Question Types for Knowledge Acquisition in LLM-based Dialogue Systems: Experiments with Simulated and Human Users

Kazunori Komatani[1], Ryu Takeda[1] and Mikio Nakano[1,2]

[1]SANKEN, University of Osaka, Ibaraki, Osaka, Japan
[2]C4A Research Institute, Inc., Setagaya, Tokyo, Japan
{komatani@, rtakeda@, mikio.nakano@ei.}sanken.osaka-u.ac.jp

Abstract

To acquire knowledge from users through dialogue, systems must decide not only what to ask but also how to ask it. Since users may not always be willing to answer such questions, question formulation affects both user experience and the information obtained. Prior work has examined question types in controlled, template-based settings, but it remains unclear whether similar effects are observed in LLM-based dialogues. We investigated the effects of question type on knowledge acquisition through experiments with both an LLM-based user simulator and crowdsourced human participants. We compared three question types: implicit questions, explicit questions, and wh-questions. The results showed no substantial differences in user annoyance among the question types. In contrast, the question types differed in how well they elicited correct responses: wh-questions were less effective, whereas implicit and explicit questions performed comparably. The findings suggest that candidate-guided question forms are useful when the system has a plausible candidate answer, whereas wh-questions may be appropriate when the system lacks sufficient confidence to ask a more specific question.

1 Introduction

Dialogue systems sometimes need to handle information that is not readily available in advance, such as unknown terms (Waki et al., 2025) or user-specific knowledge. To operate robustly in such situations, systems need to acquire knowledge through interaction with users. This is especially important for interactive systems used in everyday environments, where it is unrealistic to prepare all necessary knowledge in advance. Under such conditions, the system must decide not only what to ask but also how to ask it, because question formulation affects users' impressions as well as the information obtained through dialogue.

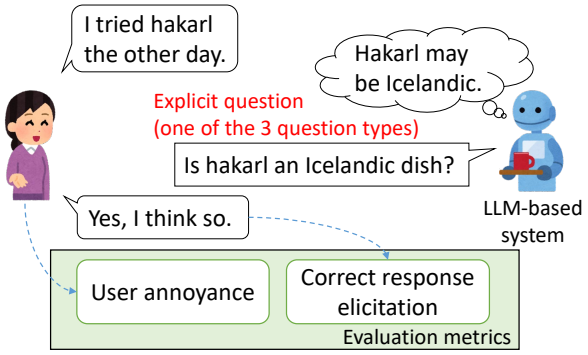

Figure 1: Example of one question type examined in this study and the two evaluation metrics.

Our previous work examined how question types affect users' impressions in knowledge acquisition dialogue (Komatani et al., 2022). However, that study was conducted before the large language model (LLM) era, with predefined system utterances and fixed dialogue flows, and may therefore not fully reflect recent dialogue systems that generate utterances dynamically based on context. More broadly, Finch and Choi (2020) have shown that dialogue system evaluation can depend substantially on the evaluation protocol itself, including whether systems are assessed in static settings or through interactive dialogue. Taken together, these points make it necessary to re-examine how question types affect user impressions and the information obtained in more dynamic LLM-based dialogue settings.

In this study, we re-examine the effects of question types in LLM-based dialogue systems. More specifically, we ask two research questions: (1) how question type affects user annoyance in LLM-based knowledge acquisition dialogue, and (2) how question type affects the proportion of correct responses elicited in such dialogue. We build an LLM-based dialogue system and compare three question types: implicit questions, explicit questions, and wh-questions. We first conduct a prelim-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 602–613
August 2–5, 2026.

inary analysis using an LLM-based user simulator. We then conduct a human evaluation with crowdsourced participants to examine whether the observed tendencies also hold in human interactions. As evaluation measures, we use user annoyance and the proportion of correct responses. Figure 1 illustrates the task of this study using one example question type and highlights the two evaluation metrics: user annoyance and correct response elicitation.

The contributions of this paper are as follows:

- We re-examined the effects of three question types, implicit questions, explicit questions, and wh-questions, for knowledge acquisition in LLM-based dialogue systems.
- We investigated question-type tendencies through both a preliminary analysis with an LLM-based user simulator and a human evaluation with crowdsourced participants.
- We showed that, in LLM-based dialogue, differences in user annoyance among question types were small, whereas clearer differences were observed in correct response elicitation: implicit and explicit questions performed comparably, while wh-questions elicited correct responses less often.

2 Related Work

2.1 Knowledge Acquisition through Dialogue

Knowledge acquisition through dialogue has been studied as a way to address practical challenges such as incomplete knowledge bases and unknown or out-of-vocabulary terms (Purver, 2002; van Schagen and Knott, 2004; Otsuka et al., 2013). More broadly, this line of work is related to *lifelong learning* (Chen and Liu, 2018), in which systems continually acquire, update, and revise knowledge over time. In goal-oriented dialogue, systems have been designed to strategically ask users questions in order to acquire task-relevant knowledge (Pappu and Rudnicky, 2014).

