Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2022)

Comparing Experts and Novices for AI Data Work: Insights on Allocating Human Intelligence to Design a Conversational Agent

Lu Sun[1], **Yuhan Liu**[2], **Grace Joseph**[3], **Zhou Yu**[4], **Haiyi Zhu**[3], **Steven P. Dow**[1]
[1] UC San Diego,
[2] Princeton University,
[3] Carnegie Mellon University,
[4] Columbia University
l5sun@ucsd.edu, yl8744@princeton.edu, gejose ph@andrew.cmu.edu, zy2461@columbia.edu,
spdow@ucsd.edu, haiyiz@andrew.cmu.edu

Abstract

Many AI system designers grapple with how best to collect human input for different types of training data. Online crowds provide a cheap on-demand source of intelligence, but they often lack the expertise required in many domains. Experts offer tacit knowledge and more nuanced input, but they are harder to recruit. To explore this trade off, we compared novices and experts in terms of performance and perceptions on human intelligence tasks in the context of designing a text-based conversational agent. We developed a preliminary chatbot that simulates conversations with someone seeking mental health advice to help educate volunteer listeners at 7cups.com. We then recruited experienced listeners (domain experts) and MTurk novice workers (crowd workers) to conduct tasks to improve the chatbot with different levels of complexity. Novice crowds perform comparably to experts on tasks that only require natural language understanding, such as correcting how the system classifies a user statement. For more generative tasks, like creating new lines of chatbot dialogue, the experts demonstrated higher quality, novelty, and emotion. We also uncovered a motivational gap: crowd workers enjoyed the interactive tasks, while experts found the work to be tedious and repetitive. We offer design considerations for allocating crowd workers and experts on input tasks for AI systems, and for better motivating experts to participate in low-level data work for AI.

Introduction

Many AI systems rely on human-provided data for training. While data work can significantly affect the performance and robustness of AI systems, it is often undervalued compared to other aspects of AI design (Sambasivan et al. 2021). Crowdsourcing has proved to be an effective strategy to collect training data since crowd workers can serve as a cheap and on-demand source of intelligence. However, crowd workers often lack domain knowledge and sometimes may fail to accomplish complex tasks for model training (Flores-Saviaga et al. 2020; Hashimoto and Sassano 2018). To overcome this challenge, researchers have explored decomposing complex work into micro-tasks or providing just-in-time training for crowd workers (Yuan et al. 2016). Another approach is to recruit experts to perform data work; an expert's tacit knowledge could help them perform better on input tasks for specialized domains. However, in contrast to on-demand crowds, experts are harder to recruit and may be hesitant to engage with low-level data work (e.g., labeling or curating tasks). Some researchers have raised concerns that undervaluing data work – particularly by domain experts – can cause negative, cascading effects on high-stakes applications of AI. (Sambasivan and Veeraraghavan 2022).

In this paper, to understand the role of expertise and the trade-offs for data work on AI systems, we created an AI that requires different types of human input. We built a preliminary text-based conversational chatbot and a pipeline for human input tasks designed to improve the system. The goal of the chatbot was to create a virtual patient called "MemberBot" that simulates someone seeking support on 7Cups.com, an online peer-to-peer mental health service (Yao et al. 2021). The 7Cups organization sought to create an authentic, yet a low-risk way of helping volunteers practice listening and counseling before engaging with real members who need support. To train "MemberBot", we created a platform to engage humans on a range of data work with different levels of complexity, including: 1) correcting the intent classification of statements made to the chatbot, 2) adding new intent classification categories, 3) correcting the responses offered by the chatbot, and 4) authoring new lines of dialogue for the chatbot. Our platform asks participants to have an initial interaction with the chatbot and then it replays the conversation while asking for input on each step. This replay strategy allows participants to authentically evaluate the chatbot and reflect on its design before doing the data work, unlike prior work that seeks input *during* the conversation (Hancock et al. 2019) or even without trying the chatbot (Yu et al. 2016).

We conducted a comparative study to explore trade-offs in terms of performance and engagement on data work by novices and experts. We recruited 11 domain experts who had extensive experience talking with people as volunteer listeners on 7Cups, as well as 15 novice crowd workers from Amazon Mechanical Turk who had no background in counseling and no affiliation with 7Cups. As a measure of performance, independent experts rated the quality of human inputs for all tasks. A post-task survey captured participants' perceptions of the easiness and enjoyment of each task.

Our comparative analysis found that in contrast to novices, experts held longer conversations and employed more active listening strategies. Experts also proved to be more productive – adding more intents, correcting more responses, and authoring new, emotionally appropriate lines of dialogue for the chatbot. We found that novices were as good as experts for the low level data curation work of reclassifying their own input dialogue with the chatbot. In terms of engagement, the survey responses indicate that experts perceived the tasks as a whole to be easier, yet less enjoyable, than the novice crowd workers. While the crowd workers enjoyed the tasks, experts found the tasks to be tedious and repetitive.

Our work advances knowledge of how experts and novices differ in terms of engagement and performance, while doing a range of data work within our novel chatbot design platform. Novices are sufficient for low-level annotations that only require an understanding of natural language, while domain experts appear more suited for creative or generative data work. The findings provide implications for chatbot developers thinking about how to allocate different types of data work to human participants. We also discuss future work around how to create data-work pipelines that take advantage of different types of human intelligence and how to provide incentive structures to engage experts on data work.

Related Work

Human Intelligence to Support AI Data Work

Previous work has employed human-in-the-loop mechanisms to engage people in different types of data work to advance AI or ML systems (Vaughan 2017). The quality of training data provided by humans can have a significant impact on the quality of the AI system being developed (Halevy, Norvig, and Pereira 2009; Sambasivan et al. 2021). Common practices for data science workers include discovering data, capturing data, and curating data (Muller et al. 2019; Sambasivan et al. 2021). Some data work, such as wrangling, is often perceived as tedious and time-consuming (Kandel et al. 2012).

