Tree-based Interview Topic Guidance for Collecting Target Information Under Adaptive Topic Continuation/Switching

Fuminori Nagasawa and Ekai Hashimoto and Shun Shiramatsu
Nagoya Institute of Technology,Aichi,Japan
nagasawa.fuminori@nitech.ac.jp

Abstract

Interview-style dialogue for service recommendation must both elicit users' underlying needs and collect information required for recommendation. We propose a Question Tree–based guidance method that steers dialogue toward target information while allowing externally controlled topic continuation and switching. The Question Tree represents questions as nodes linked by parent–child derivational relations. For each user response, the system generates deepening and target-approaching question candidates, appends suitable candidates to the tree, and selects the next question from child or sibling nodes according to the required topic-control action. Selection uses a weighted combination of sentence-embedding similarity to the current dialogue context and to the target information. We evaluated the method in LLM-based dialogue simulations with externally controlled continuation and switching. The combined method increased the mean target-information collection rate from 10.8% to 26.6% after 10 turns and reduced cross-session variability, without a statistically significant difference from the baseline in the limited human naturalness evaluation. These results indicate that explicit tree-based topic guidance can support information collection when combined with external topic-adaptation modules.

1 Introduction

One of the ultimate goals of our research is to build a dialogue robot that naturally encourages users to talk about their own feelings and interests. Dialogue techniques that elicit users' emotions and interests form a basis for effective human-computer interaction (HCI).

As one such user-adaptive application, we focus on interviews. Interviews have been widely studied as a method to collect information through question-answer interaction (Fontana and Frey, 2005). One important interviewing technique is to promote self-disclosure by appropriately following up on topics in accordance with the interviewee's attitude (Emans, 2016). By encouraging self-disclosure, a system can elicit detailed and personal narratives from users, including their interests, memories, and opinions.

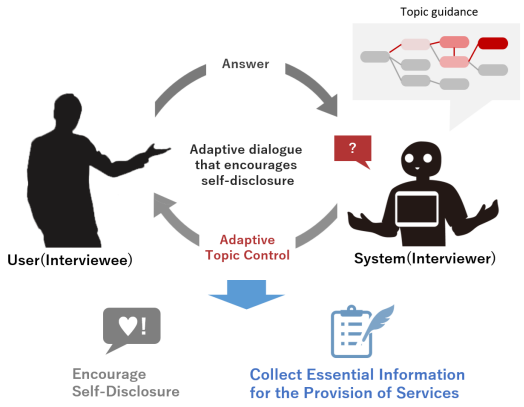

Figure 1: Research Overview: The system conducts 1-on-1interview with user, determining the next question based on a topic-guidance strategy for each user answer. This approach allows the system to reliably collect the information necessary for service provision and recommendations while simultaneously achieving other objectives, such as encouraging self-disclosure.

At the same time, for applying interviews to HCI systems, information collection is as important as the promotion of self-disclosure. To provide appropriate functions and services based on dialogue, the system must efficiently collect the information necessary for such service provision. Nagasawa et al. (Nagasawa and Okada, 2025) showed that adaptive control of topic continuation and topic shift based on user attitudes estimated from multimodal information during dialogue can encourage self-disclosure. However, because that method did not consider how to deepen a current topic or how to orient topic shifts, the conversation tended to become biased toward what the user simply wanted to talk about, making systematic information col-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 497–515
August 2–5, 2026.

lection difficult.

To address this issue, this paper proposes a Question Tree, which manages questions in a tree structure based on parent-child topical relations, and a method for guiding topics so that required information can be gradually elicited while still promoting self-disclosure. In the Question Tree, dialogue topics are arranged on a tree-structured graph according to their follow-up relations, and question selection through topic continuation and topic shift is realized as graph exploration. Based on the content of the user's response, the system uses an LLM to generate follow-up questions as well as questions that help approach the target topic, and then selects the next question from among the candidates using document-vector similarity to the target topic.

We implemented a prototype text-chat system incorporating the proposed method and evaluated its effect on the collection rate of target information through dialogue simulation experiments. We conducted multiple interview dialogues with an LLM-based dummy user while externally controlling topic continuation and topic shift at random, and analyzed how the collection rate of pre-specified target information changed after a fixed number of dialogue turns.

The contributions of this study can be summarized as follows:

- We verified that an explicit topic management and guidance method based on graph trees and text vector similarity improves the response collection rate for target information in interview dialogues while reducing variations in collection performance across dialogues.
- We analyzed the differences in target information collection between using each of the two functions comprising the proposed guidance method (generation of candidate guiding questions and vector-based topic selection) individually and using them in combination.
- We implemented a dialogue system incorporating the proposed topic-guidance method and demonstrated through simulation experiments that the proposed method functions even in situations where topic continuation and shift are externally controlled.

2 Related Works

2.1 Information Collection through Interview Dialogue

One important role expected of interview dialogue is information collection through conversation. Compared with structured interviews (Fontana and Frey, 2005), in which the order and content of questions are predetermined, semi-structured and unstructured interviews, in which questions can be arranged during the interaction, are more suitable for eliciting deeper information in response to the interviewee's reactions. Zenimoto et al. (Zenimoto et al., 2025) proposed a system utterance generation method that uses an LLM with chain-of-thought prompting to gradually guide casual conversation toward required question items so that necessary questions can be inserted naturally.