Related work has also explored dialogue-based knowledge acquisition in other settings. Hixon et al. (2015) proposed a method for acquiring relations between concepts through interaction in a question-answering system. Weston (2016) and Li et al. (2017) studied learning through dialogue in which asking questions itself supports learning. Mazumder et al. (2019) proposed a dialogue system that asks questions about triples for lifelong and interactive learning of factual knowledge, and subsequent work has further considered revising acquired knowledge and improving question selection in interactive settings (Liu and Mazumder, 2021). Recent work has also formulated dialogue-based knowledge acquisition as a sequential decision problem. For example, it has been cast as stream-based active learning, where the system learns to ask efficiently while reducing the number of questions posed to the user (Waki et al., 2025). In a related direction, Furuta et al. (2026) addressed the reduction of unnecessary confirmation questions caused by speech recognition errors.

These prior studies on dialogue-based knowledge acquisition have mainly focused on what knowledge to acquire, when to ask, and how to select questions efficiently. In contrast, our focus is on how a system should formulate a question once the target information has been determined, and how such formulations affect users in interaction.

2.2 Clarification Questions, Question Types, and User Impressions

When a dialogue system lacks necessary information, it can ask additional questions to obtain it through interaction. Such questions have been widely studied as clarification questions in dialogue systems (Aliannejadi et al., 2021; Rao and Daumé III, 2018). In information-seeking and task-oriented settings, they are used to resolve ambiguity in user requests and to improve the efficiency of information acquisition. Recent work has also examined clarification questions in open-domain dialogue, showing that asking an appropriate follow-up question can improve response quality and that the quality of such questions should itself be evaluated systematically (Aliannejadi et al., 2021). Prior work has further investigated dialogue systems that sequentially ask users for missing information in order to support decision making or response generation while reducing user burden (Rao and Daumé III, 2018).

In our setting, knowledge acquisition is realized as a clarification sub-dialogue triggered by an unknown term mentioned by the user. More specifically, the system asks a follow-up question in order to obtain information about that term through interaction. In this type of knowledge acquisition dialogue, it is important not only what information to ask for but also how to ask for it. Because the system needs users to provide information through interaction, the form of a question can directly af-

fect users' impressions.

This point is also consistent with recent findings that the formulation of clarification questions influences user satisfaction, and that poorly formulated questions can lead to confusion or frustration (Rahmani et al., 2024). Our previous work examined user impressions of different question types for lexical knowledge acquisition during dialogue, including implicit questions, explicit questions, and wh-questions (Komatani et al., 2022). However, that study was mainly conducted with fixed dialogue flows and predefined utterances, and it remains unclear whether similar effects are observed in more dynamic LLM-based dialogue.

2.3 Evaluation Gaps and User Simulation in LLM-based Dialogue Systems

Recent dialogue systems, especially those based on large language models (LLMs), can exhibit substantial variation across interactions because utterances are generated dynamically based on context. At the same time, prior work has pointed out that dialogue system evaluation depends strongly on the evaluation protocol itself, such as whether the system is assessed in static settings or through interactive dialogue (Finch and Choi, 2020). This suggests that findings obtained in controlled dialogue settings should be re-examined in more natural LLM-based dialogue environments, where simulator-based analysis alone may not be sufficient.

Recent work has also explored the use of LLM-based user simulators for evaluating dialogue systems and simulating user interactions (Sekulic et al., 2024; Algherairy and Ahmed, 2025), and recent survey work has organized this growing area under a unified framework (Ni et al., 2026). However, such simulators are still only an approximation of real users, and their behavior can depend strongly on prompting and simulation design (Sekulic et al., 2024; Algherairy and Ahmed, 2025). This motivates combining simulator-based preliminary analysis with human evaluation, rather than relying on simulator results alone, when examining question-type effects in LLM-based dialogue.

3 Question Types

We compare three types of system questions used for knowledge acquisition through dialogue: implicit questions, explicit questions, and wh-questions. We include implicit questions in the comparison because they have been studied in our previous work on dialogue-based lexical acquisition (Komatani et al., 2022) and provide a plausible way for a system to present a candidate hypothesis without making the question fully explicit. In particular, they allow the system to present candidate information while keeping the interaction relatively indirect.

Question type	Example follow-up question
Implicit	Icelandic food is distinctive.
Explicit	Is hakarl an Icelandic dish?
Wh	What country is hakarl from?

Table 1: Example follow-up questions after the user has mentioned hakarl, for acquiring the target information that hakarl is an Icelandic dish.

We focus on differences in how the questions are phrased while keeping the information to be elicited comparable across question types. Table 1 shows example follow-up questions for each type. In our system, follow-up questions of these types were generated by the LLM while taking the dialogue history and other contextual information into account, rather than selected from fixed templates.

