a casual conversation. To achieve this, we propose (1) a question-guiding method that smoothly transitions to a target question and (2) a mapping method that aligns free-form user responses with predefined questionnaire choices. By integrating these two methods, we create a questionnaire dialogue system that elicits answers to specific questionnaire items in a casual conversation (Figure 1).

Generate a system utterance that guides the conversation flow so that you can naturally ask the target question. Before generating your final utterance, provide a thought section that covers:

(1) Whether the conversation is ready for the target question,
(2) The current conversation topic,
(3) What the target question is about,
(4) How to naturally lead to the target question.

Then, output the final utterance, which must include the exact target question.

Keep your output concise. Avoid abrupt topic shifts using connective words such as "by the way," "speaking of which," etc. that may disrupt the flow.

Target question
"Have you ever made a mistake because you drank too much at a bar?"

Current context
User: "Good evening. It's already late at night."
System: "Good evening, indeed. Do you have a camera besides your smartphone?"
. . .
System: "That's something they'll enjoy even when they grow up. Speaking of which, I recently lost my smartphone and it was a disaster."
User: That's serious! Were you able to find it?

Thought
Let's think step by step.
Is the conversation already suitable for asking the target question?: Not yet
What is the current topic?: Possibly about losing one's camera or phone
What is the target question about?: A mistake due to drinking too much at a bar
How to transition naturally?: We could say that the phone was lost because of drinking too much at a bar. After responding in a way consistent with the conversation, we can mention that we lost our phone due to overdrinking at a bar and then ask whether the user has had a similar experience.

Output
System: "I drank too much at a bar and left it behind. These days it's hard to live without a smartphone, right? Have you ever made a mistake because you drank too much at a bar?"

Figure 2: Example of Question-Guiding Prompt C with CoT (This prompt was originally in Japanese and translated into English by the authors.)

3.1 Question-Guiding with Chain-of-Thought

In question-guiding, the system needs to smoothly transition the conversation toward the target question a few turns in advance. Following the method of Horiuchi and Higashinaka (2021, 2023), we aim to generate three consecutive system utterances: (1) a first guiding utterance, (2) a second guiding utterance, and (3) a final utterance that explicitly asks the target question. To implement this, we

Please output which choice from the following question best matches the user's answer.
The user's response may contain spelling mistakes, omissions, or be somewhat ambiguous.
Map it to the closest possible choice whenever possible.

Output only the choice.
If the user's response does not match any choice, output "Invalid response."

Question
{question} (Choices: {choices})

User's response
{response}

Figure 3: Example prompt for mapping user response to predefined choices (This prompt was originally in Japanese and translated into English by the authors.)

created three types of prompts:

Question-Guiding Prompt A Used for the first utterance to make a small transition toward the target question.

Question-Guiding Prompt B Used for the second utterance to make a more significant transition toward the target question.

Question-Guiding Prompt C Used for the third utterance to actually ask the target question.

Because question-guiding requires a nuanced understanding of both the current context and the semantic aspects of the target question, simple prompts may not suffice. We therefore use a CoT prompt in which the system is guided to reason step by step before generating the final utterance. Specifically, we instruct the LLM to output four pieces of reasoning before generating the utterance: (1) whether the conversation is ready for the target question, (2) the current topic, (3) the meaning of the target question, and (4) how to naturally transition toward the target question. An example of Question-Guiding Prompt C is shown in Figure 2.

Moreover, it is known that including concrete examples in prompts improves performance on various reasoning tasks (Brown et al., 2020), and this effect has been shown to hold true for CoT (Kojima et al., 2022). In this study, we thus include five CoT examples with intermediate reasoning within the prompts.

3.2 Mapping User Responses to Questionnaire Choices

Once the user's response to a question is obtained via question-guiding, it must be mapped to the predefined choices of the questionnaire. However, a user's free-form response in casual conversation will not necessarily match the prepared options, and in some cases, the user may not answer the question at all. Simple pattern matching is thus likely to be inadequate.

To address this, we use an LLM to associate a free-form user response with the predefined choices. The prompt contains (1) the question text, (2) the predefined choices, and (3) the user's response, with instructions to decide whether the user's response corresponds to any choice or if it is effectively "no valid answer." We also inform the LLM that the user's response may contain typos, omissions, or ambiguities. An example prompt is shown in Figure 3.