In contrast, our goal is to develop a method that can explicitly manage topic transitions through a data structure, enabling robustness to variation in guidance paths and the number of turns required to approach target topics, while also being integrable with other dialogue adaptation techniques.

2.2 Topic Management in Interview Dialogue

Emans et al. (Emans, 2016) describe appropriate follow-up of topics according to an interviewee's attitude as one of the important interviewing techniques.

Hashimoto et al. (Hashimoto et al., 2025) developed a slot-filling dialogue system for career interviews with nurses, in which the system forms hypotheses from users' responses and generates new questions accordingly, thereby realizing topic deepening through dynamic slot generation with an LLM. In this study, we similarly treat questions as slots, but manage them according to parent-child topic relations so that the method can be combined with self-disclosure-promoting approaches based on continuation/shift control.

2.3 User Needs Extraction by Dialogue Systems

Methods for extracting user needs through dialogue include approaches that infer functions matching ambiguous user utterances (Tanaka et al., 2021) and methods that select the best question for concretizing a user's request (Zhao and Eskenazi, 2016). However, because these methods rely on prior knowledge to refine requests and utterance content, their applicable domains are limited. For open-

domain dialogue, previous studies have proposed asking clarification questions when user utterances are ambiguous (Kuhn et al., 2022) and generating prerequisite questions for reaching a final question (Fu and Du, 2025). However, when many additional questions are asked, such methods may hinder user self-disclosure. Our goal is to extract user needs in a domain-independent manner while enhancing self-disclosure so that the system can obtain needs that are closer to the user's genuine intentions.

In works such as GuessWhat?!(de Vries et al., 2017) and FATA(Fu and Du, 2025), the importance of the system gradually transitioning from more abstract questions to more specific ones was highlighted. Our proposed topic-guidance method likewise adopts a gradual approach to the target information. However, we specifically develop and evaluate a method that remains effective even when external modules intervene in topic control during the process of approaching the target information.

2.4 Self-disclosure and Topic Adaptation

In addition to gathering necessary information, several studies have been conducted on methods to encourage self-disclosure in order to elicit more in-depth information from users. Self-disclosure is defined as "revealing information so that others can understand oneself"(Altman, 1973), and it has been noted that encouraging self-disclosure plays a crucial role in improving satisfaction and providing appropriate support(Papneja and Yadav, 2024; Apostol et al., 2025).

As dialogue systems designed to encouraging self-disclosure, Mitsuno et al.(Mitsuno et al., 2024) proposed a dialogue system that gradually shifts to deeper topics over time; and Nagasawa et al. (Nagasawa and Okada, 2025) proposed a system that estimates the user's emotions based on nonverbal cues and adaptively adjusts the topic. While these adaptive topic control methods can be applied to various dialogue systems, they have faced challenges in terms of information gathering because the topics collected tend to be biased toward what the user wants to talk about. One of the objectives of this study is to leverage these topic adaptation techniques to guide the conversation toward the information the system seeks, thereby extending the benefits of these methods to real-world applications.

Clavel et al. (Clavel et al., 2022) emphasized that dialogue technologies that engage users through various forms of emotion modeling while helping them accomplish tasks enhance the system's usefulness. This study is positioned as a technology that aids in task accomplishment when utilizing emotion modeling techniques to realize such integrated systems.

3 Method

To realize dialogue that balances adaptive topic follow-up and information collection according to the user's attitude, we constructed an interview dialogue system that manages system questions using a tree-structured graph. The system determines the next question by exploring the graph for each dialogue turn(see Figure2).

3.1 Question Selection Based on Exploration of a Question Tree

Questions are arranged on a tree-structured graph as nodes. Each node has parent and child nodes, and the tree is organized such that child nodes correspond to questions that further probe the topic of the parent node. The root node, which serves as the starting point of exploration, is given in advance as the initial question that starts the dialogue. Additional questions are generated in real time from the user's responses and added to the Question Tree.

3.2 Adding Questions to the Question Tree

When the user responds, the system generates candidate questions derived from that topic using an LLM and adds them as child nodes of the current question in the Question Tree. The prompt given to the LLM includes instructions for generating questions for topic continuation and topic shift, along with the user utterance and the status of responses to previous questions.

At this stage, the system generates two types of candidate questions:

Deepening questions: A deepening question continues the current topic and seeks more specific information. It emphasizes smooth topical continuity from the source topic while expanding related subtopics. Using keywords such as nouns and verbs appearing in the user's response, the system generates questions that ask for more concrete details or impressions. To ensure topical diversity among deepening questions, the prompt is designed to generate questions corresponding to 5W1H and to probe types for qualitative interviews (Hurst, 2023).

