On Replacing Humans with Large Language Models in Voice-Based Human-in-the-Loop Systems

Shih-Hong Huang, Ting-Hao 'Kenneth' Huang

College of Information Sciences and Technology, The Pennsylvania State University
201 Old Main, University Park, PA 16802, USA
{szh277,txh710}@psu.edu

Abstract

It is easy to assume that Large Language Models (LLMs) will seamlessly take over applications, especially those that are largely automated. In the case of conversational voice assistants, commercial systems have been widely deployed and used over the past decade. However, are we indeed on the cusp of the future we envisioned? There exists a social-technical gap between what people want to accomplish and the actual capability of technology. In this paper, we present a case study comparing two voice assistants built on Amazon Alexa: one employing a human-in-the-loop workflow, the other utilizes LLM to engage in conversations with users. In our comparison, we discovered that the issues arising in current human-in-the-loop and LLM systems are not identical. However, the presence of a set of similar issues in both systems leads us to believe that focusing on the interaction between users and systems is crucial, perhaps even more so than focusing solely on the underlying technology itself. Merely enhancing the performance of the workers or the models may not adequately address these issues. This observation prompts our research question: What are the overlooked contributing factors in the effort to improve the capabilities of voice assistants, which might not have been emphasized in prior research?

Introduction

For a long time, the argument for (nearly) real-time crowd-powered systems—employing online workers to operate computer components for quick judgments or predictions, mimicking actual computer components—was to fill the social-technical gaps of current automatic technologies, enabling researchers to study interaction problems or user needs that can only be studied when particular systems or services exist. For example, Huang, Chang, and Bigham suggested in Evorus (Huang, Chang, and Bigham 2018), a crowd-powered conversational assistant:

> Such automated [general conversational assistant] systems only become useful once they can completely take over from the crowd-powered system. Such abrupt transition points mean substantial upfront costs must be paid for collecting training examples before any automation can be tested in an online system. ... we explore an alternative approach of a crowdpowered system architecture that supports gradual automation over time. (Huang, Chang, and Bigham 2018, p. 1-2)

This was essentially based on the stance: "AI is not there (yet); let's use humans for now". A classic example is VizWiz (Bigham et al. 2010), a nearly real-time crowd-powered system that uses online crowd workers to quickly respond to visual questions sent from blind users or those with visual impairments. Way before AI could reliably answer arbitrary visual questions, VizWiz enabled large-scale research into the visual challenges that blind users encounter daily. Other impactful examples include Chorus (Lasecki et al. 2013), Soylent (Bernstein et al. 2010), and Scribe (Lasecki et al. 2012). A series of techniques and tools were created to make these systems possible, such as the retainer model (Bernstein et al. 2011) and quikturkit (Bigham et al. 2010), or even scalable (Huang and Bigham 2017). Some projects, such as Evorus (Huang, Chang, and Bigham 2018) and VizWiz, took a step further to argue that the insights gained from studying these imminent problems could, in turn, improve today's computer technology.

Such arguments were, unspokenly and probably subtly, challenged by the recent rise of large language models (LLMs). Evidence has shown that ChatGPT exceeds the performance of online crowd workers in labeling and writing tasks (Törnberg 2023; Gilardi, Alizadeh, and Kubli 2023; Cegin, Simko, and Brusilovsky 2023); many crowd workers use ChatGPT for their tasks. A recent study showed that aggregating 20 crowdsourced labels from a well-executed, carefully designed MTurk pipeline still falls short of GPT-4's labeling accuracy (He et al. 2024). Some even view LLMs as a precursor to Artificial General Intelligence (AGI), which can achieve or surpass human performance in millions of tasks (Bubeck et al. 2023). All of these, collectively, hint that LLMs might be closing the social-technical gaps that justified the creation of real-time crowd-powered systems more effectively than human crowd workers. If true, researchers could replace crowd workers with LLMs in these systems, significantly reducing engineering efforts and financial costs while still being able to study the same set of interactive problems and user needs. In fact, some studies have directly compared crowdsourcing workflows with LLM workflows, seeking to identify what elements of crowdsourcing can be adapted to LLM workflows and what

cannot (Wu et al. 2023; Grunde-McLaughlin et al. 2023).

