Proceedings of the Design Society, Volume 6: DESIGN 2026
https://doi.org/10.1017/pds.2026.10583

LLM-based voice chatbot surveys as an alternative to post-experience questionnaires: probe-controlled, ultra-short field interviews

Fujian Ding✉, Yuki Taoka and Momoko Nakatani

Institute of Science Tokyo, Japan

✉ ding.f.b02e@m.isct.ac.jp

ABSTRACT: Chatbot-based surveys offer low-burden, in-situ data collection, yet unconstrained LLMs often drift from research aims. We conducted 359 ultra-short, post-experience voice interviews in a public venue to compare a framework-guided LLM, an unconstrained LLM, and fixed questions. The guided approach produced significantly longer responses than fixed questions and yielded the richest diversity of process-specific accounts. These findings show that probe control is essential for eliciting actionable, experience-grounded feedback in real-world, time-limited settings.

KEYWORDS: design methods, human-centred design, interviews, artificial intelligence (AI), probing

1. Introduction

As information technology advances, services are increasingly used within diverse contexts. This trend has elevated the importance of User-Centered Design (UCD) and engineering design, where identifying latent user needs is critical for product innovation (Huang et al., 2025). Questionnaires are a common data collection method in UCD, valued for their cost-effectiveness and rapid large-scale administration (Evans & Mathur, 2018). Typically, questionnaires are designed to capture both quantitative and qualitative data, comprising closed-ended questions for straightforward quantification and open-ended questions to elicit detailed experiences and impressions (Chen, 2017).

However, questionnaires cannot probe responses in real time, risking missed insights (Jacobsen et al., 2025). This absence of interaction can also lead to low-quality responses or non-responses. Furthermore, the effort required for open-ended text entry is burdensome, often causing survey fatigue and participant dropout (Porter et al., 2004). This reflects a common HCI trade-off: in situ methods such as experience sampling and mobile sensing improve temporal fidelity but often increase burden or constrain response depth (Van Berkel et al., 2017).

In response to these limitations, the potential for chatbots to "probe" qualitative survey questions has been explored, with numerous studies employing conversational agents that follow a predefined flow to pose questions (Xiao et al., 2020). This line of research suggests that chatbot-based surveys, through their interactive format, can foster greater self-disclosure and gather higher-quality data than static questionnaires (Kim et al., 2019; Xiao et al., 2020). Recent advancements in Large Language Models (LLMs) further enhance this potential, enabling "in-the-moment" interviews that can elicit deeper user insights and latent needs comparable to human interviewers (Liu et al., 2025; Korn et al., 2025).

However, a significant constraint has been identified: leaving the dialogue entirely to an uncontrolled LLM often fails to collect the specific data that researchers require (Li & Cai, 2025). Design research currently lacks a systematic framework that codifies when and what to probe (Wilson, 2013), which constitutes the research gap this paper addresses. This stems from the fact that interviewing effectiveness still hinges on the irreproducible tacit knowledge of individual expert interviewers (Hahn & Uppa, 2013).

https://doi.org/10.1017/pds.2026.10583 Published online by Cambridge University Press

Therefore, building on a user interview framework that systematizes designers' tacit interview strategies (Ding et al., 2026), this study proposes an LLM-driven dialogue method that uses this framework to control probing in ultra-short voice interviews. Our aim is to combine the convenience of questionnaires with the experiential depth of interviews through a reproducible probing mechanism. Specifically, we investigate LLM-based, ultra-short (four-question) voice interviews conducted immediately after user experiences—a context in which users have limited time and traditional chatbot surveys may be impractical. To validate our proposed method (a voice-based dialogue system that takes the framework as input), we address the following Research Questions (RQs):

- RQ1: Compared to conventional open-ended questionnaires, can an LLM-based dialogue method collect longer responses and more diverse content?
- RQ2: When conducting a voice-based questionnaire using an LLM, what differences emerge in the dialogue based on the presence versus the absence of the proposed framework?

2. Related research

2.1. User information gathering by questionnaire-based survey

Capturing subjective user experience is valuable for hypothesizing system requirements. Methods that investigate this experience after use are crucial techniques within the evaluation and improvement phases of the UCD process. Questionnaires enable researchers and practitioners to capture user concerns using a series of closed-ended or open-ended questions. Open-ended questions are valued for providing deep insights into user experiences and sentiments and are also useful for understanding the rationale behind responses to closed-ended items (Chen, 2017).

However, several factors are known to degrade the quality of open-ended responses. Participants can be overwhelmed by the questions, leading to dropout, or they may respond hastily without deep consideration, degrading response quality. The typing burden of open-ended items adds cognitive load (Chen, 2017), leading participants to skip or minimize answers (Krosnick, 1991).