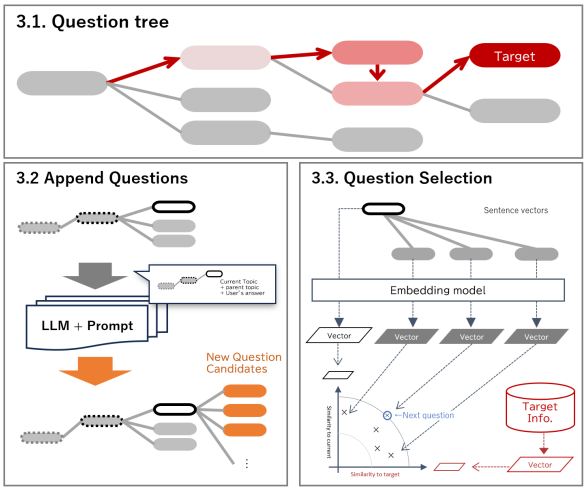

Figure 2: Overview of tree-based topic guidance. In Question-Tree, topics to be asked about are arranged as nodes in a tree. Follow-up questions are added based on the tree structure, and when selecting a question, an embedding vector is used to determine the next question from the candidates.

Approach questions: An approach question is generated so as to move from the current topic toward the target information. It is designed to connect the current topic, the user's response, and the target information through a shared subtopic. When multiple target information items exist, the system generates at least one approach question for each target item.

After generating these two types of candidate questions, the system filters out contextually unnatural ones and adds the remaining questions to the Question Tree as child nodes of the current topic.

3.3 Guided Question Selection

From the set of question candidates arranged on the Question Tree, the system selects the next question so that the dialogue remains loosely connected to the current topic while approaching a target topic expected to yield the desired information. The magnitude of a topic transition is estimated based on the cosine similarity of sentence embeddings. For each candidate question, the system computes its similarity to the current topic, S_c, and its similarity to the target topic, S_t. In this study, for the topic description of a candidate question, S_c was calculated as the similarity to the immediately preceding system-user question-answer pair, and S_t was calculated as the similarity to the description of a single piece of target information. The next question is then selected as the candidate that maximizes their weighted sum. Let r_c be the weight for S_c and r_t be the weight for S_t. The score for each candidate question is calculated as follows: $S_{blended} = r_c S_c + r_t S_t$.

When multiple pieces of target information exist, $S_{blended}$ is calculated for each one, and the question candidate with the highest $S_{blended}$ among all of them is selected.

Topic continuation and topic shift are realized by switching the pool from which candidate questions are explored. When continuing a topic, the next question is selected from the child nodes of the current topic. When shifting the topic, the next question is selected from the sibling nodes of the current topic, that is, other child nodes of its parent node. The selected question is then set as the new current topic, and the system generates an utterance that is presented to the user in a way that remains natural and consistent with the dialogue history.

3.4 Collecting Answer Values for Target Information

The dialogue log between the system and the user is analyzed, and if it contains references to the questions generated in Section 3.2 or to the target information, the corresponding answer is extracted and collected as an answer value. In this study, descriptions of unanswered items, the dialogue log, and task instructions were provided to an LLM, which generated the answer status in JSON text format. The JSON output was then parsed and stored as collected answer values.

4 Experimental Settings

To evaluate the effectiveness of the proposed question-guiding strategy, we constructed a text-based dialogue simulation environment and conducted experiments.

4.1 Simulation Environment

As the simulation environment, we implemented an interviewer system equipped with the proposed question-guiding strategy, a dummy interviewee system, and a text chat application for input and output, and conducted mock interview dialogue experiments. For the interviewer system, we used the open-weight model openai/gpt-oss-20b (OpenAI, 2025), running on a PC (Intel Core Ultra i7 265KF + NVIDIA GeForce RTX5090). For sentence embeddings, we used openai/text-embedding-ada-002 via the OpenAI API.

Interviewer system and chat application: The system was implemented as a web application that

Table 1: Conditions and settings

Condition	Candidate Questions	Selection Guidance
(A)Baseline	Deepening only	No
(B)+Approach	Deepening+Approach	No
(C)+Selection	Deepening only	Yes
(D)+All	Deepening+Approach	Yes

takes a user utterance and a topic continuation/shift flag as input and returns the next question as text. To ensure reproducibility, the temperature parameter of the GPT model used in the interviewer system was fixed at 0.0 for all calls.

Dummy interviewee system: As a participant simulator for the interview experiment, we implemented a dummy interviewee that generates responses using GPT. The dummy interviewee was given only an instruction to respond as an interview participant and the dialogue log as the prompt. No fixed persona was assigned, with the aim of allowing diverse personas to emerge randomly. The temperature parameter of the GPT model was set to 0.8, so that diverse responses could be obtained while maintaining conversational consistency.

4.2 Experimental design and procedure

To evaluate the effectiveness of the two functions that constitute the proposed method—(1) differentiation of generated questions and (2) guided node selection—we conducted dialogue simulations under the following four conditions:

To simulate a setting in which topic continuation and topic shift are externally controlled, the continuation/shift decision was randomly determined. The ratio was adjusted so that continuation and shift occurred at $50\% : 50\%$. To minimize the influence of the answer-value collection module on the results, the prompts and settings used in that module were fixed across all conditions.

The dialogue began with the fixed system question: "Hello! What has been going on with you recently?" After that, the dialogue automatically proceeded with the dummy interviewee system for 10 exchanges, after which it ended automatically.

We conducted 100 simulation sessions for each condition. To reduce the influence of topical bias on the results, we pre-generated 100 initial responses from the interviewee and shared the same set across all conditions, thereby matching the starting points of the dialogues. Furthermore, for the two parameters (r_c, r_t) used in the question selection for (C)+Selection and (D)+All, equal weights $(r_c = r_t = 0.5)$ were set in this study to ensure balance.

