Mic Drop or Data Flop? Evaluating the Fitness for Purpose of AI Voice Interviewers for Data Collection within Quantitative & Qualitative Research Contexts

Shreyas Tirumala, Nishant Jain & Danny D. Leybzon
VKL Research, Inc.
San Francisco, CA
{shreyas,nishant,danny}@vkl.ai

Trent D. Buskirk
Professor, Provost Fellow of Data Science
Old Dominion University
Norfolk, Virginia
tbuskirk@odu.edu

Abstract

Transformer-based Large Language Models (LLMs) have paved the way for "AI interviewers" that can administer voice-based surveys with respondents in real-time. This position paper reviews emerging evidence to understand when such AI interviewing systems are fit for purpose for collecting data within quantitative and qualitative research contexts. We evaluate the capabilities of AI interviewers as well as current Interactive Voice Response (IVR) systems across two dimensions: input/output performance (i.e., speech recognition, answer recording, emotion handling) and verbal reasoning (i.e., ability to probe, clarify, and handle branching logic). Field studies suggest that AI interviewers already exceed IVR capabilities for both quantitative and qualitative data collection, but real-time transcription error rates, limited emotion detection abilities, and uneven follow-up quality indicate that the utility, use and adoption of current AI interviewer technology may be context-dependent for qualitative data collection efforts.

1 Introduction

1.1 Background

Since Vaswani et al. (2017)'s seminal work on transformers revolutionized natural language processing, Large Language Models (LLMs) have become increasingly commonplace tools to parse and produce human language. As such technologies have matured, survey methodologists, too, have begun to explore their potential (Lerner, 2024; Buskirk et al., 2025). One area of inquiry, in particular, involves understanding the feasibility of using artificial intelligence (AI) interviewers that leverage LLMs for data collection.

Automated systems have long been studied as a means of improving data quality while reducing costs. Years of social desirability bias research suggest that, particularly for sensitive topics, respondents may be more forthcoming in automated, self-administered surveys than in traditional interviews with other humans (Gribble et al., 1999; Cooley et al., 2000; Kreuter et al., 2008; Goldstein et al., 2025). The open question, however, is what logic such automated systems should use to elicit the highest quality responses.

Although early automated systems were more akin to rules engines that followed predefined logic to decide how to respond, more recent work has explored more expressive "chatbots" that use LLMs to communicate with respondents dynamically using a text-based chat interface. Kim et al. (2019) show that employing such technologies resulted in lower

rates of "satisficing" (selecting the easiest possible valid option) among respondents. Later work by Rhim et al. (2022), Chopra & Haaland (2023), and Barari et al. (2025) support these findings, with Rhim and colleagues observing that more conversational chatbots yielded higher levels of respondent self-disclosure – at the expense of slightly elevated social desirability bias.

In recent years, however, voice interfaces have received newfound attention. Even before the advent of transformer-based LLMs, audio interviews had been explored as a means of increasing self-disclosure (Lind et al., 2013; Lucas et al., 2014; DeVault et al., 2014). Voice is a particularly compelling modality because some respondents may be more forthcoming via voice than via web forms (Galasso et al., 2024). Additionally, audio-based interviews allow researchers to record an entire survey interview as it is conducted, which can be a useful strategy for quality control. Gomila et al. (2017) suggest that collecting and analyzing recordings helps mitigate response fraud, a technique unavailable for text-based surveys.

Voice interviews are not a panacea, however. Mode effects may affect the types of responses collected. For instance, respondents may be more likely to round numerical answers via voice compared to text-based input methods (Schober et al., 2015). Moreover, from an operational perspective, some types of questions are easier to communicate textually than orally; for example, multi-select questions (i.e., "Select all that apply") often involve large numbers of answer choices that are cumbersome to be heard than read. Such questions also pose data quality concerns. Some researchers have questioned whether respondents tend to over-index on the first or last answer choices presented in voice surveys (i.e., primacy or recency effects), especially as the number of choices provided increases (Le Bigot et al., 2013).

Recent work by Leybzon et al. (2025), Lang & Eskenazi (2025), and Wuttke et al. (2024) establishes that AI interviewers powered by LLMs can indeed conduct voice-based interviews. This begs a natural question: When and how should contemporary AI voice interviewers be used? In this work, we aim to address this question and examine the fitness for purpose of contemporary voice AI technologies for collecting data within both quantitative and qualitative research contexts.