Research has attempted to mitigate these drawbacks through strategies such as adding simple open-ended probe questions after closed-ended items or incorporating participant's previous data into survey questions (Velykoivanenko et al., 2024). In contrast, our work adapts human interviewing probes to respond dynamically to user utterances.

2.2. Chatbot-based surveys and in situ elicitation

Research on conversational agents for information elicitation has advanced, with applications reported in survey contexts. In the survey domain, it has been argued that chatbots can present items as personalized, conversational messages, which may enhance response quality. Empirical evidence suggests that chatbots can reduce cognitive effort required to respond, which can reduce "satisficing" (Kim et al., 2019) and this reduction is associated with higher engagement (Xiao et al., 2020).

Recent studies further highlight that LLM-powered interviews can collect "in-the-moment" qualitative data at scale (Liu et al., 2025) and foster greater self-disclosure (Papneja & Yadav, 2025). However, these studies report that the turn-by-turn interaction of these text-based chatbots increases the time required compared to the efficiency of "quick completion" afforded by questionnaires, without necessarily improving usability or enjoyment (Xiao et al., 2020; Kim et al., 2019).

This line of work also connects to HCI in situ and longitudinal methods. Experience sampling captures momentary self-reports but can impose burden, whereas mobile sensing scales to long-term behavior yet lacks subjective meaning in context (Van Berkel et al., 2017). Diary approaches add this meaning but often face attrition (Li et al., 2024). Our work contributes ultra-short voice probing as a lightweight qualitative layer immediately after experiences.

2.3. Interview probing by conversational AI

Modern LLMs exhibit strong capabilities in text understanding and are increasingly applied to interviews and surveys. For example, they can automate requirements elicitation (Korn et al., 2025) and identify novel customer needs from user-generated content (Huang et al., 2025). By contrast, in exploratory user-understanding and post-use surveys, unconstrained LLM interviews lacking explicit control of the

https://doi.org/10.1017/pds.2026.10583 Published online by Cambridge University Press

probing process may fail to obtain desired answers (Li & Cai, 2025). This motivates the introduction of principled probing strategies to improve dialogue quality.

In social-science interviewing, probes play a central role in eliciting rich, nuanced data; they are follow-up questions that surface unarticulated information and deepen understanding. Patton (2014) groups probes into three categories: Detail-oriented (e.g., who/where/when), Elaboration, and Clarification. Extending this work, Robinson (2023) proposed the DICE model (Descriptive Detail, Idiographic Memory, Clarifying, and Explanatory), reframing detail-oriented probes as descriptive (including feelings and internal states), redefining elaboration as idiographic memory for one-off events, maintaining clarifying, and adding explanatory probes to uncover reasons behind actions and thoughts. However, these models specify probe types but provide limited guidance on when to probe, which is essential in dynamic interviews. To address this gap, we previously analyzed interviews with expert service designers to identify conversational cues that warrant probing and to refine probe types for LLM-based implementation. The next section briefly summarizes the resulting framework.

3. The user interview framework

To address the lack of systematic guidance on when and what to probe identified in the Introduction, we employ a user interview framework derived in our prior qualitative study of expert interviewing practices in service design (Ding et al., 2026). In that work, 11.7 hours of interview recordings with 10 professional service designers (Taoka et al., 2024) were coded using inductive qualitative content analysis. We briefly summarize its structure and how we operationalized it for LLM control.

Figure 1 illustrates how the framework was operationalized as a structured prompt that drives a stepwise control sequence for follow-up generation. The LLM initiates the session using the 'Raise Topic' category. Upon receiving user response, the model is instructed to detect predefined "Utterance Features" as candidate probing points, select one feature as the current target, choose a corresponding "Direction of In-depth Exploration", and generate a single follow-up aligned with that direction. If no features are detected in the current utterance, the prompt specifies a fallback to features from earlier turns or to introducing a new topic. "Questioning" and "Listening Techniques" are included only as stylistic guidance and are not modeled as control variables.