This paper presents an interesting case study to offer our viewpoint within this discussion. We compared two systems built on top of Amazon's Echo, a commercial smart speaker integrated with Amazon Alexa. One, created by us, involved humans in the loop to hold the conversation (Huang et al. 2022), and the other, created by a team of researchers from Johns Hopkins University and Northeastern University, used an LLM to hold the conversation behind the scenes (Mahmood et al. 2023). Through this comparison, we argue that while last-mile interaction issues such as conversation cut-offs and speech recognition problems persist across both human-powered and LLM workflows, each approach requires distinct considerations. Specifically, certain accommodations must be made when replacing human workflows with LLMs. LLMs may introduce new challenges that are rarely encountered by human workers, and vice versa.

A Tale of Two Systems

In this paper, we observe and compare two similar systems: one powered by humans and the other by an LLM.

ECHOPAL: A Human-in-the-Loop Voice Assistant

ECHOPAL is a human-in-the-loop voice assistant based on Amazon Echo introduced by Huang et al. in 2022 (Huang et al. 2022). When a user talks to ECHOPAL, Echo records the audio and turns the speech into text through the built-in automatic speech recognition (ASR) system. The transcribed text is then sent to the backend of ECHOPAL and presented to the human worker as a message in the worker interface, where the worker can see the transcribed message and a set of possible alternative transcriptions generated by the system. Furthermore, ECHOPAL also automatically generates and presents a list of suggested responses to the worker. As each alternative transcription has its own suggested responses, the worker can click on the transcription to switch between them. Search support sites, such as Google Search or Google Weather, are also provided in the interface. With the support of these technological features, the worker was required to produce each response within 25 seconds. The worker's response is then sent back to the Echo device, where a built-in text-to-speech (TTS) system reads out the message to the user. Users were allowed to chat freely with ECHOPAL without constraints on the conversation topic.

ECHOPAL inherited the spirit of Chorus and Evorus, which aimed to use human intelligence to fill the gaps between what users want to use conversational assistants for and what conversational assistants are capable of, allowing researchers to study interaction questions that could not otherwise be studied (Lasecki et al. 2013; Huang, Chang, and Bigham 2018). In the user study, Huang et al. used participants as both users and workers (instead of hiring online workers to be the workers.)

CHATGPT-IN-ECHO: An LLM-in-the-Loop Voice Assistant

CHATGPT-IN-ECHO was presented by Mahmood et al. in 2023, with the aim of exploring the constraints and opportunities of LLM-powered voice assistants (Mahmood et al. 2023).[1] In the same period of time, many other projects also appeared to connect LLMs, especially ChatGPT provided by OpenAI via API, with voice assistants such as Alexa or Google Assistant (Yang et al. 2024; AndroidAuthority 2023).

When a user talks to CHATGPT-IN-ECHO, Echo records the audio and turns the speech into text through the built-in automatic speech recognition (ASR) system. The transcribed text is then sent to a primary middleman API; this middleman API is used to handle the time-out constraint set by Alexa Skill. If it does not receive a message from ChatGPT within a certain amount of time, it will initiate filler/small talk to maintain Alexa Skill activity. A secondary middleman API is used to communicate with ChatGPT. Upon receiving the relayed user transcription from the first middleman API, it will request ChatGPT via API call with the latest user message. A shared database of conversation history and progress achieved communication between the primary and secondary middleman API. Once the secondary middleman API receives the ChatGPT response, it will update the conversation history. The primary middleman API will monitor the database for conversation updates and send the ChatGPT message to the user through Alexa TTS. Three tasks from different scenarios were tested: medical self-diagnosis, creative planning, and discussion with opposing stances.

What Makes Them Comparable

ECHOPAL and CHATGPT-IN-ECHO were both run on Amazon's Echo devices, a commercial voice-enabled smart speaker designed to assist users via voice. In the original study, ECHOPAL ran on Echo (2nd generation), and CHATGPT-IN-ECHO ran on Echo Dot. Both devices had no monitors and used voice as the only communication channel with users, aided with a ring light to show the state of the device, *e.g.*, listening, processing, or deactivated. Note that Alexa Skill did not (and still does not) support audio streaming and only provided the transcribed text of the users' speech to the backend for developers. The transcribed texts were fed directly to the backend without utilizing the intent architecture of Alexa Skill. The conversation logs between the user and the voice assistants were stored in the database hosted outside the Alexa Skill.