1.2 Definitions

While the technical language we enumerate here may be familiar to computer scientists and survey researchers, we provide a selected list of terms and definitions to facilitate a common language for analysis.

1.2.1 Interactive Voice Response (IVR)

IVR is a technology that uses a set of pre-recorded audio messages to communicate with a user over the phone (Corkrey & Parkinson, 2002). Most IVR systems rely on touch-tone inputs (i.e., numbers dialed on a keypad) to record user selections and responses. Based on a user's inputs, the system selects its next response using pre-defined logic (i.e., a decision tree or flowchart). IVR is commonly used in both customer service and telephone survey contexts.

1.2.2 AI Interviewer

In this paper, we will use the term "AI interviewer" to refer to a system that leverages computational techniques to query, hear, understand, respond and interact with survey respondents via audio without relying entirely on preset scripts. This evaluation will focus on nascent audio-based systems rather than text-based chatbots.

Advanced contemporary systems typically comprise the following primary components as described here and illustrated in Figure 1:

- **Real-time Automated Speech Recognition (ASR):** Sometimes referred to as Speech-to-Text (STT), these systems transcribe respondent audio. ASR provides the "ears" of an AI interviewer, enabling them to hear respondents.
- **Large Language Models (LLMs):** Transformer-based LLMs are the "brains" of an AI interviewer; they interpret respondent input in order to make decisions mid-conversation and decide what to say back.
- **Speech Synthesis:** Also called "Text-to-Speech" (TTS), these technologies serve as the "mouth" of an interviewer system by producing human-sounding voices to speak back to a respondent based on text sent to it by an LLM.

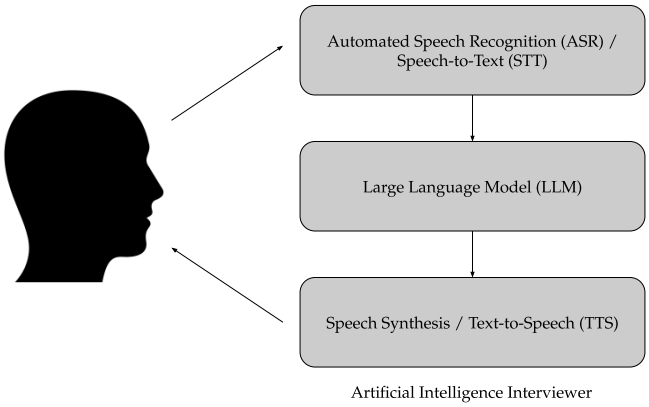

Figure 1: Standard AI interviewer architecture

As depicted in Figure 1, by connecting ASR systems to LLMs and speech synthesis technologies, a conversational AI agent can replicate a human's ability to ask questions, hear answers, and create appropriate responses. Although additional software infrastructure components (e.g., text output safeguards such as profanity filters, phone call deployment automation) are typically used to connect AI interviewer components together, we consider such technologies to be outside the scope of this analysis, as they do not meaningfully impact an AI interviewer's core capabilities.

Though preliminary work has explored the usage of models to generate output audio directly without an intermediate conversion to text (see Chu et al. (2023); Tang et al. (2024); Xu et al. (2025)), the feasibility of such speech-to-speech models in AI interviewer systems has not yet been successfully demonstrated. To date, such models introduce too much latency for real-time conversational contexts (Cui et al., 2025), though architectural improvements have been proposed in research settings to address this limitation (Défossez et al., 2024).

1.2.3 Quantitative Data Collection

Quantitative surveys typically involve a pre-defined list of closed-ended questions and pre-set lists of valid answer choices/criteria. Quantitative surveys often use multiple choice questions, rating scale questions (e.g., Likert), numerical questions, and binary yes/no

questions, but may also request some limited open-ended responses. Respondent data is typically aggregated and analyzed for quantitative results rather than qualitative themes (Groves et al., 2011).