In the context of conversational agents, training a chatbot to converse like a human requires substantial data work. One major aspect is to improve the chatbot's ability to "understand" text input such that it can correctly identify the intention behind the dialogue (Hashimoto and Sassano 2018). This aspect includes data curation work on classification – as in, noticing incorrect classifications and specifying more coherent intentions. For example, previous research asked people to label data or to fix labels on conversations to improve the NLP (Wang, Hoang, and Kan 2013; Yu et al. 2016).

Prior work has also demonstrated a range of strategies for leveraging human intelligence to *generate* responses for chatbots (Li et al. 2016) for the goal of creating natural and contextually aware chatbot dialogue for a variety of situations. For example, the Protochat system invited crowds to interact with a chatbot to generate more conversational topics (Choi et al. 2021). The self-feeding chatbot asked crowds to interact with a chatbot and then immediately evaluate and rephrase the chatbot responses (Hancock et al. 2019). In our study, we aim to extend prior work by breaking down and comparing how novices and experts perform tasks to improve chatbots. Unlike the prior work, we explore a replay model where participants first "play" with the chatbot before providing any input. This replay model can potentially help participants make better decisions on AI model design (Holstein et al. 2020).

Novices versus Experts for AI Data Work

Crowdsourcing has proved to be an effective strategy for accomplishing some types of data work. Previous research showed that novice crowd workers can produce meaningful content or ideas with enough guidance (Kim and Monroy-Hernandez 2016; Yuan et al. 2016). For example, researchers found that novice crowds can create new product ideas but they typically show lower feasibility compared to professionals' ideas (Poetz and Schreier 2012). Previous work also finds that with sufficient guidance, such as rubrics that provide scaffolding, novice workers can write feedback that is rated nearly as valuable as experts (Yuan et al. 2016).

In the context of chatbot design, researchers have also leveraged crowdsourcing for a range of tasks from labeling conversational data to rewriting chatbot responses (Weston 2016; Li et al. 2016; Yu et al. 2016; Choi et al. 2021). For example, Fantom recruits crowd workers to author utterances when the system lacks valid responses in a given historical context (Jonell et al. 2018). Other research asks crowd workers to evaluate the chatbot performance by improving the responses. For example, Li et al. explored a method for evolving existing dialogue scenarios by paying crowd workers to augment the chatbot responses as they proceed with the conversation (Li et al. 2016). Likewise, the "self-feeding" chatbot asks the user for feedback when the conversation takes a misstep, but only when it predicts that a user had an unsatisfactory interaction. The Protochat system also employs a more targeted approach by asking crowds for specific input on the chatbot performance, such as adding a conversation branch or generating a new topic that the chatbot can discuss (Choi et al. 2021). These studies establish that chatbot developers can recruit crowds to systematically perform relevant data work. However, the research also raises concerns about the quality of input provided by crowd workers. Input provided by crowd workers might be ambiguous and not as comprehensive compared to domain experts (Wauck et al. 2017; Hashimoto and Sassano 2018). Crowds often have extrinsic monetary motivations (Mason and Watts 2009) and insufficient expertise (See et al. 2013) which might be especially critical in a conversational context that requires subtlety and tacit domain knowledge, such as peer counseling (Chi, Glaser, and Farr 2014).

Instead of crowds, other researchers have explored ways to engage experts with advanced skills and domain knowledge in data work. Previous studies invite community experts who have relevant experience to engage with data or perform related tasks (André et al. 2013; Bhardwaj et al. 2014; Heimerl et al. 2012; Chan, Dang, and Dow 2016). For example, the Cobi project invited committee members to perform micro-tasks that provide thematic data on papers

that could be useful for organizers to create paper sessions at the conference (André et al. 2013). The community sourcing project motivated experts to perform data work by strategically situating a vending-machine-style kiosk in a public space frequented by the subject matter experts.

Some prior work has started to unpack the tradeoffs of using novice crowds vs. experts for data work. For example, Flores et al. investigated the performance of paid crowd workers and volunteers on a content curation task (Flores-Saviaga et al. 2020). Results indicate that volunteer collaborators are more effective at open-ended tasks, while paid crowd workers are more effective at decomposed tasks. Our research extends this work by exploring differences between novices and experts on a range of tasks. We created a platform to directly engage domain experts and crowd workers on different types of data work – low level data curation tasks and high level data generation tasks (Muller et al. 2019). The goal of the current study is to understand their trade-offs in terms of performance and engagement to inform future pipelines that seek to leverage both novices and experts.

Chatbot Design and Development

Our paper explores questions about how domain experts versus crowd workers engage with, perform on, and perceive various data work in the context of designing a mental health chatbot. We partnered with 7Cups[1] which is an online mental health community that provides free online therapy and support to people experiencing emotional distress. Members who have mental health problems connect via text-based chat with volunteer listeners to get support on a variety of mental health issues, including depression, break-ups, anxiety, and self-harm. We worked with the 7Cups organization to develop a "MemberBot" that simulates a member with mental health concerns seeking help, such as experiencing a break-up. 7Cups volunteer listeners must first pass a short training program on active listening before they are eligible to chat with real members. We designed this patient-like "MemberBot" to help volunteer listeners practice their active listening skills before interacting with real members.