ECHOPAL and CHATGPT-IN-ECHO are similar also in terms of schematic. Both utilized behind-the-scenes architecture to handle the conversation history, and Alexa Skills interacted with the backend responder regarding the user message. ECHOPAL had human workers—hence the need for a worker interface—and helper functions to help the workers. CHATGPT-IN-ECHO utilized LLM and therefore needed to handle the API call.

Alexa API Time-Out. To our knowledge, the current default time limit for Alexa Skills to process responses is about

[1]Mahmood et al. did not name their system. We used CHATGPT-IN-ECHO as a placeholder name for communication purposes.

8 seconds, which is fixed for most cases. CHATGPT-IN-ECHO employed middleman APIs to handle the time-outs by inserting fillers and small talks before the 8-second time-out occurred and continuing the conversation after responses were received from ChatGPT. ECHOPAL, on the other hand, does not face the same issue because it was developed under an earlier version of Alexa Skill, and extending the time-out value to 25 seconds was possible. However, we believe that it is possible to modify ECHOPAL's architecture to fit the current Alexa Skill configuration and adopt techniques used by CHATGPT-IN-ECHO.

Humans vs. LLMs Comparison

We are the authors of ECHOPAL (Huang et al. 2022), and the following comparison was made by us reading through the paper of CHATGPT-IN-ECHO in detail (Mahmood et al. 2023). We also looked into our experimental records to compare the details, which might include some nuanced information that we did not mention in our original paper.

Humans' Problems That LLMs Do Not Have

Long Response Latency. One of the main drawbacks for ECHOPAL is the long latency for the users, where they have to wait for an average of 17.68 seconds to get quality responses from Alexa. However, the same response time was considered too short for the workers at the backend to provide quality responses. Most of this latency comes from human workers needing to type; some latency comes from ASR and internet transmission time. Meanwhile, CHATGPT-IN-ECHO is powered by GPT-3.5, which can produce thousands of words in seconds.

To some extent, integration with LLMs might make future voice assistants that aim to hold longer, open-ended conversations one step closer to reality: it reduces the unreasonably long latency of human-powered voice assistants, it makes such systems more scalable and affordable to build than recruiting human workers, and it bypasses the speed limits set by human typings.

LLMs' Problems That Humans Do Not Have

Oversharing and Repetitiveness. CHATGPT-IN-ECHO raised concerns regarding voice assistants providing repetitive information and oversharing, which was not observed in ECHOPAL's study. Repetitive information provided by ChatGPT was reported in the scenarios tested by CHATGPT-IN-ECHO. Since voice users cannot review previous conversations compared to text conversations, repetitive information causes further inconvenience during the conversation. The phenomenon was particularly true in medical-related conversations; repetitive information can occur even if ChatGPT was prompted not to repeat certain warning information. OpenAI's policy likely causes this, and it might not be easily fixable by outside developers. Additionally, CHATGPT-IN-ECHO encountered a challenge with ChatGPT's oversharing, where the responses provided by ChatGPT was too overwhelming for users to absorb. Even when prompted to give brief responses under 100 words, the higher information density in text interaction, which ChatGPT is based on, can overwhelm users in a voice setup.

Problems That Both LLMs and Humans Have

Cut-Offs. We define cut-offs as Alexa stopping listening to the users' speech while users are in the middle of their speech or intend to keep speaking. In ECHOPAL, cut-offs were attributed to (1) users having longer pauses between words, causing Alexa to think the user's speech ended, and (2) Alexa entered listening mode after its own speech and stopped listening if users did not speak in time. CHATGPT-IN-ECHO faced similar problems, such as (1) partial listening, where partial speech was captured in users' speech, (2) interruption, where Alexa interrupts the users' speech mid-sentence, and (3) pauses between words when the users organize thoughts.

ASR Errors. ASR errors were one of the main obstacles in ECHOPAL, since the audio of users' speech was not obtainable and only the transcribed text by Alexa was received. Different alternative transcriptions were provided to help human workers to tackle this problem. CHATGPT-IN-ECHO also faced ASR problems but had the inherent ChatGPT recovery mechanism, such as apologizing or asking the user to clarify. We suspect the typo fixing of the text-based system, which ChatGPT is good at, does not work as well for voice-based communication. This is likely because similar sounds do not translate to good guessing by ChatGPT.