4 Experiment

We conducted experiments to evaluate both the question-guiding ability and the accuracy of the questionnaire responses collected by our system. All experiments, including the demonstration experiment described later in Section 5, were conducted with appropriate institutional review board approval.

4.1 Questionnaire Dialogue System

We implemented a questionnaire dialogue system that, for a specified question and timing, conducts the question-guiding described in Section 3.1. In utterances where no guiding is performed, the system engages in casual conversation. Once the user responds to the target question, we map the response to one of the questionnaire's predefined options by using the method described in Section 3.2. Note that the language of the system is Japanese.

4.2 Evaluation of Question-Guiding Ability

To evaluate the validity of our question-guiding method, we first conducted an evaluation using general questions. Following (Horiuchi and Higashinaka, 2021, 2023), we prepared 50 general questions on a variety of topics, including those that are difficult to ask, such as "Do you have any close friends?" or "Have you ever borrowed money to buy something?" to evaluate the question-guiding ability. Next, we collected text-based question-guiding dialogues with the questionnaire dialogue system on a crowdsourcing platform[1].

To examine the effectiveness of the CoT prompts, we implemented two versions of the system:

[1] https://crowdworks.jp/

System	Dialogue Naturalness	Dialogue Consistency	Dialogue Comprehension	Dialogue Interest	Dialogue Satisfaction	Question Naturalness
gpt-4o 0-shot	3.74 (0.82)	3.90 (0.78)	4.00 (0.87)	3.24 (0.99)	3.68 (0.86)	2.26 (1.07)
gpt-4o CoT 5-shot	3.82 (0.86)	3.86 (1.10)	3.90 (0.98)	3.22 (0.97)	3.52 (1.04)	$\mathbf{3.16}$ $(\mathbf{1.20})^*$

Table 1: Subjective evaluation results (average and standard deviation in parentheses) of question-guiding dialogues. * indicates statistically significant difference from gpt-4o 0-shot ($p < 0.01$). Steel-Dwass test (Dwass, 1960) was used for multiple comparisons.

System	Dialogue Naturalness	First Guiding Effectiveness	Second Guiding Effectiveness	Question Naturalness
Human	3.37 (1.21)	2.73 (1.28)	3.10 (1.26)	3.29 (1.39)
gpt-4o 0-shot	2.99 (1.19)	2.68 (1.16)	3.03 (1.22)	2.86 (1.35)
gpt-4o CoT 5-shot	$\mathbf{3.45}$ $(\mathbf{1.19})^*$	$\mathbf{3.16}$ $(\mathbf{1.14})^{*,\dagger}$	$\mathbf{3.65}$ $(\mathbf{1.02})^{*,\dagger}$	$\mathbf{3.46}$ $(\mathbf{1.31})^*$

Table 2: Results of evaluating question-guiding ability (average and standard deviation in parentheses). * indicates statistically significant difference from gpt-4o 0-shot ($p < 0.01$). † indicates statistically significant difference from human ($p < 0.01$). Steel-Dwass test (Dwass, 1960) was used for multiple comparisons.

gpt-4o 0-shot (baseline) GPT-4o-based system without CoT or examples.

gpt-4o CoT 5-shot GPT-4o-based system with CoT, including five examples of question-guiding in the prompts.

We recruited a total of 25 crowdworkers. In each dialogue, the crowdworker and the system produced 11 utterances each (22 in total). The timing of asking the target question was randomly chosen between the system's 5th and 10th utterances. Each crowdworker participated in a total of four dialogues, engaging in two dialogues with gpt-4o 0-shot and two dialogues with gpt-4o CoT 5-shot, presented in random order. A topic (e.g., fashion, games) was randomly assigned to the crowdworker, and the conversation started with that topic. Consequently, each system conducted 50 dialogues, for a total of 100 dialogues. Each crowdworker provided ratings on a 5-point scale for the following items after each dialogue, with one exception: Question Naturalness was rated after completing all dialogues, focusing only on the target question and its preceding context.

Dialogue Naturalness Were the system's responses natural in the context of the dialogue?

Dialogue Consistency Were the system's responses consistent and free of contradictions?

Dialogue Comprehension Did the system appear to understand your utterances appropriately?

Dialogue Interest Was the topic discussed in the dialogue interesting?

Dialogue Satisfaction Were you satisfied with the overall dialogue?

Question Naturalness Was the question asked in a natural flow of conversation?