Chatbot Design Process

To help us better understand the needs and context of the MemberBot, we conducted semi-structured interviews with ten experienced listeners on 7Cups for one hour each. We captured their design ideas and collected possible scenarios for different mental health issues. Three researchers on our team collaboratively summarized their ideas and made choices for a chat scenario and MemberBot's personality. From this needfinding interview results, we created the following persona and backstory for our MemberBot named Andrew with a backstory: Andrew is disappointed and frustrated because of his recent break-up with his girlfriend. Andrew is 22 years old and his ex-girlfriend is 20. They were together for 6 months before his girlfriend ended the relationship 2 weeks ago. After the breakup, Andrew is upset because his ex lied about her feelings and abruptly broke up with him. During the conversation, Andrew uses storytelling to reveal his heartbreak.

[1] www.7Cups.com

Chatbot Implementation

Coding existing chat data After signing a data-sharing agreement with 7Cups, we got access to a completely anonymized dataset of 99.6 million messages exchanged between different volunteer listeners and members seeking support. First, we searched for existing chatrooms that were primarily in English, contained the keyword "breakup", consisted of at least 50 messages, and had a high rating as 5 of 5 that rated by the members. Then, we selected 30 existing chats that had contexts similar to the persona described above.

To train the chatbot's natural language understanding (NLU) model, we developed a coding schema to code dialogue behaviors based on the existing Motivational Interviewing Skill Code (MISC), a widely used schema to code dialogue behaviors in counseling conversations (Miller et al. 2003; Cao et al. 2019). Our coding schema contains 14 behavioral categories. The definitions and examples of each intent are also shown in Table 3. We coded 392 question-answer pairs in total from the sample of 30 existing chats and used these as training data for our NLU model. The 392 responses were paraphrased in order to make them align with the breakup story and to construct the response database.

Model architecture We designed a web-based interface to support chat between "MemberBot" and listeners (see Figure 1 for an example conversation). We used the open-source Rasa framework to implement MemberBot as a retrieval-based chatbot (Bocklisch et al. 2017). Rasa is an open source natural language processing framework that we can translate messages into intent categories for the chatbot to understand Retrieval-based chatbots learn to map user utterances to a system response based on large datasets of human-human conversations. Memberbot's NLU model is an intent classifier that first categorizes the intention of users' input and then provides a response with the highest prediction score associated with that intent.

We trained the NLU classifier on a set of features as input, including SpaCy pre-trained word embeddings, a regex featurizer that can create vector representations using regular expressions, and two count vectorizers using a bag-of-words representation. Character n-grams are set to 1 and 3 respectively as lower bound and upper bound. For intents classification, we used the multitask architecture called DIET (Dual Intent Entity Transformer) Classifier in Rasa. The model took dense and sparse features as inputs. Dense features are from pre-trained embeddings and sparse features are from the training data. Then the classifier will learn the boundaries between groups of user inputs to classify different intents. The output of the NLU model is the predicted confidence scores for each item in the intents dictionary with respect to the current user message.

After classifying the user's intent, the Rasa response selector maps the response with corresponding intents and then selects the best response with the highest probability from the response database. These responses in the database

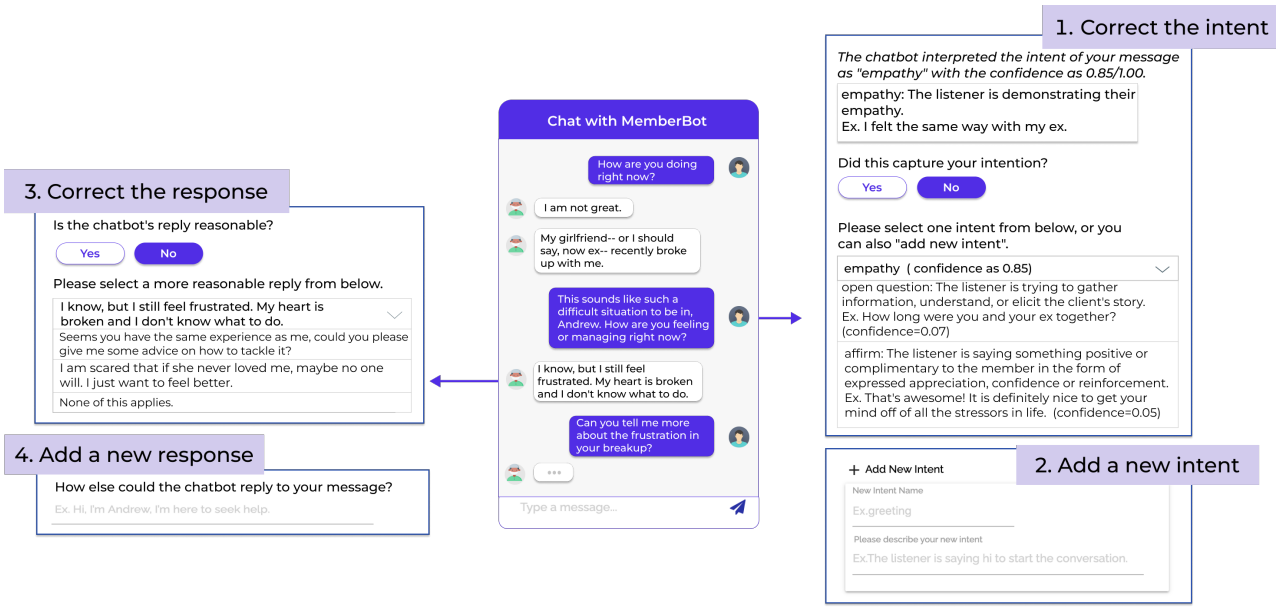

Figure 1: Example chat and design pipeline: After participants chat with the MemberBot, we replay their conversation and ask them to engage in four tasks. Participants completed these four tasks for each chat message.

are all pulled from prior chat logs and paraphrased by the research team. If the highest confidence score is less than the threshold (less than 0.1 indicating very low confidence) or the difference between the top two highest scores is less than the threshold, the fallback response will be triggered by replying "Sorry, I don't understand.".