1.2.4 Qualitative Data Collection

Qualitative data collection takes many forms, including in-depth interviews, focus groups, cognitive pre-testing, and more. Many methods of qualitative data collection also leverage a pre-defined list of questions or topics; however, they are more likely to feature longer open-ended questions that invite respondents to expand upon their beliefs, as well as unscripted follow-ups (Adams, 2015). Such methods have historically leveraged human judgment (e.g., emotional state evaluations of respondents, semantic evaluations of how vague or ambiguous answers are) to determine which topics require additional investigation mid-conversation. The importance of human intervention has made such conversations difficult to automate.

2 Evaluation Criteria

We will assess the fitness for purpose of IVR and AI interviewer technologies for data collection within both quantitative and qualitative research contexts. Specifically, using qualitative evaluations of fit, we examine whether the data collected by each system meets standard research objectives across two quality domains: Input/Output and Verbal Reasoning (Bosch & Maslovskaya, 2023). Table 1 lists key questions we evaluate.

Data Collection Context	**Quality Domain**	
	Input/Output: *Can a system adequately receive answers and reply back to respondents?*	**Verbal Reasoning:** *How intelligently can a system alter the conversation based on previous user input?*
Quantitative Data Collection: *Surveys with pre-set lists of questions and answers; these typically involve closed-ended questions*	Are questions asked and responses given faithful to the original survey design? Are answers collected faithful to a respondent's intent?	What types of conversational choices can each system make based on a respondent's answers (e.g., branching logic, clarifications, corrections)? How do the conversational choices that each system can make impact data quality?
Qualitative Data Collection: *Semi-structured interviews with several open-ended questions and unscripted follow-up questions*	How well can each system handle open-ended questions? Can each system identify emotional context and vary its output accordingly?	Can each system probe — that is, produce unscripted follow-up questions? If so, are these follow-up questions useful for improving response quality and depth?

Table 1: Comparison of input/output and verbal reasoning domains across data collection contexts

3 Capability Evaluations

In the following section, we will assess both IVR and AI interviewer systems. For each of the quality domains listed in Table 1, we will begin by providing an overview of each

system's capabilities followed by discussions of its fitness for purpose for both quantitative and qualitative research.

3.1 Interactive Voice Response (IVR)

To understand the functionality of more advanced AI interviewing systems, we must first understand the well-established technologies they are intended to replace. Automated phone-based surveys have been traditionally handled by Interactive Voice Response (IVR) systems.

3.1.1 Input/Output

Capability Overview: The most commonly used systems enable respondents to select answer choices by dialing numbers on a phone keypad via a touch-tone interface. These systems play back pre-recorded audio samples based on respondent selections. More recent systems supplement the touch-tone interface with ASR and leverage text-to-speech to varying degrees of success, though such systems are less common (Inam et al., 2017). In many cases, a recorded message must play back in its entirety before a response will be accepted.

Fitness for Purpose (Quantitative Data Collection): The touch-tone interface is straightforward to use for respondents, but poses an issue for survey researchers. Due to limitations of touch-tone interfaces, quantitative survey questionnaires must often be adapted for use with IVR systems. Question wordings are often changed to add additional instructions; for example, to ask about a respondent's gender identity, rather than asking "Do you currently identify as a man, a woman, or in some other way?", Amaya (2021) asked the following question instead: "Do you currently identify as a man? Press one. For a woman, press two, or for some other way, press three." Although such changes may be acceptable in some contexts, doing so may complicate comparative analysis between modes in mixed-mode projects.

Fitness for Purpose (Qualitative Data Collection): Structural challenges prevent effective qualitative data collection via IVR. Asking open-ended questions, for instance, requires researchers to perform post-hoc transcriptions of audio recordings of each open-ended response (Amaya, 2021). Focus groups and multi-party interviews are also not feasible with IVR systems as touch-tone interfaces are designed for one individual to use at a time. Given the pre-recorded nature of IVR systems, it is not possible to vary output back to a respondent based on emotional cues, though post-hoc analysis of voice recordings may be used to identify these after the survey.

3.1.2 Verbal Reasoning

Capability Overview: IVR systems typically have limited verbal reasoning capabilities; mid-conversation decision-making must be pre-defined with any desired follow-ups or responses recorded in advance (Kim et al., 2011).