Table 1 shows the results. In terms of Question Naturalness, the gpt-4o CoT 5-shot system significantly outperformed the gpt-4o 0-shot baseline, demonstrating the effectiveness of chain-of-thought prompting for question-guiding. In contrast, with respect to other dialogue quality metrics such as Dialogue Naturalness and Dialogue Consistency, the two systems achieved similarly high scores. This finding suggests that the overall dialogue quality is maintained even when question-guiding is performed.

The above subjective evaluation assessed the naturalness of the question within the dialogue as a whole but did not directly evaluate whether the conversation was appropriately guided. Therefore, we conducted a separate crowdsourced evaluation of these 100 dialogues to measure how naturally each system guided the conversation. For comparison with human question-guiding, we also included 50 human-human question-guiding dialogues on the same 50 questions we separately collected. Thus, we evaluated a total of 150 dialogues. A total of 30 crowdworkers participated in this evaluation, with each crowdworker evaluating 15 dialogues. Consequently, each dialogue was evaluated by three crowdworkers, each of whom was presented with the dialogue leading up to the target question, along with the corresponding target question, and rated them on a 5-point scale for the following items:

User:	I often buy generic brand products because they're cheaper. Prices have gone up lately.
System:	They are a good deal. Are there any products you've noticed have gone up in price recently?
User:	Yogurt. It used to be less than 100 yen last year, but now it's around 120 yen.
System:	That's quite an increase. When prices go up, managing the household budget can be tough. What do you do to save money on big purchases?
User:	For vegetables, meat, and fruit, I only buy at the supermarket because freshness matters. But for electronics, I usually buy from Rakuten or Yodobashi because of the point rewards because they're more cost-effective.
System:	It's great when you get point rewards on big purchases like electronics! **Have you ever borrowed money to buy something?**

(a) Example of question-guiding dialogue by gpt-4o 0-shot

User:	I try not to think about work so I can relax.
System:	I see. Sometimes shopping helps relieve stress, doesn't it?
User:	With prices going up, shopping can be stressful.
System:	That's tough indeed. Maybe splurging on something big once in a while could be nice. Have you bought anything expensive recently?
User:	I bought a slightly expensive toy for my child at Christmas.
System:	Christmas gifts are special, right? I once wanted something so bad that I borrowed money to buy it. **Have you ever borrowed money to buy something?**

(b) Example of question-guiding dialogue by gpt-4o CoT 5-shot

User:	If you're regularly replacing them, that gives a clean and tidy impression. That's nice.
System:	I tend to save on fashion, but I often spend more on food. Do you have something like that?
User:	I might be the type to spend more on computer-related things.
System:	I see. What is the most expensive item you have related to your computer?
User:	The computer itself, but I am planning to buy a high-capacity external battery soon. That might end up being the most expensive.
System:	That sounds expensive. **Have you ever borrowed money to buy something?**

(c) Example of question-guiding dialogue by human

Table 3: Examples of question-guiding dialogue for asking target question "Have you ever borrowed money to buy something?" Target question is shown in bold. (These dialogues were originally in Japanese and translated into English by the authors.)

Dialogue Naturalness How natural are the three consecutive turns from the first guiding utterance to the target question?

First Guiding Effectiveness Did the first guiding utterance help to ask the target question naturally?

Second Guiding Effectiveness Did the second guiding utterance help to ask the target question naturally?

Question Naturalness How natural is the target question in the context?

Table 2 shows the results. Overall, gpt-4o CoT 5-shot outperformed gpt-4o 0-shot. This indicates the importance of CoT prompts in generating natural and contextually appropriate question-guiding. Surprisingly, gpt-4o CoT 5-shot slightly outperformed human performance. This result can be attributed to the fact that question-guiding can be challenging even for humans.

Table 3 shows example dialogues from each system. An analysis of dialogues with low question naturalness revealed that, in some cases, the question not only appeared abrupt but was also asked even though the user had already provided an answer. Since our question-guiding dialogue system asks the target question at a predetermined turn between its 5th and 10th utterances, it sometimes asks a redundant question even when the answer has already been implicitly or explicitly given earlier in the dialogue. Addressing this will be an important direction for future work.