4.3 Analysis

To evaluate whether the proposed dialogue-guidance strategy can effectively guide questions even under externally imposed topic continuation and topic switching, we tested two hypotheses.

4.3.1 Effect on target information collection

The first hypothesis is that the proposed question-guidance method will enable the collection of more target information. To test this hypothesis, we compared the collection-rate (the number of responses received relative to the pre-set target information) under each simulation condition.

For each condition, we conducted 100 sessions consisting of 10 dialogue turns and calculated the proportion of target information collected. We then computed the average target-information collection rate at the end of the dialogue for each condition and analyzed the variability across sessions.

In this experiment, as a challenging proxy target set with varying elicitation difficulty, we used the 24 self-disclosure topics included in the self-disclosure scale of (Niwa and MARUNO, 2010): Level 1, hobbies; Level 2, difficult past experiences; Level 3, non-critical personal weaknesses; and Level 4, negative aspects of one's personality or abilities. For all sessions, we recorded the answer status of the slots at the end of the session and compared collection rates across self-disclosure levels.

4.3.2 Effect on dialogue naturalness

Our second hypothesis was that the proposed method affects the perceived naturalness of dialogue. To examine whether intervention in topic transitions introduced unnatural impressions, we conducted a human evaluation of the dialogue logs obtained from the simulations. For the evaluation of Section 4.3.1, we randomly selected 4 sessions each from the dialog logs collected during the evaluation. The questionnaire consisted of 10 items related to relevance, consistency, and the naturalness of topic transitions, selected from prior work on dialogue naturalness ((Mehri and Eskenazi, 2020), (Higashinaka et al., 2022)). Each evaluator was presented with a set consisting of one dialogue log and the questionnaire and rated each item on a seven-point Likert scale (for the questionnaire items, see Table 5 in Appendix). For each questionnaire item,

we conducted a Wilcoxon signed-rank test across conditions. To minimize order effects, logs from different conditions were mixed, and the evaluators were not informed which condition each log corresponded to.

5 Results

5.1 Effect on target information collection

Table 2 shows the results of the information collection rates in the interactive simulation, as well as the test results against the baseline and effect sizes for (B), (C), and (D).

Among the four conditions, Condition (D)+All achieved the highest collection rate ($p_{(D-A)} = 0.002 < 0.05$) , whereas Condition (A)Baseline yielded the lowest. Both (B)+Approach and (C)+Selection also outperformed the baseline ($p_{(B-A)} = 0.009 < 0.05$), ($p_{(C-A)} = 0.010 < 0.05$). These results suggest that each module constituting the proposed method contributes to improving the efficiency of information collection, and that combining the two modules leads to further gains.

The variability of collection rates across sessions was largest under (A)Baseline and smallest under (D)+All, with (C)+Selection showing the second-smallest variability. Because larger standard deviation indicates greater fluctuation in information collection performance across dialogue sessions, these results suggest that the proposed method not only improves target-information collection rates but also stabilizes performance across sessions. In particular, the question-selection module appears to contribute substantially to improving cross-session stability.

5.2 Analysis of the collected information

As shown in Table3, the collected slots were concentrated in self-disclosure Levels 1 and 2. In contrast, Levels 3 and 4 were not collected under any condition. One possible reason is that the proposed method selects target slots to approach based on similarity to the current topic, which may have caused the system to preferentially approach target slots that were already semantically close to the current topic.

5.3 Effect on dialogue naturalness

Table 4 shows the median scores, median differences, and test results for the dialogue-naturalness evaluation comparing Condition (D)+All and Condition (A)Baseline. No statistically significant differences from the baseline were observed for any of the dialogue-naturalness measures. These results suggest that, at least in the present simulation setting, we did not observe a statistically significant negative effect of the proposed method on dialogue naturalness.

6 Conclusion

In this paper, to realize dialogue that both promotes self-disclosure and enables the system to collect necessary information, we prototyped an interview dialogue system that manages topics using a Question Tree and dynamically generates questions on that tree, and evaluated its question-guiding function through LLM-based simulated dialogue experiments in text chat. In the Question-Tree, question items are arranged as nodes in a tree-structured graph based on parent-child relations among topics. Based on the user's response, the system generates two types of question candidates: deepening questions that further probe the current topic, and approach questions that guide the dialogue toward the target topic. It then selects as the next question the candidate that has high cosine similarity in sentence-embedding space to both the target topic and the current topic.

We conducted dialogue simulation experiments using a text-chat-based mock dialogue system and an LLM-based dummy interviewee. The results showed that even when topic continuation and topic shift were not predictable in advance by the system but were externally imposed, the proposed method achieved a higher target-information collection rate than the baseline, collecting on average 26.6% of the 24 pre-specified target-information items within 10 turns.

Future work includes evaluating the effect of the proposed method on information-collection performance in interview dialogues with real human participants, as well as investigating how the method affects dialogue naturalness and users' impressions of the system.

7 Limitations

This section discusses the interpretation of the results and their limitations.