Study Method

To explore the trade-offs between experts and novice crowd workers, we conducted a study to gather empirical data on the differences in behavior, perceptions, and performance when interacting with and providing input to improve a conversational agent. Our comparative study collected data on the following research questions: how do experts (experienced listeners) and novices (crowd workers) differ in terms of how they converse with the chatbot (**RQ1**), how they perform input activities designed to improve the chatbot (**RQ2**), and, how they perceive the difficulty and enjoyment of different activities (**RQ3**)?

Participants

We recruited two types of participants: domain experts and novice crowd workers. All participants were over 18, spoke English, and were compensated $20/hr for time. The extrinsic rewards were held constant to maintain parity among all participants.

Experts (E1 – E11): With the help of the 7Cups community team, we recruited 11 domain experts from the 7Cups.com listener community. The listeners' tenure on 7Cups varied from one to five years, and represented diverse backgrounds: from years of crisis hotline experience, to professional experience in a mental health-related field, to recently transiting from being a 7Cups member to a listener. Their collective experience and knowledge could help shape the "MemberBot" to emulate real users while maintaining the training purpose. In terms of motivation, most expressed interest in helping to improve the MemberBot to benefit the the 7Cups community.

Novices (N1 – N15): We recruited 15 novice crowd workers from the crowdsourcing platform Amazon Mechanical Turk with task acceptance rates higher than 95 %. All recruited crowd workers had no experience as a professional counselor or interacting with a chatbot related to mental health. They also had no experience interacting with the 7Cups community.

Procedure

First, the system presents participants with a consent form and an overview of the whole study. Next, an embedded chat window appears with MemberBot saying "Hi, I am Andrew.", inviting the participant to add their own text-based chat messages. Participants were asked to provide support to the MemberBot "Andrew" who just experienced breakup issues for at least 3 minutes. After 10 minutes, participants see a pop up message telling them the chat session has ended. Next, participants were asked to what extent the chatbot and its story were coherent, comprehensive, natural, and well aligned with the written persona description.

Next, the system then replays the conversation turn-by-turn, while participants perform four activities for each exchange (one statement by the user and a response from the chatbot): 1) correcting the intent classification, 2) adding new intent categories, 3) correcting MemberBot responses, and 4) creating new responses. To help participants prepare for each task, onscreen instructions introduced the concepts of intents and actions.

Figure 1 shows a detailed example of the four data tasks. The user writes: "This sounds like such a difficult situation to be in Andrew. How are you feeling or managing right now?" The chatbot classifies the user's input as "empathy" (meaning the participant is demonstrating empathy) with the confidence score as 0.85 out of 1.00. For the **Correct the intent** task, the interface shows the question "Did this capture your intention?" to ask the user whether "empathy" captures the intent of their message. If the user selects no, the interface will show a drop-down list of other potential intents predicted by the chatbot model, including an explanation and examples to the side. Intents are displayed in descending order according to their confidence score, excluding intents with confidence scores less than 0.01. If none of the existing intent labels apply, users do the **Add a new intent** task where they add a new intent category with an explanation to describe the meaning behind their message.

When the chatbot responds, users are asked whether the response is reasonable. If they select yes, the system moves on to the next line of dialogue in the replayed conversation. If the user selects no, this system shows the **Correct the response** task where the users selects from a drop-down list one of ten potential responses based on a confidence estimate. If the user feels that none of these responses apply, the system prompts the **Create a new response** task, where the user can author a new chatbot response. After review, the response is added to the corpus under the corresponding intent.

After completing the tasks, participants were encouraged to think about the general design of the chatbot and identify ways to further improve the chatbot. Participants were asked to reflect on each activity as well as the entire experience.

Data Collection and Analysis

User activity data During the chat phase, we collected messages between participants and the MemberBot. We explored RQ1 by comparing the behavioral patterns during the chat. Each participant's engagement was measured by the number of messages they exchanged with the MemberBot, the length of the messages sent by the participant, and the time they spent chatting with the MemberBot. To quantify and compare each dialogue's behaviors, the research team qualitatively coded the intent of each of their chat messages.

To answer RQ2, we collected participants' inputs during activities, including: the intent labels they evaluated and created, and the responses they evaluated and authored. We measured the frequency and quality of their inputs on each task to compare the experts' and novices' performance. For frequency, we measured the number of inputs each participant provided on each activity. Furthermore, we measured the quality of the inputs they provided on the intent classification activity by incorporating their inputs during the chat and then retraining the NLU model to test the effect of their inputs on the intent classification performance.

We measured the quality of their responses by coding whether the response was coherent with the previous messages, whether the response contained emotional disclosure, and whether the response offered new storylines. To evaluate open domain conversation, prior researchers used human judgment on coherence as a metric (Dziri et al. 2019). Previous research pointed out that the self-disclosure of feelings can provide high-quality information about the communicator and facilitate the development of social relationships (Yang, Yao, and Kraut 2017; Tamir and Mitchell 2012). Hence, we evaluate how much emotional disclosure is in the responses authored by the participant. Another metric we used was the number of new storylines that participants authored in their responses, reflecting their creativity (Choi et al. 2021).

Post survey data To explore RQ3 about their perceptions of different activities, we asked each participant to rate the difficulty and enjoyment level from 1 – 5 for each activity (1 is extremely hard and 5 is extremely easy and 1 is not enjoyable and 5 is very enjoyable). We included open-ended questions to collect reasons for their ratings.