Mapping Method	Matching Method	2 Choices	4 Choices	5 Choices	Overall
gpt-4o	Exact	0.737 (70 / 95)	0.476 (20 / 42)	0.667 (46 / 69)	0.660 (136 / 206)
	Partial	—	0.786 (33 / 42)	0.870 (60 / 69)	**0.791** (163 / 206)
Human	Exact	0.842 (80 / 95)	0.533 (24 / 45)	0.635 (47 / 74)	0.706 (151 / 214)
	Partial	—	0.822 (37 / 45)	0.838 (62 / 74)	**0.836** (179 / 214)

Table 4: Results of agreement between dialogue-based answers and questionnaire-based answers from crowdsourcing (Responses judged not to contain answers to question by mapping were excluded from this calculation.)

4.3 Application to Medical Questionnaires

We next compared how accurately the system-collected responses matched the answers from paper-based questionnaires. Specifically, we examined 22 medical questionnaire items related to health and social engagement (Fried et al., 2001; Dent et al., 2019), for example: "Do you often cough or choke on liquids like tea or soup? (Yes/No)" and "How often do you feel that you do not have social relationships? (Never, Seldom, Sometimes, Always)." This questionnaire is designed to help detect health deterioration in older adults without access to professional caregivers in their surroundings and facilitate appropriate interventions, thereby contributing to the extension of their healthy life expectancy. The questionnaire has been widely used in Japan, and its effectiveness has been empirically validated (Murayama et al., 2020).

We collected dialogues between the system and human on the above crowdsourcing platform. We recruited a total of 66 crowdworkers. In each dialogue, the crowdworker and the system produced 20 utterances each (40 in total). The system asked a target question every 6 to 9 turns, resulting in two target questions per dialogue. Each crowdworker participated in a total of two dialogues, engaging in one dialogue with gpt-4o 0-shot and one dialogue with gpt-4o CoT 5-shot, presented in random order. A topic (e.g., fashion, games) was randomly assigned to the crowdworker, and the conversation started with that topic. After completing all dialogues, each crowdworker answered the corresponding paper-based questionnaire items. As a result, we collected a total of 132 dialogues and 264 responses to the target questions, with each of the 22 questions being asked 12 times.

We then mapped the user responses (i.e., utterances immediately following the target question) to the predefined questionnaire choices using the mapping method described in Section 3.2. Here, we used gpt-4o as an LLM and compared these dialogue-based answers with the questionnaire-based answers. While the questionnaire included one free-form question and one multiple-choice question, we excluded these two questions from the evaluation to focus on single-choice questions. As a result， we collected 240 (i.e., 20 questions being asked 12 times) pairs of dialogue-based answer and questionnaire-based answer. We computed the agreement separately for items with 2 choices, 4 choices, and 5 choices. For those with the 4 or 5 choices, we also computed a "partial-match" score, wherein options like "Very applicable" and "Applicable" are considered a match.

To evaluate the performance of the mapping method using gpt-4o, we also conducted a manual mapping by recruiting 12 crowdworkers. This mapping task does not require medical expertise; therefore, we employed general crowdworkers for this evaluation. Each crowdworker was presented a question and the corresponding user response and was asked to map the response to the most appropriate choice. They also classified whether the response contained an answer to the question (see Section 6). For each question–response pair, three crowdworkers performed the mapping. When two or more crowdworkers selected the same option, that option was taken as the representative choice. When all three crowdworkers selected different options, the middle option was used as the representative choice for that response.

The upper half of Table 4 shows the results of gpt-4o. Overall, the dialogue-based answers achieved an exact-match agreement of 0.660 across all items and a partial-match agreement of 0.791. Hence, while there remains room for improvement, the results indicate that the proposed method can obtain questionnaire answers reasonably close to those obtained via paper-based questionnaire.

The lower half of Table 4 shows the results of the manual mapping. Compared with the gpt-4o-based mapping, the manual mapping resulted in a slightly higher agreement rate, with a difference of approximately 4%. This indicates that gpt-4o is capable of performing the mapping with accuracy comparable to that of human annotators. An analysis of responses that failed to be mapped correctly revealed that some user responses were clearly inconsistent with the paper-based answer, while others were ambiguous and could correspond to multiple options. Eliciting more appropriate responses during the conversation remains a key challenge for future work.