Table 2: Target Information Collection Rate per Dialogue Strategy (End of Turn 10)

Condition	**collection-rate**			
	Avg.(%)	**SD**	$p_{(X)-(A)}$	$r_{(X)-(A)}$
(A)Baseline	10.8	8.40	–	–
(B)+Approach	24.3	3.42	0.009	+0.72
(C)+Selection	20.5	2.99	0.010	+0.66
(D)+All	26.6	2.02	0.002	+0.75

Table 3: Average number of responses collected per target slot by level of self-disclosure

Condition	Average filled slots count Self-disclosure level			
	Lv1	Lv2	Lv3	Lv4
(A)baseline	2.59	0.00	0.00	0.00
(B)+Approach	4.21	0.30	0.00	0.00
(C)+Selection	4.90	0.03	0.00	0.00
(D)+All	6.00	0.39	0.00	0.00

7.1 Limitation of simulated experiments

The results reported in this paper were obtained through LLM-based dialogue simulations. We adopted this setting in order to secure a sufficient number of trials for exploring appropriate parameter combinations, to compare conditions under matched topical starting points, and to minimize the influence of factors other than dialogue content, such as real-time response latency. By adopting a random continuation/shift scenario that assumes user affect is unpredictable, we evaluated the quality of question-topic selection by the system in isolation. Applying the insights obtained from these simulations to real human participants and evaluating their effects in actual interaction remains future work.

7.2 Target information and Self-disclosure

The results in Section 5.2 show that all the slots for which responses were obtained in this simulation involved low levels of self-disclosure; there were no instances of high self-disclosure (Levels 3 and 4). Possible reasons for this include the fact that the number of dialogue turns was too small, making it easy for the conversation to remain confined to topics involving shallow self-disclosure, and that semantic similarity-based methods made it difficult to reach deeper levels of self-disclosure. This may suggest that shifting to topics requiring deep self-disclosure tends to result in a decrease in cosine similarity. Further investigation is needed. This suggests a limitation of the proposed method: if topics are selected based solely on cosine similarity, the system cannot move on to questions requiring a high level of self-disclosure until it has exhausted all topics requiring a low level of self-disclosure. A future research question is whether it is possible to successfully transition to deeper topics of self-disclosure when combining topic adaptation with actual interactions, or whether improvements to topic guidance that take depth into account are necessary.

In this study, we established target criteria linked to self-disclosure scales to verify their effectiveness across a wide range of topics; however, it should be noted that these criteria are somewhat detached from real-world interview settings. To conduct detailed future research on simultaneously achieving topic guidance and self-disclosure, we should prepare specific application scenarios and conduct experiments using target criteria that more closely reflect actual conditions.

7.3 Caution regarding dialogue naturalness

The results in Section 5.3 suggest that we did not observe a significant negative impact of the proposed method on dialogue naturalness in the present simulations. However, this should not be interpreted as evidence that the two conditions are equivalent. In this paper, due to constraints related to securing annotators, the evaluation was limited to only a very small portion of the entire simulation log. To demonstrate the absence of harmful effects more rigorously, future work should either detect significantly positive effects relative to the baseline or collect more realistic data through dialogue experiments with human participants, recruit a sufficient number of evaluators, and conduct an explicit equivalence test (Stanton, 2021).

7.4 Privacy risks

As pointed out by Nagasawa and Okada (Nagasawa and Okada, 2025), dialogue systems that promote

Table 4: Evaluation results of the naturalness of the dialogue between (D) and (A)
(↑: larger is better, ↓: smaller is better. When the median fell between two adjacent scale points, the midpoint is reported.)

	Question	Score(Median) (A)Baseline	(D)+All	(D)-(A)	$p_{(D)-(A)}$
F1	Appropriate responses ↑	5.0	5.5	0.50	0.068
F2	Maintenance of context ↑	5.0	6.0	-0.25	0.715
F3	Absence of contradictions ↑	5.5	5.5	0.00	0.068
F4	Continuity between topics ↑	5.5	5.5	0.00	0.715
F5	Natural topics transition ↑	5.5	5.5	0.25	0.465
F6	Unnatural topics transitions ↓	0.0	0.0	0.00	0.584
F7	Consistency in conversation ↓	0.0	0.5	0.25	0.068
F8	Smoothness of conversation ↓	0.0	0.5	0.25	0.068
F9	Appropriate responses to topics ↑	5.5	5.5	0.25	0.465
F10	Providing new topics ↑	4.0	2.5	-0.75	0.465

self-disclosure are inherently tied to the risk of eliciting private information, and therefore require transparency and safety in their design. Compared with approaches in which the entire topic-guidance process is generated by an LLM, our method is less flexible because it relies on simpler logic. However, it has the advantage that the topic being targeted at each utterance-generation step can be explicitly tracked by the algorithm, which improves transparency. Moreover, because our approach uses vector-space similarity to steer the dialogue toward target topics, it could also be extended to steer the dialogue away from topics that should not be approached. Such an extension may make it possible to avoid psychologically traumatic topics or privacy-sensitive topics while still enabling topic deepening. How to maximize the benefits of combining explicit control logic such as ours with the flexibility of LLMs, and how to evaluate such hybrid systems, remains future work.