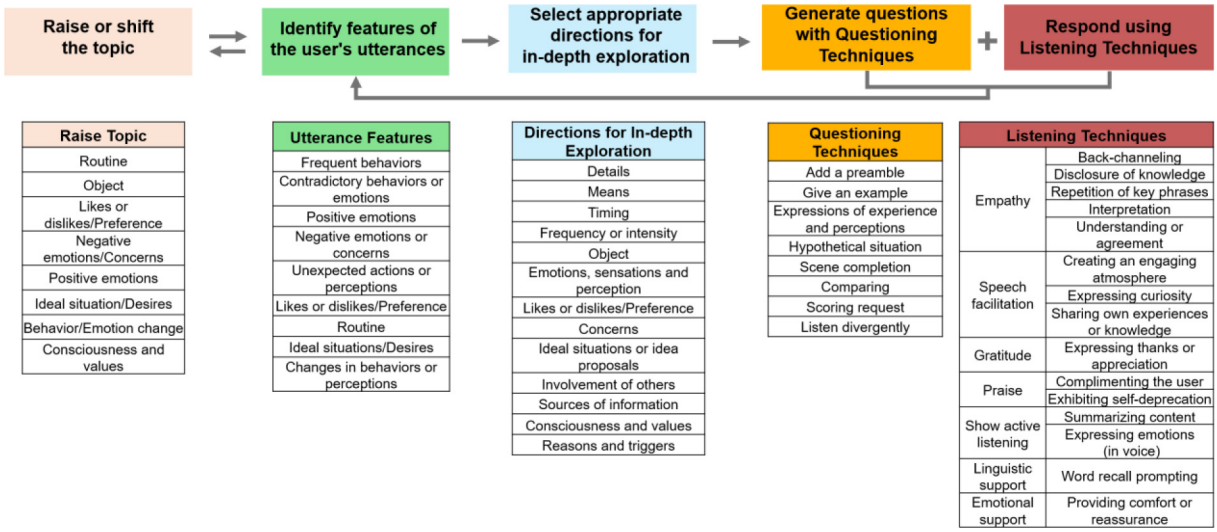


Figure 1. The user interview framework

4. Experimental setup and analysis procedure

4.1. Participants and context

To validate the proposed framework-guided LLM interview method, we conducted a field experiment in a large, crowded public exhibition venue where visitors interacted with a short motion-based activity using an exercise-support system (ESS). Immediately after using the ESS, visitors aged 7 to 90 years were invited to participate in an ultra-short interview. All visitors received a brief explanation and provided informed consent. The study protocol was approved by the Institutional Review Board of the

https://doi.org/10.1017/pds.2026.10583 Published online by Cambridge University Press

authors' institution (Approval No. 2024182). In total, 359 participants took part (mean age 48.1 years, SD 19.2, median 55; the most frequent age band was 50 to 60 years, 27.9%). Because environmental noise was unavoidable in this public setting, we used an external speaker and microphone. Participants were assigned in near-equal numbers across conditions (A:115, B:114, C:130).

4.2. Voice-based interview system overview

The iOS tablet app used the same UI across conditions: participants answered one fixed opening question followed by three probe turns; each question was shown on screen and spoken via text-to-speech (TTS), and responses were captured via a microphone button and auto-submitted after STT transcription. We used Apple Speech (SFSpeechRecognizer) for speech-to-text (STT) transcription, sent transcripts with a predefined prompt to the LLM API (gemini-2.5-flash-lite-preview-09-2025) with temperature set to 0.2, and synthesized outputs using Google Cloud Text-to-Speech (ja-JP-Neural2-B).

4.3. Procedure and experimental design

We opened with a fixed question "How was your experience with the exercise-support system? Please share your impressions." followed by three probing turns. The UI, STT, and TTS were identical across conditions; only the question-generation differed. Participants were informed that the system was AI-driven. As shown in Figure 2, follow-ups were produced by one of three pipelines: **(A) Framework-guided LLM**, which applied a three-step control sequence: feature detection, direction selection and question generation that derived from our framework; **(B) Simple LLM**, which used a shared role prompt and generated three questions freely; and **(C) Fixed Questions**, predefined based on Baxter et al. (2015) and shown in Table 1. For (A) and (B), the shared prompt defined the model's role "You are a service designer eliciting users' experience impressions; generate questions to investigate their experience", included a brief description of the ESS, and incorporated the latest user utterance and dialogue history. To reduce time-of-day and venue-traffic confounds, we counterbalanced condition blocks across days by rotating the order of A, B, and C while keeping block durations constant. Sessions concluded with an LLM-generated brief closing remark.

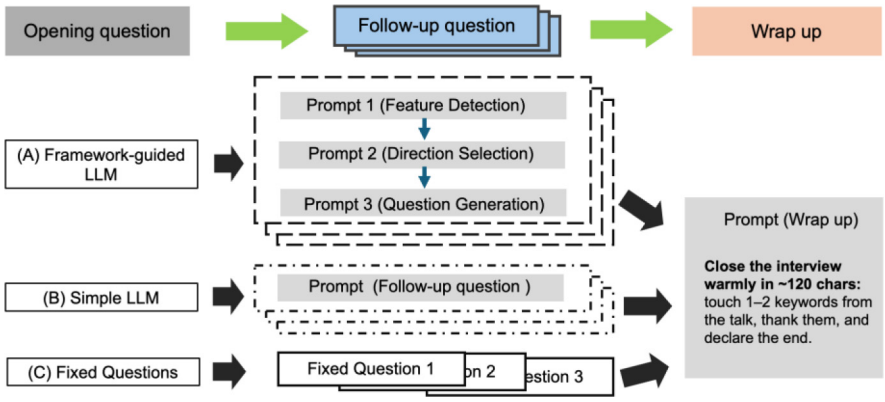


Figure 2. Dialogue flow and question-generation pipelines across conditions (A–C)