5 Demonstration Experiment

To investigate real-world feasibility, we conducted a two-week demonstration experiment in which older adults interacted daily with our questionnaire dialogue system. As in Section 4.3, we used the same 22 medical questionnaire items. The system was deployed as a spoken dialogue system implemented with an Amazon Echo Show, adopting the persona of a friendly college student (see Figure 4). We used the speech recognition and speech synthesis provided by Amazon Alexa Skill.

We recruited 11 older adults (all female), 6 aged 65–74 and 5 aged 75 or older. Over a two-week period, each participant conversed with the system at least twice a day. The system used gpt-4o CoT 5-shot to perform question-guiding every 6 to 9 turns while engaging in casual conversation using gpt-4o-mini in other turns. Prior to the experiment, we did not inform participants that medical questions would be asked; we revealed this only after the experiment. After the two-week period, they completed both a medical questionnaire and a system evaluation questionnaire. In the system evaluation, they rated dialogue satisfaction, dialogue naturalness, and perceived frequency of questions on a 5-point scale. Many participants interacted with the system more than three times per day, resulting in a total of 379 dialogues and 5,967 utterances throughout the experiment.

We then mapped the obtained dialogue responses to the questionnaire choices and compared them with the questionnaire-based answers. In total, we collected 434 pairs of dialogue-based answer and questionnaire-based answer. To evaluate the performance of the mapping method using gpt-4o, we also conducted a manual mapping. This mapping was carried out by two of the authors of this paper, following the same procedure as described

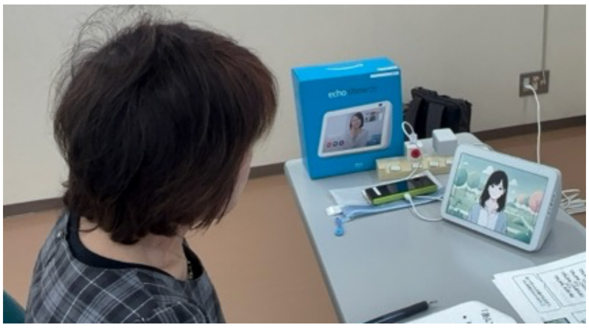

Figure 4: Dialogue system used in demonstration experiment

in Section 4.3. If the two authors selected different options, the more conservative option was used as the representative choice for that response. For example, for the question "Compared to six months ago, has it become harder for you to eat hard foods? (Yes/No)," the representative choice was set to "No."

Table 5 shows the results. The exact-match agreement was 0.580, but the partial-match agreement reached 0.716. While these scores indicate moderate agreement, they were lower overall compared with the text-based crowdsourcing data. Compared with the manual mapping results, the human annotators achieved slightly higher agreement than gpt-4o, with a difference of approximately 3%. This suggests that gpt-4o is capable of mapping with accuracy comparable to that of humans, even for data from the demonstration experiment.

Finally, we report the results of the system evaluation questionnaire. Regarding dialogue satisfaction, all participants answered "the conversations were enjoyable." In terms of dialogue naturalness, more than half of the participants answered "the system's responses felt natural," and only one participant answered "the system's responses felt unnatural." As for the number of questions asked, the majority of participants answered "it was just right," although a few answered "it was slightly more than expected." Since the system asks not only target questions but also general questions as part of casual conversation, one possible reason for this perception is that the system tended to ask many questions overall. These results suggest that the proposed dialogue system achieved high levels of user satisfaction and naturalness and that there were no major issues in the quality of the conversations. Notably, none of the participants realized that the medical questions had been asked.

Mapping Method	Matching Method	2 Choices	4 Choices	5 Choices	Overall
gpt-4o	Exact	0.601 (91 / 150)	0.592 (45 / 76)	0.531 (52 / 98)	0.580 (188 / 324)
	Partial	—	0.842 (64 / 76)	0.786 (77 / 98)	**0.716** (232 / 324)
Human	Exact	0.628 (86 / 137)	0.514 (37 / 72)	0.505 (51 / 101)	0.561 (174 / 310)
	Partial	—	0.861 (62 / 72)	0.812 (82 / 101)	**0.748** (232 / 310)

Table 5: Results of agreement between dialogue-based answers and questionnaire-based answers of demonstration experiment (Responses judged not to contain answers to question by mapping were excluded from this calculation.)