7.5 Language and Models

In this study, we developed a Japanese-language dialogue system and employed a self-disclosure scale tailored specifically for Japanese. When conducting dialogue experiments with actual human participants, it is necessary to be mindful of the impact of language and culture on the results. Furthermore, since topic induction depends on sentence vectors, the results may also be influenced by the embedding model. Conducting further research that accounts for differences arising from the models used remains a task for future studies.

7.6 Computational cost and processing time

In the present study, the system generated multiple question candidates and then filtered them to determine the next question. This design incurs a large number of LLM calls per dialogue turn and therefore substantial computational overhead. One reason for using a local LLM in this study was that the required number of calls would otherwise exceed practical API usage limits. For practical deployment, it will be necessary to improve efficiency by generating more targeted questions and reducing wasteful generation. Our analysis of processing time suggested that the most computationally expensive step was the generation of a large number of question candidates(For details, see Table 6 in the Appendix.). Because the current method first stockpiles many candidates and only later selects among them, many generated questions are never used, and there is no guarantee that the ideal guiding question is always included among them. Methods such as Vec2Text (Morris et al., 2023), which generate text from a target embedding vector, may help address this issue. Integrating such methods is left for future work.

References

Irwin Altman. 1973. Social penetration: The development of interpersonal relationships. *Rinehart, & Winston*.

Alina Elena Apostol, Kellie Turner, Rosa Hoshi, and Aimee Pudduck. 2025. Contributory factors to self-disclosure in clinical supervision: A meta-ethnography. *Clinical Psychology & Psychotherapy*, 32(2):e70068. E70068 CPP-4424.R3.

Chloé Clavel, Matthieu Labeau, and Justine Cassell. 2022. Socio-conversational systems: Three challenges at the crossroads of fields. *Frontiers in Robotics and AI*, 9.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ben Emans. 2016. *Interviewing: Theory, techniques and training*. Routledge, London.

Andrea Fontana and James H Frey. 2005. The interview. *The Sage handbook of qualitative research*, 3(1):695–727.

Chuanruo Fu and Yuncheng Du. 2025. First ask then answer: A framework design for ai dialogue based on supplementary questioning with large language models. *arXiv preprint arXiv:2508.08308*.

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 *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1562–1584. Association for Computational Linguistics.

Ryuichiro Higashinaka, Tetsuro Takahashi, Sota Horiuchi, Michimasa Inaba, Shiki Sato, Kotaro Funakoshi, Masato Komuro, Hiroyuki Nishikawa, Mayumi Usami, Takashi Minato, Kurima Sakai, and Tomo Funayama. 2022. Dialog system live competition 5. *JSAI SIG-SLUD*, 96:19.

Allison Hurst. 2023. Interviewing. *Introduction to Qualitative Research Methods*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Seiya Mitsuno, Midori Ban, Hiroshi Ishiguro, and Yuichiro Yoshikawa. 2024. Deepening conversations over time: A chatbot with a topic depth estimation model for gradually engaging in deeper chats. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pages 1354–1361.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

Fuminori Nagasawa and Shogo Okada. 2025. *Can Adaptive Interviewer Robot Based on Social Signals Make a Better Impression on Interviewees and Encourage Self-Disclosure?*, page 35–43. Association for Computing Machinery, New York, NY, USA.

Sora Niwa and Shun'ichi MARUNO. 2010. Development of a scale to assess the depth of self-disclosure. *Japanese Journal of Personality*, 18(3).

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Hashai Papneja and Nikhil Yadav. 2024. Self-disclosure to conversational ai: A literature review, emergent framework, and directions for future research. *Personal and ubiquitous computing*, pages 1–33.

Jeffrey M Stanton. 2021. Evaluating equivalence and confirming the null in the organizational sciences. *Organizational Research Methods*, 24(3):491–512.

Shohei Tanaka, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Arta: Collection and classification of ambiguous requests and thoughtful actions. *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Yuki Zenimoto, Mariko Yoshida, Ryo Hori, Mayu Urada, Aiko Inoue, Takahiro Hayahsi, and Ryuichiro Higashinaka. 2025. Development of a question-guided survey dialogue system. *Proceedings of the 31st Annual Conference of the Association for Natural Language Processing*.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–10, Los Angeles. Association for Computational Linguistics.

Appendix

Table 5: Questions presented to evaluators as part of the naturalness of dialogue assessment (translated; original in Japanese)

Key	Items	Question	reverse-scored
F1	Appropriate responses	Did the system respond appropriately to the user's most recent utterance?	No
F2	Maintenance of context	Did the system's utterances maintain a natural context throughout the entire dialogue?	No
F3	Absence of contradictions	Did the system handle information mentioned during the conversation (such as conditions or proper nouns) without contradiction?	No
F4	Continuity between topics	When changing the topic, did the system provide a smooth transition to maintain the context?	No
F5	Natural transitions between topics	Were topic shifts appropriate in light of the user's utterances and the flow of the conversation?	No
F6	Unnatural transitions between topics	Did topics change suddenly or abruptly, making the conversation feel unnatural?	No
F7	Consistency in conversation	Were many of the system's utterances unrelated to or out of sync with the flow of the conversation?	No
F8	Smoothness of conversation	Did the conversation fall into repetition or dead ends?	No
F9	Appropriate responses to topics	Was the system able to respond appropriately to topics raised by the user?	No
F10	Providing new topics	Was the system able to provide new information regarding topics selected by the user?	No