Table 1. Fixed questions used in condition C

Turn	Question
1	What was the most impressive point you noticed during the experience?
2	What aspects of the operation or visual movement did you find "good"?
3	Were there any points about the system you just experienced that you felt should be improved?

4.4. Data analysis

To assess whether the three approaches differed in response volume, we quantified each participant's utterance length. We counted characters per participant per condition, treating half-width and full-width characters as single characters and excluding punctuation. Because the character-count data did not meet

https://doi.org/10.1017/pds.2026.10583 Published online by Cambridge University Press

normality, we tested group differences using the Kruskal–Wallis test. Given a significant omnibus result, we performed Dunn's pairwise comparisons with Bonferroni adjustment.

To examine whether applying the framework yielded qualitative differences, we analyzed all dialogue logs across the three conditions (N = 359). First, we applied deductive coding to follow-up questions in conditions (A) Framework-guided LLM and (B) Simple LLM, using a predefined codebook **"Probing-direction"** derived from the framework's "Directions of In-depth Exploration" shown in Table 2. Second, we deductively coded all participant utterances across the three conditions using a predefined **"Utterance-features"** codebook derived from the framework's "Utterance Features" shown in Table 3. Emergent recurring themes were incorporated by extending the codebooks according to predefined criteria, and coding decisions were iteratively cross-checked to maintain consistency.

Table 2. Probing-direction codes for follow-up questions

Code	Definition
Details: Idiographic	Follow-up questions that vertically drill down into an experience the user has already mentioned, asking for concrete episodes, situations, or step-by-step descriptions.
Emotions/Sensations	Questions that newly ask about emotional states or bodily sensations about experience.
Timing	Questions that ask when it happened, or about the temporal order or specific moment.
Ideas/Ideal situations	Questions that invite future-oriented ideas, ideal scenarios, desired features.
Reasons/Triggers	Questions that ask why something was experienced in a certain way, probing causes, mechanisms, motivations, or triggers behind the user's evaluation or feeling.
Concerns	Questions that explicitly elicit worries, pain points, obstacles related to activity/system.
Frequency/Intensity	Questions that probe how often it happens, how strong it is, or how large the impact is.
Lateral probe	Questions that horizontally broaden the topic, asking for other examples, aspects, or items beyond the current focus, rather than deepening the episode.

Table 3. Utterance-features codes for participant responses

Code	Definition
Positive emotions	Utterances that express positive affect toward the experience.
Negative emotion/ concerns	Utterances that express negative feelings (e.g., pain, fatigue, annoyance) or concrete worries about the activity or system.
Ideas/Ideal situations	Utterances that propose new features, ideal scenarios, potential uses, or imagined benefits of the system in the future.
Likes or dislikes/ Preference	Utterances that state what the user personally liked or disliked, or preferences regarding content, pace, style, or setting.
Behavioral/Cognitive changes	Utterances describing a change in the user's behavior, awareness, or way of thinking resulting from the experience.
Contradictory behaviors / emotions	Utterances that combine seemingly conflicting elements (e.g., positive and negative feelings or outcomes) within the same statement.
Refusal	Utterances indicating non-cooperation, lack of opinion, or unwillingness to elaborate.

After coding, we aggregated code counts by condition. Probing-direction distributions for Conditions A and B and utterance-feature distributions for Conditions A–C were compared using chi-square tests of independence. We report Cramér's V as an effect size and, where tests were significant, used standardized residuals to interpret which categories differed by condition. For brevity, we display only codes that account for at least 2% of coded instances in any condition; rarer codes were included in the chi-square analyses but are omitted from Tables 2 and 3 and from Figure 4.

https://doi.org/10.1017/pds.2026.10583 Published online by Cambridge University Press

5. Results

5.1. Overview of the collected data

Figure 3 shows utterance length by condition, A (Framework-guided LLM; n=115), B (Simple LLM; n=114), and C (Fixed Questions; n=130). Median lengths were A = 112, B = 110.5, C = 84, indicating that LLM-based methods tended to elicit longer responses than the Fixed Questions. Because lengths were non-normal and sample sizes differed, we used a Kruskal-Wallis test, which indicated a significant difference (H = 7.29, p = .026). Dunn-Bonferroni post hoc tests showed A > C (p = 0.029), with no other significant pairwise differences (Table 4).