Response Type	Crowdsourcing	Demonstration Experiment
1. Explicit	499 (62.4%)	456 (52.5%)
2. Implicit	167 (23.2%)	136 (15.7%)
3. Vague	71 (9.9%)	114 (13.1%)
4. Unrelated	33 (4.6%)	100 (11.5%)
5. Incomplete	0 (0.0%)	62 (7.1%)
Total	720	868

Table 6: Distribution of response types

6 Analysis of User Response

In dialogue, users do not always provide a direct answer to the question being asked. Therefore, we examined whether participants actually provided valid answers in their responses. We manually classified each user's response to a question into five response types:

1. Explicit Response Response that provides a clear and direct answer to the question. (e.g., Q: "Compared to six months ago, has it become harder for you to eat hard foods?" A: "No, not at all.")

2. Implicit Response Response that does not provide a clear, direct answer, but the answer can be inferred. (e.g., Q: "How many meals do you usually have?" A: "I just have coffee in the morning, and then regular lunch and dinner.")

3. Vague Response Response that references the question's topic but provides no inferable answer. (e.g., Q: "Do you find food to taste good?" A: "I like rich flavors.")

4. Unrelated Response Response that is unrelated to the question, offering no relevant content. (e.g., Q: "Do you sometimes cough or choke when drinking tea or soup?" A: "What's the weather going to be like today?")

5. Incomplete Response Response that is truncated or syntactically broken, making it impossible to determine an answer. (e.g., Q: "Do you ever feel isolated from others?" A: "Well$\cdots$")

The crowdsourced data was annotated by 12 crowdworkers, while the demonstration data was annotated by two of the authors. Annotators were shown each question and its corresponding response, and they classified the response as one of the above five response types. For the crowdsourced data, each question-response pair was annotated by three crowdworkers, whereas for the demonstration data, each pair was annotated by two authors.

Fleiss' k was 0.407 for the crowdsourced data, indicating a moderate level of agreement, and Cohen's k was 0.616 for the demonstration data, indicating a substantial level of agreement. In the crowdsourced data, annotators frequently disagreed between Explicit Response and Implicit Response, which contributed to the lower k value. This issue may be mitigated by providing clearer annotation guidelines to the crowdworkers.

Table 6 shows the distribution of annotations aggregated across all annotators. In the demonstration experiment, the proportion of valid responses was lower than in the crowdsourced data, likely due to real-world constraints such as speech recognition errors and casual or off-topic replies in speech-to-speech interaction. This indicates the difficulty of reliably obtaining valid answers in speech-to-speech real-world settings.

7 Conclusion

We proposed a question-guiding method and a response mapping method for naturally collecting answers to desired questionnaire items through casual conversation. Our experiments demonstrated that a chain-of-thought prompt enables the system to guide questions smoothly. Further, both a human evaluation and real-world demonstration indicated that the system's dialogue-derived answers showed relatively high agreement with those from conventional questionnaires.

Future directions include dynamically selecting both the question timing and the target question on the basis of conversation flow, as well as robustly tracking previously answered items to avoid repetition. It is also necessary to ask follow-up questions when a clear answer to the target question is not obtained in order to ensure more reliable response collection. In addition, previous studies show that older adults tend to be particularly talkative in the interaction with dialogue systems, which can cause issues with speech recognition and language understanding (Wolters et al., 2009; Vipperla et al., 2009; Georgila et al., 2010); we would like to examine whether such issues occurred in our experiment. Finally, since this study focused solely on single-choice questions, handling multiple-choice questions and free-form questions remains an open challenge. Improving speech recognition and handling off-topic or incomplete replies are also crucial for enhancing real-world performance.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 24K14769, 23H00493, and 24K05433. We sincerely appreciate the invaluable cooperation of the residents and staff members of Kita Ward, Nagoya and Toyoyama Town, Aichi.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proc. NeurIPS*, pages 1877–1901.

E. Dent, J. E. Morley, A. J. Cruz-Jentoft, L. Woodhouse, L. Rodríguez-Mañas, L. P. Fried, J. Woo, I. Aprahamian, A. Sanford, J. Lundy, F. Landi, J. Beilby, F. C. Martin, J. M. Bauer, L. Ferrucci, R. A. Merchant, B. Dong, H. Arai, E. O. Hoogendijk, and 17 others. 2019. Physical Frailty: ICFSR International Clinical Practice Guidelines for Identification and Management. *Journal of Nutrition, Health & Aging*, 23(9):771–787.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, and 3 others. 2014. SimSensei kiosk: a virtual human interviewer for healthcare decision support. In *Proc. AAMAS*, page 1061–1068.