Conversation Breakdowns. Both Cut-offs and ASR errors can cause the interaction to break down, meaning the current interaction session will either be (1) terminated by Alexa without the users' acknowledgment or (2) require initiating a recovery mechanism in order to continue the interaction. The only solution to address the first type of breakdown is for the user to re-initiate the Alexa Skill, effectively restarting the interaction. One way to make the re-initiation easier is by better maintaining the previous interaction history, but it still requires the user to restart the interaction actively. For the second scenario, recovery within the same interaction session is possible. Recovery can be initiated by the users or by the voice assistants. In the first category of breakdown, users will need to restart the interaction with the voice assistants actively. In the second category, the voice assistants can assist in the recovery process within the same session. However, most recovery happens when users repeat the previous sentence that failed to be correctly processed or move on to the next sentence and discard the failed sentence.

One feature of echo devices is the light ring that indicates the state that Alexa is in. When the device is in "listening" mode, the light ring will illuminate and shimmer. ECHOPAL reported that users' familiarity level with Echo devices' interaction pattern significantly impacts their ability to handle cut-offs. Those who are more familiar with the Echo interaction patterns encountered fewer cut-offs.

CHATGPT-IN-ECHO described users' mental model while interacting with voice assistants in multi-turn scenarios. They suggested that actions users take after conversation breakdowns indicated the gap between users' mental models and voice assistants' capabilities. The light ring on the device was not discussed in CHATGPT-IN-ECHO's paper.

Long responses are not suitable. ECHOPAL suggested that long text responses are not suitable for voice application since it takes a lot of time to read out the full paragraph and interrupt the flow of voice conversation. CHATGPT-IN-ECHO on the other hand limited the response of ChatGPT to 100 words in prompt but still face the problem of voice assistant oversharing and overwhelming the users.

Discussion

Time vs. Quality Trade-offs

We identified a trade-off between the quality of responses provided by ECHOPAL and the time required to generate the response. In order to provide higher-quality responses, workers needed more time. Participants who were assigned to be the workers that operated ECHOPAL suggested that the 25-second response time was not enough. Enforcing a shorter response time would likely lead to a decrease in response quality. Grunde-McLaughlin et al. proposed the design space for LLM chaining by adapting crowdsourcing workflows (Grunde-McLaughlin et al. 2023). The trade-off between time and quality we observed in ECHOPAL also exists in this design space.

For LLM-based systems such as CHATGPT-IN-ECHO, which utilized API calls instead of waiting for human workers to provide responses, we speculate such trade-offs to be less prevalent.

Effort Put in Prompt Engineering

While LLM workflow can offer lower latency and cost compared to crowdsourcing workflow, efforts need to be directed toward prompting LLMs to generate more diverse responses. While it is possible to get diversified responses by querying crowd workers multiple times, simply prompting LLMs multiple times can generate responses with less diversity. Additionally, instructions and subtasks suitable for crowdsourcing workflows may be overly complicated and may not suit LLMs. Therefore, simpler subtasks and instructions are required to utilize LLM capabilities effectively.

Design Implication

We propose that the **interaction aspect for voice assistants' workflows should be emphasized** while the workflow is being designed. How users interact with the system and their situations play important roles. For example, ECHOPAL and CHATGPT-IN-ECHO both had users sit in front of the echo device alone in a quiet room. Considering real-life scenarios, it is likely that users will interact with the voice assistants in less than ideal situations, namely (1) where background noise and external interruption of the conversation are present compared to in-lab studies and (2) visual cues such as on Echo devices (light ring) can be less obvious or does not apply, especially in voice only interaction such as phone calls.

Echo is Not Designed for Long Conversations. The enforced eight-second timeout for Alexa Skills makes performing tasks with longer processing time hard and requires workarounds. Alexa Skills are structured with specific tasks in mind. Freestyle speech without a predefined topic or speech structure does not integrate well with the current Alexa infrastructure. Longer speeches are also not favored by Alexa, partly due to its strictly turn-based nature, where only one side can speak at a time. From the users' perspective, it is unclear when the device will stop "listening"; thus, making longer speeches feels less secure and more prone to cut-offs. There is also no means for users to edit their speech after Alexa stops listening; all edits can only happen after Alexa replies to the message and begins listening again in the next turn. Combined with the lack of indication for users regarding the actual transcribed text, it becomes challenging to backtrack and edit previous speech, especially if it is lengthy and difficult to pinpoint where the edit is needed. From Alexa's standpoint, the only way to deliver a message is through TTS. However, not all messages are suited to be delivered in one long speech, as users may lose focus in such scenarios.