Fitness for Purpose (Quantitative Data Collection): The lack of dynamic response capabilities in IVR systems inhibits their ability to conduct quantitative surveys. Respondents cannot usually clarify answer choices, correct previous responses, or repeat specific portions of a question (Kim et al., 2011). Unlike trained human interviewers, IVR systems also do not typically offer respondents encouragement mid-survey to facilitate completion (Amaya, 2021).

These limitations may impact data quality and completion rates. The inability to clarify or repeat portions of a question opens the door to order effects. Particularly for lengthy questions, respondents may forget answer choices, creating a bias toward the first or last few options spoken. Even after controlling for such order effects, Dillman et al. (2009) find that IVR respondents tend to give more extreme positive responses than other modes, and Amaya (2021) notes higher rates of both misreporting and item non-response.

Fitness for Purpose (Qualitative Data Collection): Because IVR systems lack the ability to dynamically react to a respondent's speech input, it is not possible to conduct any interviews that necessitate contextually-determined follow-up questions. Such functionality is crucial for in-depth interviews and cognitive pretesting. Accordingly, IVR systems have little utility for this specific type of qualitative data collection.

3.2 AI Voice Interviewers

Contemporary AI interviewers address many, but not all, of the issues raised by IVR. The verbal reasoning capabilities that LLMs provide to these systems enable follow-up questions and complex decisionmaking. In the last year, Leybzon et al. (2025), Lang & Eskenazi (2025), and Wuttke et al. (2024) have all been able to successfully and independently conduct quantitative voice surveys using AI interviewers. These works serve as examples of AI interviewer capabilities at this time and form the core of our analysis.

Because research in this field remains nascent and many questions are still unanswered, we augment our assessment of AI interviewers with insights drawn from the broader machine learning literature.

3.2.1 Input/Output

Capability Overview: Unlike the touch-tone interface of IVR systems, AI interviewers use ASR and speech synthesis to collect respondent input and vocalize responses.

Contemporary ASR systems have achieved relatively high-levels of accuracy. Kuhn et al. (2024) find that for state-of-the-art ASR systems, English word error rates hover around 5%, though heavily accented speech may decrease performance. Further research is needed to understand how error rates vary for other languages.

On the speech synthesis side, modern systems are advanced enough to reproduce accented English speech (Zhou et al., 2024). Many speech synthesis systems allow for AI interviewers to select voices on a per call basis. This opens the possibility for intelligent voice selection based on call-specific factors (e.g., individual respondent demographics). Though such experiments have not been conducted to date, work by Lilley et al. (2025) suggests that doing so may produce similar effects to varying which human interviewers administer surveys.

The accuracy of ASR and the sophistication of speech synthesis systems allow AI interviewers to react intelligently to respondents using voices that sound relatively human. Such systems have been shown to be robust enough to handle respondent interruptions and pauses mid-survey (Leybzon et al., 2025).

Two notable blind-spots for such systems, however, are emotion detection and expression. Because AI interviewers to-date have converted audio to text via ASR, the tone of the speaker must be inferred based on text. Even if emotion detection were to be implemented in such systems, Alhussein et al. (2025) find that existing AI-based emotion detection systems yield mixed results. Vocal expression of emotion is also difficult for modern speech synthesis systems. Though theoretical frameworks have been proposed to enable artificial voices to vary both emphasis and prosody, consumer-grade speech synthesis tools implementing these frameworks have not yet been studied in a survey context (Seshadri et al., 2021).

Fitness for Purpose (Quantitative Data Collection): The findings of Leybzon et al. (2025), Lang & Eskenazi (2025), and Wuttke et al. (2024) suggest that word error rates for modern ASR systems are at least sufficient to conduct quantitative surveys, however some caveats do exist.

Kuhn et al. (2024) observe that although word error rates are low for state-of-the-art systems, real-time/streaming transcription error rates can be significantly higher (˜10.9% on average). Indeed, Wuttke et al. (2024) and Leybzon et al. (2025) both note that respondents reacted to minor speech recognition errors mid-survey. However, problems with erroneous transcriptions can be somewhat mitigated by post-processing recorded audio after the survey is over.

Speech synthesis also appears to be adequate for quantitative use cases – no AI interviewer study to date has remarked on specific issues associated with TTS systems, even with their limited ability to express emotion or emphasis. Importantly, in comparison to IVR, an AI interviewer can recite question text as-is, without requiring wording changes to suit the medium; that is, both human and AI interviewers can ask the same set of questions.