Qualitative coding of intents and responses Three researchers conducted the intent coding for messages sent by experts and novices. Due to the subjectivity in assessing a participant's input related to intent, researchers performed independent coding so that we could establish a "ground truth" in the intents classification of all the participants' chat messages. Each message was coded for one intent based on the intent definition, as shown in Table 3. The research team iteratively trained and tested on 10% samples (30 messages) of the messages repeatedly until an acceptable inter-rater reliability was reached for the code (a Cohen's Kappa above 0.80 pairwise). Then, three researchers split the remaining data evenly and labeled the intents independently.

The team also conducted qualitative coding of the new responses created by experts and novices. The team read each new response and simultaneously coded three different binary variables: (1) is the response coherent with the previous message? (Dziri et al. 2019), (2) does the response offer new story information (as in new ideas for personal information and backstory for the MemberBot)? (Yang, Yao, and Kraut 2017), (3) does the response disclose how the MemberBot might feel at that moment, either positive or negative emotions ? (Yang, Yao, and Kraut 2017) The team iteratively trained and tested on 10% samples (40 messages) of the responses randomly selected from all participants until an acceptable inter-rater reliability was reached for each aspect (pairwise Cohen's Kappa above 0.80).

Results

RQ1: Experts and Novices Conversed Differently with the Chatbot

Table 1 showed how experts and novices exchange messages with the chatbot. We ran a non-parametric test - Mann–Whitney U test on each activity to test the difference between experts and novices (Hettmansperger and McKean 2010). We found that experts exchanged significantly more messages (33.9 messages) with the chatbot than novices (28.4 messages). Furthermore, experts spent significantly more time chatting with the MemberBot($t = 393.3$ seconds) than novices ($t = 338.2$ seconds). Messages sent by the experts (12.8 words on average) are significantly longer than

messages sent by the novices (8.0 words on average). To sum up, experts are more actively engaged with the chatbot.

average chat interactions	Experts	Novices	
avg # of messages exchanged	33.9 (9.9)	28.4 (12.7)	*
avg # of messages posted	14.5 (4.0)	13.8 (6.3)	*
avg length of messages	12.8 (7.3)	8.0 (6.4)	**
avg time spent (seconds)	393.3 (117.2)	338.3 (177.2)	*

Table 1: Overview of messages exchanged in real-time chat by 11 experts and 15 novices. The table lists average frequency with standard deviation. The p-value shown in the rightmost column indicates the significance score of the Mann–Whitney U test between experts and novices. Experts are more actively engaged than novices in the chat. p-value significance codes: <0.0001 '***', < 0.001 '**', < 0.01 '*'

Experts actively listened; novices gave advice For each intent, we built a linear regression model to compare experts' and novices' conversational behaviors. As shown in Table 3, experts significantly asked more open questions and reflected significantly more with the MemberBot, providing significantly less specific advice to the MemberBot or confronting the MemberBot. Correspondingly, novices gave significantly more advice to the MemberBot, but closed the conversation formally significantly less.

During the chat, experts, as 7Cups listeners, were able to easily employ the active learning strategies they learned from 7Cups in their chat. Contrastingly, novices even with the short training at the beginning, still lacked experience of being a listener, which differentiated their chatting behaviors with experts. One of MemberBot's common responses to listeners is to ask for advice, where MemberBot will respond "Seems you have the same experience as me, could you please give me some advice on how to tackle it?". Experts were more likely to tell the MemberBot that they feel uncomfortable giving advice than novices, as seen in their higher frequency of messages with the intent "avoid advice". For example, expert E3's response is "As a listener here, I try not to give advice, but I would be happy to talk more with you about this situation. Then, hopefully we can come up with what kind of actions that would work best for you in your situation.". However, novices tend to provide specific advice such as telling the MemberBot to spend time with their friends or work out. For example, novice N3 provided a specific response by saying "I think it's easier to get through if you have other things to get your mind off it. Like spend time with friends and working out."

RQ2: Novices Can Correct Intents; Experts Authored Better Responses

After participants finished their chat with the MemberBot, they were asked to work on four tasks: (1) correcting the intent; (2) adding a new intent (3) correcting the response (4) creating a new response. Both experts and novices were asked to correct the intents of their messages predicted by the chatbot. To evaluate whether they behave significantly differently on four tasks, we built a linear regression model for each activity. As shown in Table 2, experts provided significantly more inputs than crowd workers on two activities: creating new intents and correcting responses.

completed tasks	Experts	Novices	
correct the intent	3.0 (1.4)	2.1 (2.2)	-
add a new intent	2.6 (3.0)	1.0 (2.1)	**
correct the response	8.7 (4.4)	4.4 (3.8)	**
create a new response	10.4 (7.1)	8.9 (4.5)	-
Total messages in the chat	33.9 (9.9)	28.4 (12.8)	***

Table 2: Summary statistics on each of the four data-related tasks for the experts and novices. The table lists the average frequency with standard deviation. A linear regression is conducted using the activity-completion frequency as the dependent variable, the count of each participant's total messages as control variable and whether participant type is expert or crowd as experiment variable. p-value indicates the significance of the co-efficient of participant type, as shown on the rightmost column. p-value significance codes: 0.0001 '***', 0.001 '**', 0.01 '*'

Correcting intents: experts and novices both provided input that could improve NLU predictions Participants contributed to the intent classification in two ways: (1) they provided more samples of dialogue thus expanding the training set; (2) they provided input on how the NLU classifies their own speech. First, to measure whether participants' inputs can improve the chatbot's NLU model by expanding the training set, we incorporated the chat messages provided by experts and novices as additional training samples on a historical dataset and trained three models: a baseline model, a model that incorporated participants' messages split into single sentences with participants' labels and a model that incorporated sentences with researchers' labels.