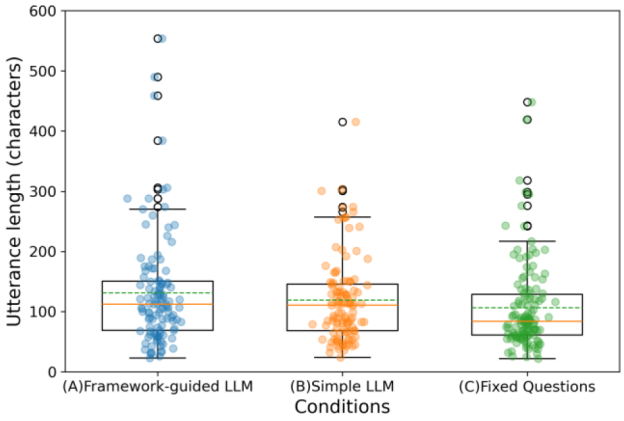


Figure 3. Box plots of utterance length by condition (A–C)

Table 4. Pairwise comparisons of utterance length across conditions

Comparison		Z	p (two-sided, adjusted)
(A) Framework-guided LLM	(B) Simple LLM	0.67	1.000
(A) Framework-guided LLM	(C) Fixed Questions	2.59	0.029
(B) Simple LLM	(C) Fixed Questions	1.89	0.176

5.2. Distribution of questions and utterances

First, we classified LLM-generated probes using the **probing-direction** codebook (Table 2). We used eight categories (seven framework-derived plus one emergent category, "Lateral probe") and coded 408 probes for A and 404 for B (Figure 4-left). A was dominated by "Details: Idiographic" (56.4%, 230), whereas B produced more "Ideas/Ideal situations" (37.1%, 150) and more "Lateral probe". The distributions differed significantly (χ^2 (8) = 165.57, p < .001) with a large effect (Cramér's V = 0.45). Residual analysis indicated that A generated more "Details: Idiographic", "Emotions/Sensations", and "Timing", while B generated more "Ideas/Ideal situations" and "Lateral probe"; "Concerns" was almost absent in B (0.2%).

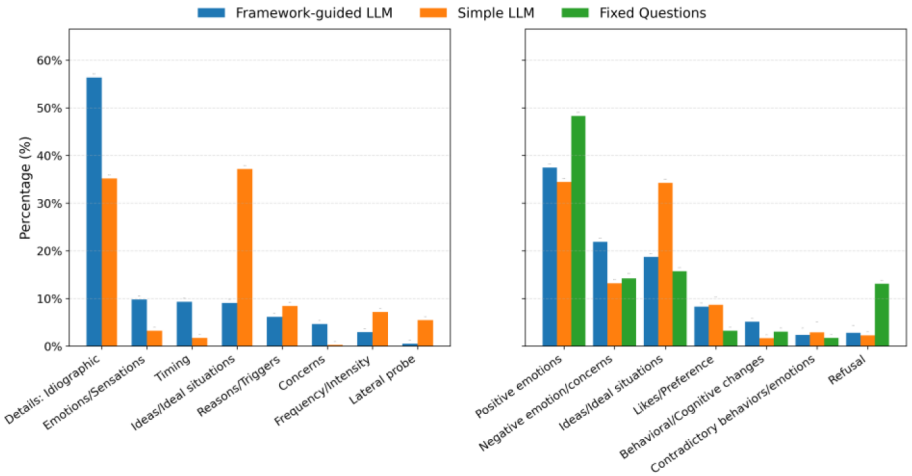


Figure 4. Probing directions (A vs. B) (left); utterance features (A–C) (right)

https://doi.org/10.1017/pds.2026.10583 Published online by Cambridge University Press

Second, we classified user responses using the utterance-features codebook (Table 3) and added one emergent category, "Refusal". Coding totals were A = 470, B = 485, and C = 528 (Figure 4-right). The distributions differed significantly (χ^2 (16) = 165.83, $p < .001$) with a small to moderate effect (Cramér's V = 0.236). Residual analysis showed that B elicited more "Ideas/Ideal situations", A elicited more "Negative emotions/concerns" and "Behavioral/Cognitive changes", and C elicited more "Positive emotions" and "Refusal", while yielding fewer "Ideas/Ideal situations" and "Likes or dislikes/Preference".

Focusing on substantive content, we computed diversity on non-Refusal utterances using Shannon entropy (H) and its effective number of categories exp(H), alongside evenness (J) and the Gini-Simpson index. Diversity ranked A > B > C: A exp(H) = 5.03 (H = 1.62), B = 4.53 (H = 1.51), and C = 3.63 (H = 1.29). A sensitivity check including Refusal preserved the same ordering (A 5.46 > B 4.88 > C 4.52).