Fitness for Purpose (Qualitative Data Collection): High streaming transcription error rates and a lack of emotion detection/expression pose issues for qualitative data collection. Unlike quantitative data collection methods, qualitative research often involves a high proportion of open-ended questions that lack pre-defined lists of valid responses. Without a list of valid responses to check against, transcription errors cannot easily be corrected. This is especially problematic because open-ended questions also typically yield higher average response lengths than closed-ended multiple choice questions; for instance, asking respondents to elaborate on their political beliefs requires more explanation than asking them to classify their political party affiliations. Longer responses increase the expected number of transcription errors per survey. Further complications arise in focus groups and multi-party interviews as ASR performance declines when two or more speakers are audible (Addlesee et al., 2020).

Emotion detection and expression are also more important in qualitative data collection contexts than they are in quantitative contexts. Building rapport is crucial for maximizing self-disclosure in qualitative interviews, and AI interviewers may need emotion detection and expression skills to do so (Sun et al., 2021). A respondent's emotional state also serves as an important input for verbal reasoning; without the emotional context of a response, the quality of follow-up questions generated by an AI interviewer may be lower. It remains to be seen whether AI interviewers can understand a respondent's emotional state enough to vary intonation, emphasis, and follow-up questions effectively.

3.2.2 Verbal Reasoning

Capability Overview: Work by Zeng et al. (2023) suggests that the verbal reasoning capabilities of LLMs were enough to navigate the conditional branching and termination logic of a semi-structured interview. Leybzon et al. (2025) demonstrate this in the field by using an LLM-powered AI interviewer to administer the SSRS Opinion Panel Omnibus, a 123-question survey that contained a mix of closed-ended questions (multiple-choice, Likert scale, yes/no) and open-ended questions, along with conditional branching/termination logic. This survey took respondents approximately 30 minutes to complete. Both Lang & Eskenazi (2025) and Wuttke et al. (2024) also find, independently, that an AI agent was able to successfully navigate shorter survey instruments without major issues. Notably, AI interviewers can respond to requests for clarification or repetition of questions and answer choices, as well as apply branching logic based on the semantic content of open-ended responses. Such features are unavailable in IVR systems.

LLMs also display contextual awareness. When presented with unclear or ambiguous answers, AI interviewers have been shown to re-prompt respondents intelligently. For example, Lang & Eskenazi (2025) find that AI interviewers could detect and respond to "numeric answer[s] that [were] out of [a valid] range" and Leybzon et al. (2025) note that AI interviewers can be instructed to re-prompt respondents who answer with just "Liberal" when answer choices are "Somewhat Liberal" or "Very Liberal." Beyond re-prompting, some research has been conducted on the ability of LLMs to generate intelligent follow-up questions with inconclusive assessments of question quality and utility (see Wei et al. (2024), Geisen (2024), Kuric et al. (2024), Cuevas et al. (2024), and Chopra & Haaland (2023)). LLMs can also contextually introduce humor, small talk, and affirmation, similar to a human interviewer, though respondent feedback is mixed (Yun et al., 2023).

Fitness for Purpose (Quantitative Data Collection): Leybzon et al. (2025), Lang & Eskenazi (2025), and Wuttke et al. (2024) all observe that their AI interviewers successfully conducted quantitative surveys with conditional branching logic. Leybzon and colleagues also highlight the AI interviewer's capability to randomize both question order and answer choice order. Although LLMs have been noted to hallucinate (i.e., respond with "plausible, but non-factual content", see Huang et al. (2025)), to date, none of the three AI voice interviewer

studies report any instances of the interviewer inventing questions/answer choices or asking questions out of order. This suggests that AI interviewers perform at least as well as IVR systems with respect to navigating survey instruments.

AI interviewers also have additional verbal reasoning capabilities that may improve data quality. AI interviewers can facilitate clarifications and corrections upon request. Even without a respondent request, it appears that LLMs are capable of basic data integrity validation. As mentioned previously, AI interviewer systems can flag ambiguous responses and answers outside of pre-defined ranges in real-time, allowing them to re-prompt users for valid data.