3.1 Implicit question

An implicit question presents candidate information without asking about it explicitly. If the user's next response indicates acceptance, the presented information can be regarded as correct; otherwise, it may be understood as incorrect (Ono et al., 2017). In our setting, this type includes an utterance containing the correct country name for the target dish.

3.2 Explicit question

An explicit question presents candidate information explicitly and asks the user in a yes/no form whether it is correct. The user's response is therefore expected to directly indicate whether the presented information is correct or incorrect. In our setting, this type includes an utterance containing the target dish together with its correct country name.

3.3 Wh-question

A wh-question asks for the relevant information directly by using an interrogative expression, such as asking which country the target dish is from. The user's response is therefore expected to provide the requested information. In our setting, this type includes an utterance containing the target dish without the correct country name.

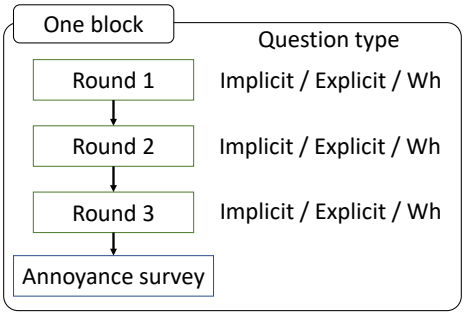

Figure 2: Overview of one dialogue block in the task.

4 Experimental Dialogue Task

The dialogue task in our experiments involved a system obtaining country information about an unknown food term through conversation with a user. More specifically, when a user mentioned an unknown food term during casual conversation, the system asked a follow-up question using one of the question types defined in Section 3.

To examine the effect of question phrasing, we compared the three question types while keeping the target information to be elicited comparable across question types. In this experimental setting, we kept the candidate country information in implicit and explicit questions correct across conditions so that differences in candidate correctness would not affect the comparison.

4.1 Task Overview

Each participant interacted with the system through repeated blocks, and Figure 2 illustrates the structure of a single dialogue block. In each block, the participant engaged in up to three rounds of interaction, each involving one unknown food term. In each round, the system could ask a follow-up question using one of the three question types: implicit, explicit, or wh, unless the relevant country information had already been provided. After completing a block, the participant rated the annoyance of the system's questions in that block.

4.2 Dialogue Flow

The flow of each round was as follows.

1. The user was given an unknown food term and instructed to mention it during the conversation.

2. When the user mentioned the term, the system asked a follow-up question about it using one of the three question types, unless the relevant country information had already appeared in the dialogue before the follow-up question was generated.

3. If the system asked a follow-up question, the user responded naturally to it.

This round-level flow was repeated within each dialogue block, after which the user provided a block-level annoyance rating.

4.3 System

We implemented a Japanese dialogue system on DialBB (Nakano and Komatani, 2024) with state-transition network (STN)-based dialogue management. The system engaged in casual conversation using an LLM until an unknown food term appeared in the user's utterance. When such a term appeared, the dialogue manager moved to a questioning state, where the system asked a follow-up question using one of the three question types.

After the user responded, the system interpreted the response and returned to the casual conversation state. The dialogue flow itself was controlled by the STN-based dialogue manager, while the wording of both the casual utterances and the follow-up questions was generated by the LLM. Details of the prompt design used to generate follow-up questions are provided in Appendix B. The user's response was interpreted by the LLM-based language understanding module in DialBB and stored for later use in determining whether the response contained the correct information.

5 Experimental Setup

This section describes the experimental setup for both the LLM-based user simulation and the human evaluation. We first describe the elements shared across the two settings and then describe the settings specific to each condition.

In both settings, the dialogue task was organized into blocks, each designed to include up to three question rounds followed by an annoyance survey. We also used a shared set of 30 unknown food terms. Each term was paired with its corresponding country information, which served as the target information to be acquired through dialogue. Items in each session were randomly sampled from this shared pool.

5.1 Evaluation Metrics

To evaluate the effect of question type, we used two measures in both the LLM-based user simulation

and the human evaluation.

The first measure concerned user annoyance toward the system's questions. Following our previous work (Komatani et al., 2022), annoyance was rated once after each dialogue block rather than after each individual question. This design was intended to reduce the burden of repeated rating during the dialogue. We analyzed the collected annoyance ratings using a regression model and used the resulting weights as a summary of the effect of question type on perceived annoyance:

$$annoyance = w_0 + w_1 Impl + w_2 Expl + w_3 Whq.$$

Here, *Impl*, *Expl*, and *Whq* denote the numbers of implicit, explicit, and wh-questions included in the block, respectively. The prior study compared five question types by additionally distinguishing whether the content of implicit and explicit questions was correct or wrong. By contrast, because the country information was fixed as correct in our setting, we compared only three question types.