5.3. Qualitative results: growth in depth and specificity

In Condition A (Framework-guided LLM), we observed follow-up chains that progressively deepened initially vague user statements (e.g., affect, sensations, ideas) into specific, actionable accounts. Each session comprised eight turns (four questions), labeled T1–T8; AI denotes the system question and USER denotes participant responses, with T1–AI as the shared opening question. In the representative example (Table 5), the vague emotion *"It was fun"* was concretized through probing down to a specific, actionable behavior.

In Condition B (Simple LLM), as shown in Table 6, similar openings were frequently met with probes that steered the conversation toward future-oriented ideals or feature proposals rather than deepening the experience. In the paired example following *"It was fun"*, turns from T5 onward diverted toward the participant's intent to continue and desired features (e.g., *"various mountains"*)—topics reflecting individual speculation rather than a deeper probe of the experienced emotion itself.

Table 5. Dialogue example (condition A, participant in their 50s)

Label	Transcript
T2–USER	It was fun.
T3–AI	What part specifically was fun? (Response omitted)
T5-AI	What specific changes did you notice at that moment? (Response omitted)
T7-AI	What kind of specific movements were there?
T8-USER	I enjoyed tracking how a pointer-like element moved together when I moved my hands up or to the side.

Table 6. Dialogue example (condition B, participant in their 60s)

Label	Transcript
T2-USER	It was fun.
T3-AI	Which part, specifically, was fun? (response omitted)
T5-AI	Do you feel like continuing this in the future? (response omitted)
T7-AI	What kind of visuals would you like to exercise with?
T8-USER	Let's see, I think it would be nice if there were various mountains.

In contrast, the Fixed Questions (C) condition, which lacks adaptive probing, resulted in simple, single-exchange responses to the "impression", "good points" and "improvement points" prompts. Consequently, utterances were often vague and concretization was limited. Furthermore, we observed many refusals to answer, such as *"I don't know"* (Table 7, Figure 4-right).

5.4. Qualitative results: recovery after refusals

In Condition A, when participants produced refusals (e.g., *"nothing in particular," "I don't know"*), the system often reconnected to a previously raised utterance feature to resume the dialogue. For example:

T6–USER: "Nothing in particular."

https://doi.org/10.1017/pds.2026.10583 Published online by Cambridge University Press

Table 7. Example user responses (condition C, participant in their 70s)

Label	Transcript
T2-USER	I think it's good.
T4-USER	I'd rather look at the scenery than run alone.
T4-USER	When I get tired, I'll go slowly.
T8-USER	It was only for a little while, so I don't know.

T7-AI:"Understood, 'Nothing in particular.' Could you please tell me a bit more about that feeling you mentioned earlier, about 'being able to go at your own pace'? [. . .]"

In contrast, when faced with a refusal, the Simple LLM (B) was observed to "jump" to an unrelated topic (often, a generic question about "Ideas/Ideal situations") rather than probing the original topic, resulting in a shallower investigation. For example:

T6-USER: "I don't know."

T7-AI: "Thank you. If there are any points that could be improved, or features you wish it had, to get more people to use this system, could you please tell us?"

And because Condition (C) lacks adaptive probing, refusal responses emerged, leaving the system no choice but to proceed to the next predefined question.

6. Discussion

6.1. What is the effect of introducing an LLM?

We investigated whether a real-time voice-based LLM dialogue method could collect richer data than Fixed Questions (RQ1). While the (A) Framework-guided LLM yielded significantly more utterance volume than (C) Fixed Questions, the (B) Simple LLM showed no significant difference from C. This suggests that simply introducing an LLM does not guarantee more data; rather, the dialogue control appears critical. Condition A's systematic strategy scaffolded the concretization of vague statements and enabled recovery from refusal, encouraging sustained participant engagement

Beyond volume, we also quantified the diversity of non-refusal content using Shannon entropy and its effective number of categories, diversity ranked A > B > C, with a great concentration on "Positive emotions". These results indicate that introducing an LLM appears to counteract the skew toward brief positive remarks, thereby broadening the mix of elicited categories.

The finding that the weakly controlled (B) Simple LLM did not statistically outperform (C) Fixed Questions differs from reports such as Xiao et al. (2020). largely due to context and method: their text-based, multi-turn chatbots operated with ample time, whereas our setting enforced short-turn, real-time voice interactions with a different probing implementation. Consistent with this interpretation, our qualitative observations suggest that B favored "Ideas/Ideal situations" and "Lateral" probes, shifting talk to hypotheticals rather than elaborating past experiences and may have constrained volume gains. Nevertheless, both lines of work converge on the benefit that conversational formats can promote self-disclosure (Kim et al., 2019) relative to static prompts.