Fitness for Purpose (Qualitative Data Collection): Qualitative research is often highly interactive; as we have previously highlighted, contextually aware follow-ups are often crucial for collecting high quality data. Though research has shown that LLMs can generate follow-up questions, the quality of these questions is unclear. Cuevas et al. (2024) find that LLMs tend not to probe enough to identify personalized examples or understand a respondent's specific motivations. Kuric et al. (2024) corroborate these findings in a user experience research context, observing that OpenAI's GPT-4 model performs worse at eliciting in-depth feedback than a static, pre-curated list of follow-ups created prior to the conversation. Conversely, Geisen (2024) and Chopra & Haaland (2023) both suggest that LLM-generated follow-up questions can improve answer depth. Given these varying perspectives, additional investigation may be required to understand follow-up question quality.

LLMs may be able to be tuned to improve follow-up question quality. Wei et al. (2024) show that different system prompts can dramatically impact interviewer behavior and question generation. It is possible that because most ASR systems do not perform emotion detection, improvements to the emotional context fed as input to LLMs could improve performance.

Aside from follow-up questions, further research is necessary to understand whether AI interviewers can effectively build rapport with respondents. Current literature does not yet explore this topic.

4 Conclusion

The advent of AI interviewers raises important questions about when and how to leverage these tools. AI interviewers already appear to be able to handle quantitative survey data collection and provide more functionality than current IVR systems while doing so. Early work suggests that they may be able to collect richer qualitative insights than IVR systems as well.

Our review of current research suggests a few circumstances when AI interviewers may be particularly useful:

- Human interviewer availability is limited (e.g., off hours)
- Follow-up questions are not pivotal (i.e., limited complexity of probing needed)
- Post-processing of respondent data (e.g., audio recordings) is acceptable (to minimize transcription errors)
- Emotion detection and expression are not central to the study
- Sensitive topics (e.g., money, sexuality) are involved that may trigger social desirability bias in responses to a human interviewer

In such scenarios, AI interviewers may be able to provide quantitative and qualitative insights cost-effectively.

4.1 Future Research

We are still in the early days of AI interviewer research. Although insights are beginning to emerge about the capabilities of AI interviewers, significantly more work is needed to

measure their performance – especially in relation to human interviewers. We note several key areas for further research:

- **Respondent Representativeness:** Most early work to date has been conducted on college/graduate students. It is unclear whether results about response rates and respondent experience will generalize to a general population. For example, an open question remains about whether higher-levels of antipathy towards AI will emerge as this technology becomes widespread. We still do not know which types of respondents tend to prefer AI interviewers, nor which contexts or topics such preferences are most salient for.
  - Leybzon et al. (2025) experiment with a probability-based panel meant to be representative of the US population, but sample sizes are not large enough to draw conclusive inferences across specific subpopulations.
- **Respondent Preferences:** Early signs are encouraging that respondents are willing to speak with AI interviewers. Leybzon et al. (2025) find most respondents to be at least equally comfortable speaking with AI versus a human, particularly for sensitive topics, though more research is necessary to confirm this hypothesis. Interestingly, Wuttke et al. (2024) find that respondents may have a lower tolerance for conversational latency when made aware that they were interacting with an AI interviewer; some respondents actively noted the time necessary for responses to be generated.
- **Methodological Variation:**
  - **Inbound vs Outbound Campaigns:** Most studies to date involve outbound phone call campaigns with limited inbound options. Lang & Eskenazi (2025) did offer respondents an option to connect to a web-based call on-demand, but further investigation is necessary to see whether the act of opting-into a survey meaningfully changes results. Offering voice-based interview options alongside traditional web form surveys or SMS campaigns has also not yet been fully explored.
  - **Interviewer Attributes:** Respondents may have different responses to variations in interviewer behavior (e.g., use of humor, small talk, and affirmation) as well as which accents/voices are used (see Yun et al. (2023), Lilley et al. (2025)). The impact of such changes on respondent experience feedback remains to be seen.

We acknowledge that the speed at which AI technologies have developed has been rapid; it is possible that the capabilities of AI interviewers will progress quickly in the coming months and years. Investigation of this exciting new modality will be crucial to maintain methodological rigor and inform judgments on the many ethical issues that are sure to arise from this technology's widespread usage.