Table 6: Measurement results for computation time per module

Module	Processing time(milliseconds)		
	min	average	max
Collecting Answer	960.0	1384.3	2100.0
Question Candidates Generate	150.0	691.4	1150.0
Sentence Embedding	2180.0	2431.4	2730.0
Next topic selection	1.0	2.0	2.5
System utterance Generation	230.0	272.9	340.0
Total	3521.0	4782.0	6322.5

Table 7: Details of the 24 target informations used in the experiment in Section 4(translated; original in Japanese)

Description of topic	Self-disclosure level
The user's favorite things (music, movies, fashion, etc.)	1
How the user spends their days off	1
Recent enjoyable events in the user's life	1
Things the user has been obsessed with lately	1
The user's hobbies	1
Events the user is looking forward to	1
Hobbies the user would like to try in the future	1
Times when the user received help from someone during a difficult experience	2
Efforts the user has made to overcome difficult situations	2
How the user has overcome painful experiences	2
How past painful experiences are helping the user now	2
Things the user feels are a bit of a shortcoming (being late, etc.)	3
Minor flaws the user feels they need to fix but can't seem to change (being late, etc.)	3
Times when the user feels down, even over what might be minor flaws (being late, etc.)	3
Times when the user thought, "I'm a bit of a failure," based on a specific experience (being late, etc.)	3
Times when others expressed concern about the user's minor flaws (being late, etc.)	3
Things the user worries about daily regarding their minor flaws	3
Aspects of the user's personality they really dislike (being unable to sincerely celebrate others' success, etc.)	4
An incident where a very unpleasant aspect of the user's personality came out	4
Things the user is deeply worried about regarding their abilities	4
An experience where the user failed to achieve a goal due to a lack of ability	4
Areas where the user feels inferior due to their abilities	4
An experience where the user felt limited by their abilities and became disappointed	4
An experience where the user deeply hurt someone because of their own actions	4

common_persona

You are an interviewer. You are interested in what the user interested in. Through conversation, you want to understand the user's feelings and inner thoughts. You are careful not to be rude to the user.

Figure 3: Prompt (1/7) "$common\_persona$" : common prompt for persona setting(translated; original in Japanese)

prompt_collect_answer

Based on the slot information and the dialogue log, collect the information that has been provided in response to the target slot.

Slot information is provided in the following format.

Slot key: Slot description

The dialogue log is provided in the following format.

user: User's utterance

system: Agent's utterance

- If there is information in the dialogue log that corresponds to the content of the target slot, output that information.
- Output the results in JSON format.
- Please output only information directly mentioned in the dialogue log as having been answered.
- Do not output information regarding slots not mentioned in the dialogue log as having been answered.
- Even if mentioned in the dialogue log, do not output information deemed to be only loosely related to the target slot's content as having been answered.
- The result must be accurate; do not output information based on guesswork or inference.
- Use the target slot key as the JSON key.
- Output only JSON.

Slots Information:

<all_slots_descriptions>

Dialog Log:

<dialog_log>

Output Format:

{{"slot_key": "Answered Information", ...}}

Result:

Figure 4: Prompt (2/7) "$prompt\_collect\_answer$" : Prompt used for Collecting Answer module(translated; original in Japanese)

prompt_slogtgen_common

- Please output the data in JSON format, with all items organized into a single dictionary, such as {{“what”:“What?”, ‘when’:“Time”...}}. Only single-byte alphanumeric characters are allowed as JSON keys.
- Symbols are not permitted.
- Do not include a hierarchical structure in the items.
- Ensure that each item can be answered concisely.
- Keep slot descriptions short and to the point.
- Please create slots that are not present in previous versions.
- New slots should only be those that can be directly associated with user utterances.
- Focus on words such as nouns and verbs contained in user utterances, and generate questions related to those words.
- If the generated question sounds unnatural, please skip it rather than forcing a question.
- Ensure that each new slot corresponds one-to-one with a question category.

Figure 5: Prompt (3/7) “$prompt\_slogtgen\_common$” : commom prompt for question candidate generation(translated; original in Japanese)

elaborating_categories

“Experience/Action”: “Specify what was done or what was experienced.”,
“Opinions/Values”: “The other person's way of thinking, values, and what they think”
“Emotions”: “Emotional reactions and feelings”
‘Knowledge’: “Factual knowledge possessed”
“Sensory Descriptions”: “Memories involving the five senses (sight, hearing, smell, touch, taste)”
“Background Information”:“Attributes and background”
“Who”:“Who are the people involved?”
“When”:“The time of the event”
“Where”:“The location involved”
“What”:“What was done or what happened?”
“Why”:“Motives or reasons”
‘How’:“Specific methods or circumstances”

Figure 6: Prompt (4/7) “$elaborating\_categories$” : Common prompt for generating question candidates. Here, we define categories for follow-up questions.(translated; original in Japanese)

prompt_slotgen_deepning

<common_persona>

Based on the following user utterance and slot states, and in accordance with the question categories below, please list the new questions to ask and their explanations in order to gain a deeper understanding of the user's description.