We wanted to explore whether adding more training samples with labels from experts and novices can outperform the baseline model, which was trained on 249 listeners' messages extracted from the historical dataset, on the intent classification accuracy. The research team randomly selected another 30% from the historical dataset (96 listener messages) with labels as test data. Correspondingly, we trained two new models, adding participants' messages with the researchers' labels. These models added the participants' messages as additional training samples but with the intent of each message labeled by researchers. We compared the F1-scores of the three different models. As shown in the Table 4, models that incorporated participants' sentences with researchers' labels outperform the baseline model. This comparison indicates that incorporating data from both crowd workers and experts can improve the intent classification. Also, when we compare models that incorporated participants' messages with their own labels provided while interacting with the bot, both models outperform the baseline model. Interestingly, the model that added the novices' messages with novices' labels has a higher F1-score than the model that added experts' messages with experts' labels. One hypothesis is that crowd workers' messages are usually short and have simple intents, which are easier to clas-

Intent	Definition	Example	Experts	Novices	
open question	The listener is trying to gather information, understand or elicit the client's story by asking open-ended questions.	How long were you and your ex together?	6.2 (2.9)	3.8 (3.0)	*
support	The listener is providing sympathetic, compassionate, or understanding comments that have the quality of agreeing or siding with the client.	I understand that, I believe you.	3.5 (1.1)	4.3 (3.5)	
reflect	The listener is reflecting their understanding of the information they have received from the member.	So you said you were together for 6 months.	2.4 (1.7)	0.3 (1.0)	***
empathy	The listener is demonstrating their empathy.	I'm so sorry to hear that.	1.3 (0.7)	1.6 (1.5)	
closing	The listener is ending the conversation.	Take care of yourself and your heart :)	1.2 (1.4)	0.4(0.7)	.
greeting	The listener is saying hi to start the conversation.	Hi, Andrew.	1.1 (0.5)	1.1 (0.5)	.
conventional opening	The listener is being courteous and seguing into discussion of the member's stressor.	What is going on? What's been happening?	0.8 (0.7)	0.7 (0.6)	
affirm	The listener is saying something positive or complimentary to the member in the form of expressed appreciation, confidence or reinforcement.	It is definitely nice to get your mind off of all the stressors in life.	0.7 (1.0)	0.8 (0.7)	
avoid advice	The listener avoids providing direct advice to the member	I'm not allowed to give advice to you.	0.6 (1.0)	0.0 (-)	*
facilitate	The listener is trying to get more details from the member regarding their conflict.	Can you tell me more about what happened?	0.5 (0.7)	0.7 (1.8)	
give advice	The listener is giving specific advice to the member.	I think you should go out and talk with her.	0.5 (0.7)	2.0 (1.8)	**
off-topic	The listener is abusing their role and engaging in inappropriate behavior.	Tell me about your sex life. Women are like that, stick to bots.	0.4 (1.2)	0.1 (0.2)	
confront	The listener directly disagrees, argues, corrects, or seeks to persuade.	No, you are not! You can't force her to love you.	0.2 (0.4)	0.3 (1.0)	
give information	The listener is giving information to the member, explaining something, educating or providing feedback or disclosing personal information.	Here's a link to a helpful self-help guide!	0.1 (0.3)	0.0 (-)	

Table 3: Intent categories, definitions, examples, and average frequency of intents (with standard deviation) expressed by experts (represented as E) and novices(represented as N) during the live chat with MemberBot are listed below. A linear regression is conducted using the frequency of a particular intent as the dependent variable, the count of each participant's total sentences as control variable and the participant type (S or CW) as experiment variable. p-value shown in the rightmost column indicates the co-efficient significance of participant type. p-value significance codes: < 0.0001 '***', < 0.001 '**', < 0.01 '*', < 0.05 '.'

model /training size	Experts	Novices
baseline model /249	0.517	0.517
model with researchers' labels /462	0.609	0.593
model with participants' labels /462	0.576	0.589

Table 4: F1 score of the intents classification models. Baseline model only trained on 249 messages pairs from the historical data. Models that incorporated participants' single sentences added 213 sentences from participants as additional training sample. Higher F1 score indicates better model. Models incorporated participants' inputs outperformed the baseline model.

sify, while experts' messages usually contain multiple intents. This result revealed that crowd workers' performance on correcting intents is comparable to experts.

Creating new intents: experts generated more new valid intent categories Experts created significantly more intents than novices. 11 of the 15 intents added by the novices were similar to the intents already in the coding schema. For example, novices create a new intent category for their message "Keep going" as "prompting" and explained that "The listener is prompting". However, this label of "prompting" is very similar to the "facilitate" label in the intent category, which means "the listener is trying to get more details from the member". On the contrary, experts tended to create intent categories that were distinct from existing categories with good details about the underlying intentions. For instance, one expert (E3) provided a new intent label "Encouraging self-help" to explain the detailed intents of the message "What would you tell a friend in your situation?". The intent labels created by novices were less creative and descriptive than the intent labels added by the experts.

Correcting responses: experts corrected responses more effectively. As shown in Table 2, experts corrected 8.7 responses on average which was significantly more than the corrected by the novices (4.4 responses on average). Participants corrected MemberBot's responses when the responses seemed inconsistent with the participant's prior message or that simply repeated an earlier response. For example, when the chatbot received a response saying "That's understand-

able, have a nice night, we are here if you ever need us." Then the chatbot replied "I know, but I still feel frustrated. My heart is broken and I don't know what to do". The expert corrected the response into a more consistent one saying "Thanks, I really appreciate that that's really kind of you. I guess the whole thing has just made me extra insecure about myself".