References

William C Adams. Conducting semi-structured interviews. *Handbook of practical program evaluation*, pp. 492–505, 2015.

Angus Addlesee, Yanchao Yu, and Arash Eshghi. A comprehensive evaluation of incremental speech recognition and diarization for conversational ai. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3492–3503, 2020.

Ghada Alhussein, Ioannis Ziogas, Shiza Saleem, and Leontios J Hadjileontiadis. Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis. *Artificial Intelligence Review*, 58(7):198, 2025.

Ashley Amaya. How call-in options affect address-based web surveys. *Pew Re-search Center*, 2021.

Soubhik Barari, Jarret Angbazo, Natalie Wang, Leah M Christian, Elizabeth Dean, Zoe Slowinski, and Brandon Sepulvado. Ai-assisted conversational interviewing: Effects on data quality and user experience. *arXiv preprint arXiv:2504.13908*, 2025.

Oriol J Bosch and Olga Maslovskaya. The utility of probability-based online surveys: a literature review. In *2023 AAPOR Webinar Series*. AAPOR, 2023.

Trent D Buskirk, Adam Eck, Leah von der Heyde, and Florian Keusch. Gpt, pretend you are a survey researcher: Results from a systematic literature review exploring the use of large language models within survey research. In *80th Annual AAPOR Conference*. AAPOR, 2025.

Felix Chopra and Ingar Haaland. Conducting qualitative interviews with ai. CESifo Working Paper 10666, Center for Economic Studies and ifo Institute (CESifo), Munich, September 2023. URL https://www.ifo.de/DocDL/cesifo1_wp10666.pdf.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023. URL https://arxiv.org/abs/2311.07919.

Philip C Cooley, Heather G Miller, James N Gribble, and Charles F Turner. Automating telephone surveys: using t-acasi to obtain data on sensitive topics. *Computers in Human Behavior*, 16(1):1–11, 2000.

Ross Corkrey and Lynne Parkinson. Interactive voice response: review of studies 1989–2000. *Behavior Research Methods, Instruments, & Computers*, 34(3):342–353, 2002.

Alejandro Cuevas, Jennifer V. Scurrell, Eva M. Brown, Jason Entenmann, and Madeleine I. G. Daepp. Collecting qualitative data at scale with large language models: A case study, 2024. URL https://arxiv.org/abs/2309.10187.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. Recent advances in speech language models: A survey, 2025. URL https://arxiv.org/abs/2410.03751.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pp. 1061–1068, 2014.

Don A Dillman, Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell, Jodi Berck, and Benjamin L Messer. Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (ivr) and the internet. *Social science research*, 38(1):1–18, 2009.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue, 2024. URL https://arxiv.org/abs/2410.00037.

Vincenzo Galasso, Tommaso Nannicini, and Debora Nozza. We need to talk: Audio surveys and information extraction. Technical report, CESifo Working Paper, 2024.

Emily Geisen. Prompting insight: Enhancing open-ended survey responses with ai-powered follow-ups. In *79th Annual AAPOR Conference*. AAPOR, 2024.

Risë B Goldstein, Megan A Hendrich, Randall K Thomas, and Robert Petrin. Digging deep: How diverse identities are associated with different survey mode preferences for adverse, traumatic experiences. In *80th Annual AAPOR Conference*. AAPOR, 2025.

Robin Gomila, Rebecca Littman, Graeme Blair, and Elizabeth Levy Paluck. The audio check: A method for improving data quality and detecting data fabrication. *Social Psychological and Personality Science*, 8(4):424–433, 2017.

James N Gribble, Heather G Miller, Susan M Rogers, and Charles F Turner. Interview mode and measurement of sexual behaviors: Methodological issues. *Journal of Sex research*, 36 (1):16–24, 1999.

Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey methodology*. John Wiley & Sons, 2011.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Itorobong A Inam, Ambrose A Azeta, and Olawande Daramola. Comparative analysis and review of interactive voice response systems. In *2017 Conference on Information Communication Technology and Society (ICTAS)*, pp. 1–6. IEEE, 2017.

Hee-Cheol Kim, Deyun Liu, and Ho-Won Kim. Inherent usability problems in interactive voice response systems. In *Human-Computer Interaction. Users and Applications: 14th International Conference, HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings, Part IV 14*, pp. 476–483. Springer, 2011.

Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 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, 2019.

Frauke Kreuter, Stanley Presser, and Roger Tourangeau. Social desirability bias in cati, ivr, and web surveys: The effects of mode and question sensitivity. *Public opinion quarterly*, 72 (5):847–865, 2008.

Korbinian Kuhn, Verena Kersken, Benedikt Reuter, Niklas Egger, and Gottfried Zimmermann. Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing*, 16(4):1–23, 2024.

Eduard Kuric, Peter Demcak, and Matus Krajcovic. Unmoderated usability studies evolved: Can gpt ask useful follow-up questions? *International Journal of Human–Computer Interaction*, pp. 1–18, 2024.

Max Lang and Sol Eskenazi. Telephone surveys meet conversational ai: Evaluating a llm-based telephone survey system at scale. In *80th Annual AAPOR Conference*. AAPOR, 2025.

Ludovic Le Bigot, Loïc Caroux, Christine Ros, Agnès Lacroix, and Valérie Botherel. Investigating memory constraints on recall of options in interactive voice response system messages. *Behaviour & Information Technology*, 32(2):106–116, 2013.

Joshua Lerner. The promise & pitfalls of ai-augmented survey research. NORC research library, Oct 2024. URL https://www.norc.org/research/library/promise-pitfalls-ai-augmented-survey-research.html. Accessed on June 20, 2024.

Danny D. Leybzon, Shreyas Tirumala, Nishant Jain, Summer Gillen, Michael Jackson, Cameron McPhee, and Jennifer Schmidt. Ai telephone surveying: Automating quantitative data collection with an ai interviewer. In *80th Annual AAPOR Conference*. AAPOR, 2025.

Kevin D Lilley, Ellen Dossey, Michelle Cohn, Cynthia G Clopper, Laura Wagner, and Georgia Zellou. Social evaluation of text-to-speech voices by adults and children. *Speech Communication*, 166:103163, 2025.

Laura H Lind, Michael F Schober, Frederick G Conrad, and Heidi Reichert. Why do survey respondents disclose more when computers ask the questions? *Public opinion quarterly*, 77 (4):888–935, 2013.

Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100, 2014.

Jungwook Rhim, Minji Kwak, Yeaeun Gong, and Gahgene Gweon. Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality. *Computers in Human Behavior*, 126:107034, 2022.

Michael F Schober, Frederick G Conrad, Christopher Antoun, Patrick Ehlen, Stefanie Fail, Andrew L Hupp, Michael Johnston, Lucas Vickers, H Yanna Yan, and Chan Zhang. Precision and disclosure in text and voice interviews on smartphones. *PloS one*, 10(6): e0128337, 2015.

Shreyas Seshadri, Tuomo Raitio, Dan Castellani, and Jiangchuan Li. Emphasis control for parallel neural tts. *arXiv preprint arXiv:2110.03012*, 2021.

Hanyu Sun, Frederick G Conrad, and Frauke Kreuter. The relationship between interviewer-respondent rapport and data quality. *Journal of Survey Statistics and Methodology*, 9(3): 429–448, 2021.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024. URL https://arxiv.org/abs/2310.13289.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. Leveraging large language models to power chatbots for collecting user self-reported data. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–35, 2024.

Alexander Wuttke, Matthias Aßenmacher, Christopher Klamm, Max M Lang, Quirin Würschinger, and Frauke Kreuter. Ai conversational interviewing: Transforming surveys with llms as adaptive interviewers. *arXiv preprint arXiv:2410.01824*, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Hye Sun Yun, Mehdi Arjmand, Phillip Sherlock, Michael K Paasche-Orlow, James W Griffith, and Timothy Bickmore. Keeping users engaged during repeated administration of the same questionnaire: Using large language models to reliably diversify questions. *arXiv preprint arXiv:2311.12707*, 2023.

Jie Zeng, Yukiko I Nakano, and Tatsuya Sakato. Question generation to elicit users' food preferences by considering the semantic content. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 190–196, 2023.

Xuehao Zhou, Mingyang Zhang, Yi Zhou, Zhizheng Wu, and Haizhou Li. Accented text-to-speech synthesis with limited data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1699–1711, 2024.