<prompt_slotgen_common>

Question Categories and Descriptions:

<elaborating_categories>

Current Slots: Descriptions

- slot1: description of slot1

- slot2: description of slot2

- slot3: description of slot3

⋮

Previous slots: Information obtained

- slotA : collected answer to slotA

- slotB : collected answer to slotB

- slotC : collected answer to slotC

⋮

Output Format:

{"slot_name": {"description":"Question Content",'focus_topic':"Word Focused On During Question Generation"}}, ...}}

Question from the previous system:

"*<last system utterance>*"

User utterance:

"*<last user utterance>*"

Figure 7: Prompt (5/7) "*prompt_slotgen_deepening*" : Prompt for question candidate generation (Deepening question). The gray parts are where corresponding variables and other data are stored at the time of generation.(translated; original in Japanese)

prompt_slotgen_approaching

<common_persona>

Based on the following user utterance and slot states, and in accordance with the question categories below, please list the new questions to ask and their explanations in order to gain a deeper understanding of the user's description.

<prompt_slotgen_common>

When doing so, ensure that the topics are related to the target information.

Target slots: Descriptions

<target_slot_name : target_slot_description>

Question Categories and Descriptions:

<elaborating_categories>

Current Slots: Descriptions

- slot1: description of slot1

- slot2: description of slot2

- slot3: description of slot3

Previous slots: Information obtained

- slotA : collected answer to slotA

- slotB : collected answer to slotB

- slotC : collected answer to slotC

Output Format:

{"slot_name": {"description":"Question Content",'focus_topic':"Word Focused On During Question Generation", "aimed_target":" The target slot associated with this slot "}}, ...}}

Question from the previous system:

"*<last system utterance>*"

User utterance:

"*<last user utterance>*"

Figure 8: Prompt (6/7) "*prompt_slotgen_approaching*" : Prompt for question candidate generation(Approaching question). The text in red shows the differences from 'prompt_slotgen_deepening' (translated; original in Japanese)

prompt_utterance_generation

<common_persona>

Based on the previous conversation and the content of the slot you are asking about this time, please generate the text for the next question to ask the user.

Keep the question concise, in a single sentence.

Ensure the question is relevant to the content of the slot you are asking about this time.

Base the question on the content of the user's most recent utterance.

Formulate the question in a way that fits the natural flow of the conversation and is easy for the user to answer.

Generate the question in written language.

conversation log:

<dialog_log>

Content of the slot to be asked about:

<next_topic_to_question>

Figure 9: Prompt (7/7) "$prompt\_utterance\_generation$" : Prompt for system utterance generation(translated; original in Japanese)

prompt_dummy_interviewee

As the interviewee, please answer the interviewer's questions.

Based on the constraints below and the input text, please generate natural responses to the interview questions.

Please provide concise, natural answers, just like in everyday conversation.

Please output only the response text; do not include the question or any other explanations.

Constraints:

- Your responses must not contradict the dialogue.
- Responses must be in Japanese.

- Responses must address only one topic.
- Responses should be concise, consisting of approximately one sentence.
- Output only the content of your response. Do not include the question or any other explanations.
- Respond to the interviewer's most recent question. Do not provide a response covering the entire conversation history.
- The episode you are responding to should be realistic as of March 2026.
- Answer in conversational language; do not use written language expressions such as hashtags.

History of previous interview questions and answers:

<dialog_history>

Latest question from the interviewer:

<latest_question>

Your answer:

Figure 10: Prompt for dummy interviewee(translated; original in Japanese)

Table 8: Example of dialog log (case of (D)+All)(translated; original in Japanese)

Turn index	speaker	text	topic-control
0	system	Hello! What has been going on with you recently?	-
	user	Hello. I've been working from home more often lately, so I've been spending a lot more time at home.	-
1	system	So you've been working from home more? Have you gotten used to working at home yet?	continue
	user	I've gotten used to it quite a bit, but it's a little hard to stay focused.	-
2	system	It is definitely hard to stay focused, isn't it? What do you do to clear your head?	switch
	user	I go for short walks or brew some coffee to take a break.	-
3	system	A walk and coffee are great ways to refresh. Have you picked up any new hobbies at home?	switch
	user	I recently started growing houseplants. I feel like my room has gotten a little brighter.	-
4	system	Houseplants are lovely. What kinds of plants are you growing?	continue
	user	I'm growing pothos, snake plants, and a few small succulents.	-
5	system	That's an easy-to-care-for selection. Is there anything you find difficult about caring for them?	continue
	user	I still don't quite get the timing for watering, so I tend to overwater them.	-
6	system	Getting the watering right is a common struggle, isn't it? What made you start being careful not to overwater them?	continue
	user	Actually, I've killed a plant before by overwatering it.	-
7	system	So you've had that experience of failing due to overwatering. Did anyone help you out back then?	continue
	user	I asked a friend who's into gardening, and they taught me a lot.	-
8	system	I see. Have you started anything else recently?	switch
	user	I started taking online English conversation lessons. I take them about twice a week.	-
9	system	English conversation lessons sound great. Have you started any other new habits?	switch
	user	Since I've been working from home more, I started working out.	-
10	system	Starting a workout routine is a healthy choice. How's your health been lately?	switch
	user	I'm feeling healthy. Thanks to the workouts, I feel like I've gained a little muscle.	-