Meyer Dwass. 1960. Some k-sample rank-order tests. In *Contributions to Probability and Statistics*, page 198–202. Stanford University Press.

Ahmed Fadhil. 2018. Beyond patient monitoring: Conversational agents role in telemedicine healthcare support for home-living elderly individuals. *arXiv preprint arXiv:803.06000*.

Linda P. Fried, Catherine M. Tangen, Jeremy Walston, Anne B. Newman, Calvin Hirsch, John Gottdiener, Teresa Seeman, Russell Tracy, Will J. Kop, Gloria Burke, and Margaret A. McBurnie. 2001. Frailty in older adults: evidence for a phenotype. *Journal of Gerontology: Series A, Biological Sciences and Medical Sciences*, 56(3):M146–M156.

Friedrich Geiecke and Xavier Jaravel. 2024. Conversations at scale: Robust AI-led interviews with a simple open-source platform. *Social Science Research Network*, pages 1–73.

Kallirroi Georgila, Maria Wolters, Johanna D. Moore, and Robert H. Logie. 2010. The MATCH corpus: a corpus of older and younger users' interactions with spoken dialogue systems. *Language Resources and Evaluation*, 44(3):221–261.

Prakhar Gupta, Harsh Jhamtani, and Jeffrey Bigham. 2022. Target-guided dialogue response generation using commonsense and data augmentation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1301–1317.

Ekai Hashimoto, Mikio Nakano, Takayoshi Sakurai, Shun Shiramatsu, Toshitake Komazaki, and Shiho Tsuchiya. 2025. A career interview dialogue system using large language model-based dynamic slot generation. In *Proc. COLING*, pages 1562–1584.

Sota Horiuchi and Ryuichiro Higashinaka. 2021. Learning to ask specific questions naturally in chat-oriented dialogue systems. In *Proc. IWSDS*.

Sota Horiuchi and Ryuichiro Higashinaka. 2023. Learning to guide questions in chat-oriented dialogue by using combination of question-guiding corpora. In *Proc. IWSDS*.

Eunkyung Jo, Yuin Jeong, Sohyun Park, Daniel A. Epstein, and Young-Ho Kim. 2024. Understanding the impact of long-term memory on self-disclosure with large language model-driven chatbots for public health intervention. In *Proc. CHI*, pages 1–21.

Yosuke Kishinami, Reina Akama, Shiki Sato, Ryoko Tokuhisa, Jun Suzuki, and Kentaro Inui. 2022. Target-guided open-domain conversation planning. In *Proc. COLING*, pages 660–668.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proc. NeurlPS*, pages 1–15.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proc. ACL*, pages 1036–1049.

Zhengyuan Liu, Hazel Lim, Nur Farah Ain Suhaimi, Shao Chuen Tong, Sharon Ong, Angela Ng, Sheldon Lee, Michael R. Macdonald, Savitha Ramasamy, Pavitra Krishnaswamy, Wai Leng Chow, and Nancy F. Chen. 2019. Fast prototyping a dialogue comprehension system for nurse-patient conversations on symptom monitoring. In *Proc. NAACL*, pages 24–31.

Hiroshi Murayama, Erika Kobayashi, Shohei Okamoto, Taro Fukaya, Tatsuro Ishizaki, Jersey Liang, and Shoji Shinkai. 2020. National prevalence of frailty in the older japanese population: Findings from a nationally representative survey. *Archives of Gerontology and Geriatrics*, 91:104220.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical Analysis of Training Strategies of Transformer-Based Japanese Chit-Chat Systems. In *Proc. SLT*, pages 685–691.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proc. ACL*, pages 5624–5634.

Ravichander Vipperla, Maria Wolters, Kallirroi Georgila, and Steve Renals. 2009. Speech input from older users in smart environments: Challenges and perspectives. In *Proc. UAHCI*, pages 117–126.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. NeurlPS*, pages 1–14.

Maria Wolters, Kallirroi Georgila, Johanna D. Moore, and Sarah E. MacPherson. 2009. Being old doesn’t mean acting old: How older users interact with spoken dialog systems. *ACM Transactions on Accessible Computing*, 2(1):1–39.

Jun Xu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che. 2020. Knowledge graph grounded goal planning for open-domain conversation generation. *Proc. AAAI*, 34(05):9338–9345.