Creating responses: experts authored better lines of dialogue than novices Participants added new responses to further expand the variations of the chatbot's responses depending on different contexts and emotions, or to provide a reasonable response that can answer the listeners' questions better. We first ran a Mann–Whitney U test to evaluate the difference between the length of the response provided by experts and novices (Hettmansperger and McKean 2010). Results showed that responses created by experts are significantly longer (p-value $<$ 1e-3) than novices as experts' responses contain 18.6 words on average while novices' responses contain 9.6 words on average.

To further assess the quality of new responses created by experts and novices, we coded the coherence(Demasi, Li, and Yu 2020) of the response as well as whether or not the response contained emotional disclosure and/or new story information(Chen, Wu, and Yang 2020) as described in Section . To compare the frequency of each code by each participant, we built a linear regression model for each code. Results in Table 5 showed that experts expressed significantly more emotional disclosure and added more storylines than novices when they created new responses. One reason might be that experts can more effectively simulate the scenario because of their experience talking with members on 7Cups.

Qualitative codes	Experts	Novices	
coherence	12.3(6.8)	10.2 (2.9)	-
emotional disclosure	4.0 (2.6)	2.1 (1.3)	*
new storylines	4.6 (4.4)	2.8 (2.2)	*

Table 5: Summary statistics on quality code of responses created by experts and novices. The table listed average frequency of participants with standard deviation. A linear regression is conducted using the count of each code as the dependent variable, the count of each participant's total messages as control variable and whether participant type is expert or novice as experiment variable. p-value indicates the significance of the co-efficient of participant type, as shown on the rightmost column. p-value significance codes: 0.0001 '***', 0.001 '**', 0.01 '*'

Participants tried to be empathetic with the MemberBot and express the emotion that the MemberBot might have faced with the break-up issue. We found that experts tried to add more details of the break-up storyline and disclosure more emotion. For example, when the MemberBot received a message from the listener "It must be very confusing to be met with her mixed signals and behavior. Why do you think she's beginning to distance herself from you?". Instead of following the current response that MemberBot has "I'm scared that if she never loved me, maybe no one will. I just want to feel better", E9 created news responses that

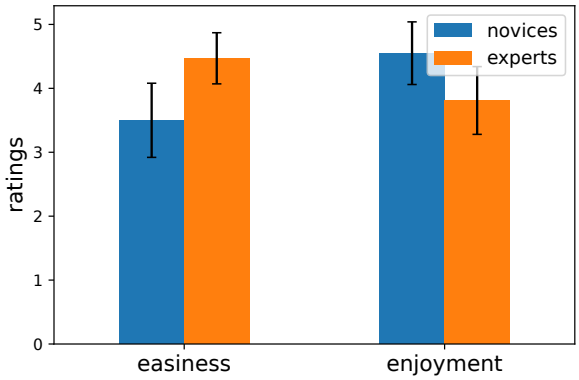

Figure 2: Average ratings of easiness and enjoyment of all four data tasks for experts and novices with standard errors bars are shown. Higher ratings indicate easier or more enjoyable. The cumulative differences across all 4 activities are significant based on the Mann–Whitney U test' result.

contained storyline in more detail – " I don't know, she just gives me the cold shoulder whenever she's in a bad mood, or that's just how she is at one point, but then the next thing you know, she's love-bombing me. I can't tell what she truly feels about me." However, some novices might keep the way MemberBot did. These new storylines authored by participants can help MemberBot expand its conversation flow and handle more diverse questions. In addition, we found experts even authored various versions of responses that MemberBot might express with a slightly different emotion. But all these responses are coherent with MemberBot's persona.

RQ3: Experts Perceived Tasks as Easier Yet Less Enjoyable Than Novices

To understand how participants react to different design activities, we asked all participants to rate their difficulty and enjoyment for each data task from 1 - 5 (1 is extremely hard / not enjoyable, while 5 represents extremely easy / extremely enjoyable) at the end of tasks. We ran a Mann–Whitney U test to evaluate the difference between the easiness ratings and enjoyment ratings provided by 11 experts and 15 novices across four tasks (Hettmansperger and McKean 2010). From the Figure 2, results showed that experts perceived data related tasks as significantly easier than novices (p-value $<$ 1e-03), while novices perceived these tasks as significantly more enjoyable than experts (p-value $<$ 1e-03).

From the survey responses, we observed that novices reflected that they need some context or expertise to answer questions even after the training, which made them find activities like creating new intents and creating new responses difficult. Some novices pointed out the reason that they think the activity is difficult: "It is time-consuming and requires creativity" or "Coming up with varied and realistic responses could be a challenge. I found myself worrying that I was being repetitive"(N8). In contrast, "Creating a new intent" was found to be hard, but engaging for experts.

Adding a new intent may be difficult because they need to think about a proper name for the intent on their own: "It is a little bit hard to add a new intent, since I need to understand the concept of the intent and think about a proper name for the intent. But this thinking process also makes it interesting" (E1). At the same time, having the option to add a new intent provides the user agency in their labeling: "What makes it interesting is that it gets your wheels turning"(E7). When we survey their motivation, experts seemed to pay more attention to the final performance of the chatbot and how could the chatbot benefit the community, saying "I am curious what will be the final chatbot looks like." (E3) On the contrary, crowd workers mentioned their motivation for the payment. Some novices reflected that "the tasks payment is fair to work on"(N9).