Conclusion

We presented a comparison between two Amazon Alexa-based voice assistants, ECHOPAL and CHATGPT-IN-ECHO, where ECHOPAL utilized human-in-the-loop workflow and CHATGPT-IN-ECHO utilized LLM to generate responses to user inquiries. Prior research suggested that crowdsourcing workflow could be adapted by LLM chaining workflow, which supported our comparison. Regarding the use case, ECHOPAL let users chat with it freely, and CHATGPT-IN-ECHO predefined a set of scenarios and tasks that needed to be accomplished. We concluded that there are problems troublesome for ECHOPAL, such as latency, and problems present for CHATGPT-IN-ECHO, such as repetitiveness and oversharing. At the same time, there is also a set of problems that were present for both workflows, namely the cut-offs, breakdowns, and ASR errors. We consider the above shared problems for human-in-the-loop and LLM workflow to be interaction-related, and simply treating them as an LLM model problem or crowdsourcing problem does not help solve them.

References

AndroidAuthority. 2023. Supercharge your Google Nest Mini with AI by swapping out its brains. https://www.androidauthority.com/google-nest-mini-unofficial-ai-experiment-3346448/. Accessed: 2024-04-10.

Bernstein, M. S.; Brandt, J.; Miller, R. C.; and Karger, D. R. 2011. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 33–42. ACM.

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 313–322. ACM.

Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; et al. 2010. VizWiz: nearly real-time answers to visual questions.

In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 333–342. ACM.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.

Cegin, J.; Simko, J.; and Brusilovsky, P. 2023. ChatGPT to Replace Crowdsourcing of Paraphrases for Intent Classification: Higher Diversity and Comparable Model Robustness. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1889–1905. Singapore: Association for Computational Linguistics.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Grunde-McLaughlin, M.; Lam, M. S.; Krishna, R.; Weld, D. S.; and Heer, J. 2023. Designing LLM Chains by Adapting Techniques from Crowdsourcing Workflows. arXiv:2312.11681.

He, Z.; Huang, C.-Y.; Ding, C.-K. C.; Rohatgi, S.; and Huang, T.-H. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. *arXiv preprint arXiv:2402.16795*.

Huang, S.-H.; Huang, C.-Y.; Deng, Y.; Shen, H.; Kuan, S.-C.; and Huang, T.-H. K. 2022. Too Slow to Be Useful? On Incorporating Humans in the Loop of Smart Speakers. arXiv:2212.03969.

Huang, T.-H.; and Bigham, J. 2017. A 10-month-long deployment study of on-demand recruiting for low-latency crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 5, 61–70.

Huang, T.-H.; Chang, J. C.; and Bigham, J. P. 2018. Evorus: A crowd-powered conversational assistant built to automate itself over time. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–13.

Lasecki, W.; Miller, C.; Sadilek, A.; Abumoussa, A.; Borrello, D.; Kushalnagar, R.; and Bigham, J. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, 23–34.

Lasecki, W. S.; Wesley, R.; Nichols, J.; Kulkarni, A.; Allen, J. F.; and Bigham, J. P. 2013. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, 151–162. ACM.

Mahmood, A.; Wang, J.; Yao, B.; Wang, D.; and Huang, C.-M. 2023. LLM-Powered Conversational Voice Assistants: Interaction Patterns, Opportunities, Challenges, and Design Guidelines. arXiv:2309.13879.

Törnberg, P. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. arXiv:2304.06588.

Wu, T.; Zhu, H.; Albayrak, M.; Axon, A.; Bertsch, A.; Deng, W.; Ding, Z.; Guo, B.; Gururaja, S.; Kuo, T.-S.; Liang, J. T.; Liu, R.; Mandal, I.; Milbauer, J.; Ni, X.; Padmanabhan, N.; Ramkumar, S.; Sudjianto, A.; Taylor, J.; Tseng, Y.-J.; Vaidos, P.; Wu, Z.; Wu, W.; and Yang, C. 2023. LLMs as Workers in Human-Computational Algorithms? Replicating Crowdsourcing Pipelines with LLMs. arXiv:2307.10168.

Yang, Z.; Xu, X.; Yao, B.; Zhang, S.; Rogers, E.; Intille, S.; Shara, N.; Gao, G. G.; and Wang, D. 2024. Talk2Care: Facilitating Asynchronous Patient-Provider Communication with Large-Language-Model. arXiv:2309.09357.