The second measure compared question types in terms of how often users correctly answered the system's questions. For each system question, we judged whether the user's response correctly answered the question and computed the proportion of correct responses for each question type. Because the country information was fixed to be correct when provided, a response was counted as correct if it expressed acceptance for implicit questions, gave an affirmative response for explicit questions, or provided the correct country name for wh-questions. This judgment was performed using the LLM-based language understanding module in DialBB.

To assess the reliability of the automatic judgments, we manually inspected a sample of 100 user responses, stratified by question type and data source (simulation or human evaluation). The manual judgments agreed with the automatic judgments in 96 of the 100 cases. The disagreements mainly involved ambiguous, affective, or question-like responses whose correctness was difficult to determine.

5.2 User Instructions

In both settings, users were instructed to engage in food-related small talk with the system. During the dialogue, task instructions were presented in brackets so that they could be distinguished from the system's ordinary dialogue utterances. In each round, the user was given an unknown food term and instructed to mention it in the conversation. Each unknown food term was also presented with a corresponding Wikipedia URL so that the user could look up the term if needed. Users were also instructed to respond naturally to the system's follow-up questions. They were told that they did not need to carry over the content or impression of the previous block.

After each block, users were asked to rate how annoying the system's questions were on a 1–5 scale. The detailed instruction is provided in Appendix A, and the same wording was used in both the LLM-based user simulation and the human evaluation to keep the rating condition consistent across the two settings.

5.3 LLM-based User Simulation Setup

Before the human evaluation, we conducted a preliminary study using an LLM-based user simulator. We used the user simulator provided with DialBB (Nakano and Komatani, 2024) with gpt-4o-2024-11-20 as the underlying LLM. The simulator was prompted to behave as a user engaging in casual food-related small talk and to respond to the system's follow-up questions. To make the simulation setting comparable to the human evaluation, the prompt was adjusted to align with the instructions given to the crowdsourced users. The simulator was given the same Wikipedia URLs as in the human evaluation to keep the instructions parallel, but no external browsing tool was used. The prompt used for the simulator is provided in Appendix C.

In total, we generated 100 simulated dialogue sessions. Each simulated session consisted of 10 dialogue blocks, resulting in 1,000 dialogue blocks in total. Although each block was designed to contain three rounds, some rounds did not require a follow-up question because the user had already mentioned the relevant country. As a result, the actual number of follow-up questions was smaller than the total number of rounds. An example dialogue with the LLM-based user simulator is shown in Figure 6 in Appendix D.

5.4 Human Evaluation Setup

We conducted a human evaluation via crowdsourcing on CrowdWorks[1] to examine whether the same question-type tendencies were observed with human users. Participants were instructed to follow

[1] https://crowdworks.jp/

Rating	1	2	3	4	5	Total
Count	297	683	18	2	0	1000

Table 2: Distribution of annoyance ratings in the LLM-based user simulation.

Implicit	Explicit	Wh	Intercept	R^2
0.121	0.132	0.131	1.363	0.015

Table 3: Regression weights for annoyance in the LLM-based user simulation.

Question type	Correct / Total	Proportion
Implicit	989 / 990	0.999
Explicit	901 / 908	0.992
Wh	842 / 894	0.942

Table 4: Correct response elicitation by question type in the LLM-based user simulation.

experimental instructions shown in brackets, avoid entering personal information, and answer the system's questions when asked. They were also informed that they could stop the task at any time. Compensation was provided through the crowdsourcing platform for completed work. The instructions were designed to align with the prompt used for the user simulator, and an example dialogue was shown to illustrate the task.

Each crowdsourced dialogue session was designed to consist of five dialogue blocks, half the number used in the simulation, to keep the task within CrowdWorks' 60-minute limit.

The human evaluation used the same dialogue system as in the simulation. The system was implemented on DialBB and deployed on Amazon Web Services (AWS). Crowdsourcing workers were allowed to participate up to twice. After excluding sessions with no recorded annoyance ratings, we retained 273 crowdsourced dialogue sessions for analysis. In total, these sessions contained 1,135 completed dialogue blocks with annoyance ratings.

6 Experimental Results

This section presents the results of the LLM-based user simulation and the crowdsourced human evaluation, focusing on user annoyance toward the system's questions and how often the questions elicited correct responses.

In the analyses by question type, we counted only follow-up questions that were actually asked by the system and answered by the user. We associated annoyance ratings with the actually asked questions in the corresponding dialogue block. When a block contained no such follow-up questions, its annoyance rating was not used in the question-type analysis.

6.1 Results with LLM-based User Simulation

Table 2 shows the distribution of annoyance ratings. Annoyance ratings were concentrated at the lower end of the scale, indicating that the system's questions were generally not perceived as highly annoying in the simulation.