Qualitatively, the LLM-based methods (A and B) elicited more multi-faceted content than C. Utterance analysis (Figure 4-right) showed that C yielded more "Positive emotions" and a significantly higher Refusal rate. In contrast, A and B layered listening responses and probes onto the user's immediate utterance, acquiring deeper information (see examples in the Section 5.3). This aligns with prior findings: in our short-turn voice setting, volume gains require guided control, whereas a conversational format alone still enhances the diversity of information compared to static prompts.

6.2. What is the value of applying the user interview framework to an LLM?

To address RQ2: What differences emerge based on the framework's presence? While utterance volume was similar between A and B, code distributions revealed clear qualitative differences regarding supported inferences (Tavory, 2020). First, Condition A supported inference about "Open Contexts," the concrete processes of user experience. Framework-guided (A) generated process-oriented probes

https://doi.org/10.1017/pds.2026.10583 Published online by Cambridge University Press

(e.g., Details: Idiographic), progressively elaborating vague statements into actionable behaviors (Table 5). This is important for reconstructing user experiences in design research.

Conversely, Simple LLM (B) favored "Ideas/Ideal situations" (Figure 4-right), risking ***"Closed Contexts"*** (Tavory, 2020). B's hypothetical questions often solicited on-the-spot opinions, making responses more likely to be artifacts of the interview situation than reflections of latent needs and thus a weaker basis for design inference. Additionally, B used "Lateral probes" more frequently than A, indicating a tendency to diverge horizontally. In contrast, A more consistently elicited episodic, condition-specific accounts of what happened, when, and why, yielding constraints and failure-related details that are better suited for early requirement reasoning than hypothetical ideals. As one indicator, A surfaced "Negative emotion/concerns" more often than B (A: 21.9% vs. B: 13.2%), signals that can reframe perspectives and stimulate innovation (McDonagh & Thomas, 2010).

This distinction is critical for preliminary engineering design. Capturing experience-based knowledge like failure modes minimizes rework costs (Mountney et al., 2007). Unlike Method B's speculative ideation, Method A provides more grounded evidence for requirement definition, which can help engineering designers mitigate development risks. Practically, our system serves as a practical tool for "reflective design documentation" (Dalsgaard & Halskov, 2012). By enforcing probes on idiographic details rather than drifting into shallow ideation, this systematic control supports deeper reflection-on-action, ensuring that data reflects lived experiences rather than on-the-spot speculations.

6.3. Limitations and future work

This study has limitations regarding evaluation, demographics, and context. First, we did not validate the utility of the data with practitioners, interpreting our results instead as proxies based on content specificity. Second, our analysis aggregates a wide age range (7 to 90 years); future work requires a more granular demographic breakdown.

Third, as a field experiment, external factors such as noise and participant haste may have influenced interactions. While randomization mitigated bias, the less adaptive Fixed Questions may be more vulnerable to environmental distractions than the Framework-guided condition. Finally, we did not systematically evaluate accessibility. Although voice and text options were available, further research is needed to assess support for users with speech or hearing impairments.

6.4. Conclusion

We evaluated ultra-short voice interviews in a public venue, comparing framework-guided LLM probing with a simple LLM and Fixed Questions. The simple LLM did not increase response volume, whereas the framework-guided condition elicited significantly more talk than Fixed Questions. It also shifted follow-ups from hypotheticals to concrete lived experience, surfacing negative feedback. While acknowledging the specific constraints of the public field setting, these results indicate that a dialogue control framework capturing designers' tacit probing logic can make LLM-mediated interviews effective and scalable. These findings suggest a division of labor in which the system handles lightweight, in-the-moment probing to secure a baseline level of qualitative detail, allowing researchers to focus on higher-level work such as sensemaking, creative ideation, and iterative system improvement.

Acknowledgement

This work was supported by JST SPRING (JPMJSP2180), JSPS KAKENHI (JP23K17155, JP24K15615), and the Nippon Foundation HUMAI Program. We are grateful to our lab members for their assistance with the experimental work.

References

Baxter, K., Courage, C., & Caine, K. (2015). *Understanding your users: a practical guide to user research methods*. Morgan Kaufmann.

Chen, Y. (2017). *The Effects of Question Customization on the Quality of an Open-Ended Question*. Nebraska Department of Education, Data, Research, and Evaluation.

Dalsgaard, P., & Halskov, K. (2012). Reflective design documentation. In *Proceedings of the Designing Interactive Systems Conference* (pp. 428–437). Association for Computing Machinery. https://doi.org/10.1145/2317956.2318020.

Ding, F., Taoka, Y., & Nakatani, M. (2026). Cue-based probing logic for qualitative inquiry in user-centered design [Manuscript submitted for publication].

https://doi.org/10.1017/pds.2026.10583 Published online by Cambridge University Press

Evans, J. R., & Mathur, A. (2018). The value of online surveys: A look back and a look ahead. *Internet Research*, 28(4), 854–887. https://doi.org/10.1108/INTR-03-2018-0089.
Hahn, Y. A., & Uppa, H. (2013). The interplay between experiential and theoretical knowledge in service design. In *Proceedings of DRS Experiential Knowledge Special Interest Group Conference* (Vol. 114).
Huang, S., Qin, H., Chan, T. T., & Wang, Y. (2025). Identifying novel customer needs from user-generated content for product development using pre-trained language model. *Journal of Engineering Design*, 1–21. https://doi.org/10.1080/09544828.2025.2504850.
Jacobsen, R. M., Cox, S. R., Griggio, C. F., & Van Berkel, N. (2025). Chatbots for data collection in surveys: A comparison of four theory-based interview probes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–21). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713316.
Kim, S., Lee, J., & Gweon, G. (2019). Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300316.
Korn, A., Gorsch, S., & Vogelsang, A. (2025). LLMREI: Automating requirements elicitation interviews with LLMs. In *2025 IEEE 33rd International Requirements Engineering Conference (RE)* (pp. 19–30). IEEE. https://doi.org/10.1109/RE63999.2025.00013.
Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3), 213–236.
Li, J., He, C., Hu, J., Jia, B., Halevy, A. Y., & Ma, X. (2024). DiaryHelper: Exploring the use of an automatic contextual information recording agent for elicitation diary study. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642853.
Li, L., & Cai, H. (2025). Applying LLM-powered virtual humans to child interviews in child-centered design. In *Proceedings of the 24th Interaction Design and Children* (pp. 1140–1145). Association for Computing Machinery. https://doi.org/10.1145/3713043.3731551.
Liu, M., Wang, T., Cohen, C. A., Li, S., & Xiong, C. (2025). Understand User Opinions of Large Language Models via LLM-Powered In-the-Moment User Experience Interviews. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 13872–13893). https://doi.org/10.18653/v1/2025.findings-acl.714.
McDonagh, D., & Thomas, J. (2010). Rethinking design thinking: Empathy supporting innovation. *Australasian Medical Journal*, 3(8), 458–464. https://doi.org/10.4066/AMJ.2010.391.
Mountney, S. L., Gao, J. X., & Wiseall, S. (2007). A knowledge system to support manufacturing knowledge during preliminary design. *International Journal of Production Research*, 45(7), 1521–1537. https://doi.org/10.1080/00207540600942300.
Papneja, H., & Yadav, N. (2025). Self-disclosure to conversational AI: A literature review, emergent framework, and directions for future research. *Personal and Ubiquitous Computing*, 29(2), 119–151. https://doi.org/10.1007/s00779-024-01823-7.
Patton, M. Q. (2014). *Qualitative research & evaluation methods: Integrating theory and practice*. Sage publications.
Porter, S. R., Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New directions for institutional research*, 2004(121), 63–73. https://doi.org/10.1002/ir.101.
Robinson, O. C. (2023). Probing in qualitative research interviews: Theory and practice. *Qualitative Research in Psychology*, 20(3), 382–397. https://doi.org/10.1080/14780887.2023.2238625.
Taoka, Y., Tanaka, T., Nakatani, M., & Saito, S. (2024). Designers' perceptions of a sensor-enabled diary method for enhancing user research. *Proceedings of the Design Society*, 4, 2267–2276. https://doi.org/10.1017/pds.2024.229.
Tavory, I. (2020). Interviews and inference: Making sense of interview data in qualitative research. *Qualitative Sociology*, 43(4), 449–465. https://doi.org/10.1007/s11133-020-09464-x.
Van Berkel, N., Ferreira, D., & Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)*, 50(6), 1–40. https://doi.org/10.1145/3123988.
Velykoivanenko, L., Salehzadeh Niksirat, K., Teofanovic, S., Chapuis, B., Mazurek, M. L., & Huguenin, K. (2024). Designing a data-driven survey system: Leveraging participants' online data to personalize surveys. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642572.
Wilson, C. (2013). *Interview techniques for UX practitioners: A user-centered design method*. Newnes.
Xiao, Z., Zhou, M. X., Liao, Q. V., Mark, G., Chi, C., Chen, W., & Yang, H. (2020). Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(3), 1–37. https://doi.org/10.1145/3381804.

https://doi.org/10.1017/pds.2026.10583 Published online by Cambridge University Press