We next examined whether annoyance differed by question type. Table 3 summarizes the regression weights for annoyance. Although the coefficient for implicit questions was slightly smaller than those for the other two types, the differences were small overall. The coefficient of determination was also very low ($R^2 = 0.015$), suggesting that question type explained only a limited portion of the variance in annoyance ratings.

We then compared question types in terms of how often they elicited correct responses. Table 4 summarizes the results. Implicit and explicit questions both elicited correct responses at very high rates: 0.999 for implicit questions and 0.992 for explicit questions. In contrast, wh-questions showed a lower proportion of correct responses, at 0.942. A chi-square test of independence showed a significant association between question type and correct response elicitation ($\chi^2(2) = 85.13$, $p < .001$).

6.2 Results with Crowdsourced Users

Table 5 shows the distribution of annoyance ratings. Compared with the LLM-based user simulation, annoyance ratings were more widely distributed in the crowdsourced evaluation.

We next examined whether annoyance differed by question type. Table 6 summarizes the regression weights for annoyance. The coefficients varied somewhat across question types, but their magnitudes were modest and the coefficient of determination was very low ($R^2 = 0.010$), suggesting that question type explained little of the variation in annoyance ratings.

We also conducted an exploratory analysis of annoyance ratings by the position of the question within a dialogue block, as shown in Table 7. For explicit questions, the mean annoyance rating was lowest at the third position. In contrast, the ratings for implicit and wh-questions remained relatively stable across positions.

We then compared question types in terms of how often they elicited correct responses. Table 8

Rating	1	2	3	4	5	Total
Count	267	369	191	245	63	1135

Table 5: Distribution of annoyance ratings in the crowdsourced evaluation.

Implicit	Explicit	Wh	Intercept	R^2
-0.113	-0.226	-0.144	2.953	0.010

Table 6: Regression weights for annoyance in the crowdsourced evaluation.

summarizes the results. Implicit and explicit questions both elicited correct responses at similarly high rates: 0.935 for implicit questions and 0.948 for explicit questions. In contrast, wh-questions showed a lower proportion of correct responses, at 0.820. A chi-square test of independence showed a significant association between question type and correct response elicitation ($\chi^2(2) = 110.06$, $p < .001$). This association mainly reflects the lower elicitation rate for wh-questions.

Thus, the crowdsourced evaluation showed the same pattern as the LLM-based user simulation: implicit and explicit questions elicited correct responses at comparable rates, whereas wh-questions did so less often.

6.3 Summary of Findings

Across both the LLM-based user simulation and the crowdsourced evaluation, differences in user annoyance among question types were small. In the simulation, annoyance ratings were concentrated at the lower end of the scale, and the regression analysis showed only limited differences among question types. In the crowdsourced evaluation, annoyance ratings showed greater variation, but the effect of question type remained small.

In contrast, clearer differences were observed in how often the questions elicited correct responses. In both settings, implicit and explicit questions performed comparably, whereas wh-questions elicited correct responses less often. Taken together, these results suggest that, in the present LLM-based knowledge acquisition setting, question type had only a limited effect on user annoyance, whereas clearer differences were observed in correct response elicitation.

7 Discussion

7.1 User Annoyance

The results suggest that question type had only a limited effect on user annoyance in the present setting. Although annoyance ratings were more widely distributed in the crowdsourced evaluation than in the LLM-based user simulation, the effect of question type remained small in both settings. These results indicate that user annoyance was not strongly determined by question type alone in this setting.

Question type	First	Second	Third
Implicit	2.60 (379)	2.50 (364)	2.59 (244)
Explicit	2.49 (346)	2.53 (336)	2.26 (250)
Wh	2.50 (384)	2.52 (349)	2.58 (258)

Table 7: Mean annoyance ratings by question type and question position within a dialogue block in the crowdsourced evaluation. Numbers in parentheses indicate the number of question–rating associations.

Question type	Correct / Total	Proportion
Implicit	942 / 1007	0.935
Explicit	904 / 954	0.948
Wh	829 / 1011	0.820

Table 8: Correct response elicitation by question type in the crowdsourced evaluation.

This limited effect contrasts with our earlier study (Komatani et al., 2022), where differences in annoyance among question types were larger. The earlier study was conducted with template-based system utterances in a more controlled setting and found that consecutive explicit questions increased user annoyance. In the present study, however, explicit questions did not lead to substantially higher annoyance, even though their surface forms were still relatively similar across instances.

Another possible factor is that question types may have differed in the amount of information-seeking effort required from participants. In the crowdsourced evaluation, answering wh-questions may have required participants to click a Wikipedia link, inspect the page, and identify the relevant country name before answering. Although participants were instructed to consider only the questions themselves, this information-seeking effort may still have influenced their ratings, especially for wh-questions. Thus, differences in annoyance among question types may reflect not only the effect of question form, but also the burden of finding the information needed to answer wh-questions.

The exploratory position-based analysis in Table 7 provides a related observation: explicit questions had the lowest mean annoyance rating at the third position, whereas implicit and wh-questions remained relatively stable across positions. One

possible interpretation is that explicit questions became less annoying at the third position because they presented a candidate answer, allowing participants to answer by confirming the candidate rather than searching for the country name themselves. Future work should use an experimental setting that better separates question form from information lookup effort.

7.2 Correct Response Elicitation

By contrast, clearer differences were observed in how often the questions elicited correct responses. In both settings, implicit and explicit questions performed comparably, whereas wh-questions elicited correct responses less often.

One possible interpretation is that implicit and explicit questions guide users more directly toward the target information by presenting a candidate answer in the question itself. This makes it easier for users to confirm the presented candidate. In this sense, these question forms allow the system to ask about its target information more specifically when it has a candidate answer to present, as in the present experimental setting.

Wh-questions differ from implicit and explicit questions in that they do not present a candidate answer. Instead, they require the user to supply the target content without such guidance. As a result, even when users broadly understand the topic, their responses may be expressed at a different level of granularity from that expected in the system's knowledge representation. Figure 3 shows an example in which the response is broadly relevant but does not provide the specific country information required by the system. For a dish such as "khao man gai," a user might describe it as Thai, Singaporean, or more broadly Southeast Asian, depending on how they conceptualize the dish. In such cases, a wh-question may elicit information at a different level of specificity from what the system intends to obtain. If the system already has a clear candidate hypothesis, asking a more specific question may therefore be a more effective way to confirm whether that candidate is correct.

Accordingly, the comparison between wh-questions and the other two question types should be understood as contrasting open elicitation with candidate-guided confirmation, rather than isolating only the effect of surface question wording. This contrast also helps explain why wh-questions elicited correct responses less often in both settings.

System: What country is khao man gai from?
User: It is popular in Southeast Asia.

Figure 3: An example of a response to a wh-question.

7.3 Implications for Question Type Selection

In dialogue systems that acquire knowledge through interaction, asking questions is useful not only for obtaining previously unknown information, but also for confirming or refining information that the system knows to some extent but is not fully confident about. This suggests that more specific question forms, such as implicit and explicit questions, can be advantageous when the system has sufficient confidence in a plausible candidate.

Although implicit and explicit questions performed comparably in the present experiments, they may still serve somewhat different design purposes. Explicit questions may be preferable when the system needs a clear confirmation or rejection of a specific candidate. Implicit questions, in contrast, may be more suitable when the system aims to strengthen an existing hypothesis while keeping the interaction relatively natural.

This does not mean, however, that wh-questions should always be avoided. When the system lacks sufficient confidence in the information it is asking about, wh-questions may still be useful as a more open way of acquiring information from the user.

Overall, these results suggest that the choice of question type should depend not only on how natural the question is, but also on how reliably it is expected to elicit correct information in the current dialogue context. This consideration is particularly important in settings where systems must obtain supervision from natural interactions rather than from dedicated annotators or tightly controlled workflows (Hancock et al., 2019).

8 Conclusion

In this study, we examined how question type affects knowledge acquisition in LLM-based dialogue systems. We compared three question types: implicit questions, explicit questions, and wh-questions, through experiments with both an LLM-based user simulator and crowdsourced human participants.

The results showed that differences in user annoyance among question types were small in both settings, whereas clearer differences were observed in how often the questions elicited correct re-

sponses. Implicit and explicit questions performed comparably, whereas wh-questions elicited correct responses less often. These findings suggest that, in LLM-based dialogue for knowledge acquisition, candidate-guided question forms may be more effective in eliciting correct responses when the system has sufficient confidence in the target information, whereas wh-questions remain useful when the system lacks such confidence.

As future work, it will be important to examine whether the same tendencies hold in other dialogue tasks and domains. It will also be important to investigate how dialogue systems should choose question types dynamically according to their confidence and the interaction context. In addition, future work should further compare cases in which the system asks about predicted country information that is correct or incorrect.

Limitations

This study has several limitations. First, our experiments focused on a specific knowledge acquisition setting in Japanese dialogues, in which the system asked about unfamiliar food terms and their associated countries. This is a simple single-relation task, and the observed tendencies may not directly generalize to more complex types of knowledge, user-specific facts, other dialogue goals, other languages, or different cultural settings. In particular, language-specific pragmatic conventions and their associated cultural backgrounds may influence how users respond to different question forms.

Second, in our experiments, the candidate country information presented in implicit and explicit questions was always correct. Therefore, the present results do not show how question-type effects may change when the system presents an incorrect candidate country. Although the effect of incorrect candidate information was examined in our previous template-based study (Komatani et al., 2022), it remains to be examined in LLM-based dialogue settings.

Ethical Considerations

This study involved crowdsourced participants. The collected data consisted of dialogue logs and impression ratings. Participants were instructed not to enter personal information, and they were informed that dialogue excerpts could be used as examples in academic papers or presentations without disclosing personal information. Because the study involved free-form text dialogue, unintended personal information could still be entered, so such information was not disclosed when presenting dialogue excerpts.

Acknowledgments

We thank the anonymous reviewers for their constructive comments and valuable feedback. This work was partly supported by JSPS KAKENHI Grant Numbers JP22H00536 and JP23K28147, and JST Moonshot R&D Grant Number JPMJMS2011, Japan.

References

Atheer Algherairy and Moataz Ahmed. 2025. Prompting large language models for user simulation in task-oriented dialogue systems. *Computer Speech & Language*, 89:101697.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4473–4484.

Zhiyuan Chen and Bing Liu. 2018. *Lifelong Machine Learning, Second Edition*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 236–245.

Takumi Furuta, Ryu Takeda, and Kazunori Komatani. 2026. Suppressing unnecessary clarification requests for unknown word acquisition in spoken dialogue using syllable-based ASR confidence. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3667–3684.

Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning knowledge graphs for question answering through conversational dialog. In *Proc. North American Chapter of Association for Computational Linguistics (NAACL)*, pages 851–861.

Kazunori Komatani, Kohei Ono, Ryu Takeda, Eric Nichols, and Mikio Nakano. 2022. User impressions of system questions to acquire lexical knowledge during dialogues. *Dialogue and Discourse*, 13(1):96–122.

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017. Learning through dialogue interactions by asking questions. In *Proc. International Conference on Learning Representations (ICLR)*.

Bing Liu and Sahisnu Mazumder. 2021. Lifelong and continual learning dialogue systems: Learning during conversation. In *Proc. Conference on Artificial Intelligence (AAAI)*, pages 15058–15063.

Sahisnu Mazumder, Bing Liu, Shuai Wang, and Nianzu Ma. 2019. Lifelong and interactive learning of factual knowledge in dialogues. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 21–31.

Mikio Nakano and Kazunori Komatani. 2024. DialBB: A dialogue system development framework as an educational material. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 664–668.

Bo Ni, Yu Wang, Leyao Wang, Branislav Kveton, Franck Dernoncourt, Yu Xia, Hongjie Chen, Reuben Luera, Samyadeep Basu, Subhojyoti Mukherjee, Puneet Mathur, Nesreen K. Ahmed, Junda Wu, Li Li, Huixin Zhang, Ruiyi Zhang, Tong Yu, Sungchul Kim, Jiuxiang Gu, and 11 others. 2026. A survey on LLM-based conversational user simulation. In *Proc. European Chapter of the Association for Computational Linguistics (EACL)*, pages 4266–4301.

Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. 2017. Lexical acquisition through implicit confirmations over multiple dialogues. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 50–59.

Tsugumi Otsuka, Kazunori Komatani, Satoshi Sato, and Mikio Nakano. 2013. Generating more specific questions for acquiring attributes of unknown concepts from users. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 70–77.

Aasish Pappu and Alexander Rudnicky. 2014. Knowledge acquisition strategies for goal-oriented dialog systems. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 194–198.

Matthew Purver. 2002. Processing unknown words in a dialogue system. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 174–183.

Hossein A. Rahmani, Xi Wang, Mohammad Aliannejadi, Mohammadmehdi Naghiaei, and Emine Yilmaz. 2024. Clarifying the path to user satisfaction: An investigation into clarification usefulness. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1266–1277.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2737–2746.

Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. 2024. Reliable LLM-based user simulator for task-oriented dialogue systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pages 19–35.

Maarten van Schagen and Alistair Knott. 2004. Tauira: A tool for acquiring unknown words in a dialogue context. In *Proceedings of the Australasian Language Technology Workshop 2004*, pages 131–138.

Issei Waki, Ryu Takeda, and Kazunori Komatani. 2025. Learning to ask efficiently in dialogue: Reinforcement learning extensions for stream-based active learning. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 431–440.

Jason Weston. 2016. Dialog-based language learning. In *Proc. International Conference on Neural Information Processing Systems (NIPS)*, pages 829–837.

A Annoyance Rating Instruction

Figure 4 shows the instruction used for rating the annoyance of the system's questions in both the LLM-based user simulation and the human evaluation. The instruction explicitly stated that users could choose higher ratings when they felt discomfort, in order to clarify how the scale should be used.

Please rate how annoying the system's questions have been in the dialogue so far.
Consider only the questions themselves.

1: Not annoying at all (fits naturally into the flow of the conversation)
2: Not very annoying
3: Neither
4: Somewhat annoying (unnatural or slightly difficult to answer)
5: Very annoying (does not fit the flow of the conversation or is difficult to answer)

Please base your rating on whether the questions fit the flow of the conversation, are easy to answer, and feel natural. If any of the questions felt awkward or uncomfortable, please do not hesitate to give a stricter rating, such as 4 or 5 rather than 2.

Figure 4: Annoyance rating instruction, translated from the original Japanese.

B Prompt Design for Follow-up Question Generation

Follow-up questions were generated by the LLM from prompts that specified the target question type together with contextual information such as the dialogue history, system persona, and current situation. In all cases, the model was also instructed to keep the tone consistent with the preceding dialogue. Separate task instructions were used for implicit, explicit, and wh-questions.

For implicit questions, the model was instructed to generate a short statement about the candidate country, containing non-doubtful content that could naturally invite user agreement.

For explicit questions, the model was instructed to ask directly whether the target food was a dish from the candidate country, without adding extra content.

For wh-questions, the model was instructed to ask directly which country the target food was from, again without adding extra content.

C Prompt for the LLM-based User Simulator

Figure 5 shows the prompt used for the LLM-based user simulator.

You are a human user talking with a dialogue system about new and unfamiliar foods.

[Goal]
- Please talk briefly and naturally about new and unfamiliar foods.
- Do not ask questions on your own very often.
- If the system asks a question, answer naturally according to its content.
- You do not need to pay close attention to consistency across the whole dialogue.

[Experimental Instructions]
- If the system's utterance contains a string enclosed in square brackets ([ ]), treat it as an experimental instruction rather than as part of the conversation itself.
- Always follow the experimental instructions with top priority.
- If an experimental instruction says something like "use a certain word within the next few utterances," use that word within a few utterances. In other words, it does not have to be used in the next utterance.
- Do not output the content of the experimental instruction itself as part of the user's utterance.
- Treat any food name included in an experimental instruction as something you are hearing for the first time.
- Use the food name specified in the experimental instruction exactly as shown. Do not paraphrase it or change its form.

[Country-name Constraints in Experimental Instructions]
- The country a dish is from is important information.
- Do not mention a country name unless you are asked about it.
- If the system asks which country the dish is from, answer the question.
- If you do not know, you must look it up using the URL shown in the experimental instruction before answering. Do not respond with only "I don't know" or "I'm curious."

[Deviations from the Conversation]
- If the system's utterance feels largely unrelated to the conversation or unnatural, respond in a way that shows your discomfort or sense of mismatch.

[Output Format]
- Output only the user's utterance.
- Do not output names, role labels, explanations, or quotation marks.
- Limit each utterance to 50 Japanese characters or fewer.

Figure 5: Prompt used for the LLM-based user simulator, translated from the original Japanese.

D Example Dialogue with the LLM-based User Simulator

Figure 6 shows an excerpt from a dialogue with the LLM-based user simulator.

System: Hello. Thank you for joining today. Let's talk about food.
[Text inside brackets is an experimental instruction. Please respond only to the rest of the dialogue.]

User: Hello! I'm looking forward to talking about new foods. Have you eaten anything interesting recently?

System: Hi! Recently, I've become really interested in foods from around the world. How about you?
[Please use the word "ftira" within the next few utterances. It does not have to continue the current topic.]
`<Wikipedia URL>`

User: I like foods from around the world. I've heard of ftira. It seems to be getting popular recently.

System: Malta is an island country in the Mediterranean and has a unique food culture. *(Implicit question)*

User: I wonder what ftira tastes like. It sounds intriguing!

System: Ftira is pleasantly spiced and quite addictive!
[Please use the word "maritozzo" within the next few utterances. It does not have to continue the current topic.]
`<Wikipedia URL>`

User: That sounds flavorful! Maritozzo also sounds interesting.

System: Excuse me, which country is maritozzo from? *(Wh-question)*

User: Maritozzo is an Italian dessert! It looks delicious with lots of cream.

System: I'd like to try an Italian sweet bun too!
[Please use the word "svid" within the next few utterances. It does not have to continue the current topic.]
`<Wikipedia URL>`

User: Cream-filled desserts are really appealing, aren't they? Have you heard of svid? I wonder what it tastes like.

System: Is svid an Icelandic dish? *(Explicit question)*

User: I think it's an Icelandic dish. I'm curious what it tastes like.

System: Svid sounds unique and interesting!
[Please rate how annoying the system's questions have been in the dialogue so far.]
(For the full instruction, see Figure 4.)

User: 3: Neither. *(Annoyance rating: 3)*

System: Now let's move on to the next dialogue and continue talking about food. Thank you in advance.
[You may forget the content and impression of the dialogue so far.]

User: Hello! I'm looking forward to talking about new foods. Have you eaten anything interesting recently?

(The same procedure was repeated for the remaining blocks.)

Figure 6: Excerpt from a dialogue with the LLM-based user simulator, showing examples of implicit, wh-, and explicit questions, translated from the original Japanese. Square brackets indicate experimental instructions.