Discussion

This comparative study explored differences between domain experts and crowd workers in terms of how they engaged in conversation, how they provide inputs in each data work, and how they perceived the difficulty and enjoyment. Providing human input to simulate a chatbot can involve both high-level data work that benefits from domain experience and low-level data work that does not necessarily require expertise. We decomposed the chatbot development process into concrete tasks and created a data work pipeline that allows both experts and novices to directly interact with the AI. Our research also extends prior research by inviting participants to have a full chatbot experience and to reflect before engaging with data work; prior work in this space typically asks people to do corrections in real-time during the chat which can break the conversational flow.

Existing research has explored the complementary value of experts and crowd workers for a single data work to improve chatbot (i.e. rewrite responses) (Hancock et al. 2019). To extend previous work, our contribution here is that we built an entire chatbot design pipeline and compared the value of two sources of human intelligence on each task. By evaluating their performance, engagement and enjoyment on different tasks, our study provided empirical evidence that crowd workers and domain experts have complementary value in different types of data work. Crowd workers can demonstrate high quality on relatively simple data curation tasks, such as correcting intents. This result aligns with previous studies that showed how crowd workers can follow instructions but produce less original content (Flores-Saviaga et al. 2020). Correspondingly, experts authored better dialogue with higher creativity and more emotion. This result resonates with previous findings that domain experts offer more than an instrumental tool for collecting datasets (Sambasivan and Veeraraghavan 2022). The results imply that expertise that often takes years to build should not be excluded from AI development.

While it may be not surprising to see that domain experts can create more creative content with higher quality, the low perceived enjoyment of experts on these tasks illustrated that chatbot developer should consider allocating different data work not only based on expertise, but also based on participants' interest, motivation and time.

Trade-offs of Recruiting Domain Experts versus Crowd Workers for AI Work

This comparative study provides empirical support for the value of contributions from both domain experts and crowd workers. A key question is what factors influence how experts and novices perform differently on data work? To help chatbot developers to make decisions about how to allocate human intelligence, we highlight three trade-offs.

- **Cost and Time to Recruit**: The cost of recruiting crowd workers is generally less than that of domain experts, given they are easier to find and hire. Correspondingly, having to work around the schedules of experts makes them harder to recruit.
- **Expertise**: In order to naturally interact with humans, chatbots have to not only accurately "understand" humans, but also appropriately "respond" to them. Interestingly, the inputs provided by crowd workers are as good as domain experts in terms of improving the chatbot's ability to predict user intents (i.e., natural language understanding). Specifically, the novices' performance on the task of "correcting the intent" was equivalent as experts. In contrast, experts were much better at creating more creative and emotional responses for the chatbot and at generating appropriate storylines. This suggests that chatbot developers might want to allocate the limited and precious expert time for data creation work to improve the bot's responses and storyline, and recruit novice crowd workers for data curation task - labeling and correcting user intents to improve bot's ability to better understand humans.
- **Motivation**: The domain experts' motivation in participating in this activity tasks stems from their interest and investment in the 7Cups community. Crowd workers' motivations are usually personal or individualistic, often incentivized by the monetary payment (Mason and Watts 2009). Interestingly, our survey responses suggested that crowd workers enjoyed the data work, while experts found the task tedious and repetitive. This uncovers a motivation gap of experts in engaging in the AI data work and outlines future work on strategies to motivate and attract experts.

Design Implications

Based on the trade-offs between the two sources of human intelligence illustrated above, a better approach could be to build a hybrid pipeline that mixes the value of both on data work. AI developers should allocate data work judiciously according to participants' cost, expertise and motivation. For example, experts could be directed to complete data generation tasks, while crowd workers could be directed to focus on low level data curation tasks that experts find tedious. Also, given the motivation gap that experts found the task less interesting, we propose ways to improve the design guideline and attract more experts.

Improve the data work pipeline to better incorporate inputs Our design method still faces several challenges in

involving experts in the data work to support automated system design. Some data work contributed by participants, although meaningful and insightful, cannot be incorporated immediately into the chatbot design. When examining the participants' inputs on data curation tasks, we noted that some of the intents they provided were not reasonable for a chatbot model to incorporate, often because these labels were long and too specific. Therefore, the interface should consider providing feedback to the experts during the activities and curating the experts' labels before incorporating them into the system. For example, developers can introduce an automatic voting system to help the chatbot make decisions (Huang, Chang, and Bigham 2018).

For data creation tasks, it is hard for chatbots or other automated systems to interpret human insights and transfer their implicit inputs directly back to the system. When participants create new responses, we observe that some participants give high-level recommendations on how to improve the response instead of giving direct responses. For example, E1 gave advice by saying "Chatbot should talk about what they have already tried to do to fix the situation". One solution to this challenge is to improve the instruction by further exemplifying the potential responses or providing prompts when the created responses are vague. Another challenge is that some responses conflict with the current storyline. The contradictions in the story might make the user who interacts with the MemberBot confused about what the MemberBot wanted to convey. A method could be implemented to reconcile any contradictions in the design, such as building a computational model to detect such contradictions and improve the consistency of the dialogues (Nie et al. 2020).

Attract experts to engage in data work for chatbot Recruiting experts can have high cost and we found it is challenging to setup with experts to start the data work. However, their expertise that accumulated within years indicated that they should not only play the role as data collector, but directly interact with data. To attract experts, communities might invite experts to do these design tasks while they interact with the chatbot or when they encounter difficulties talking with real members. Intrinsic incentives like showcasing the potential social impact and proper timing can help nudge increase the number of community experts engaged, but also improve their motivation. Another factor is to reduce the repetitiveness for experts. Chatbot developers can strategically select parts of conversations that are not satisfactory or where members encounter repetitive responses, and then invite experts to provide more targeted input for those messages specifically.

Limitations and Future work

Our work has several limitations. Ideally, we would incorporate the inputs from each source and then deploy the designed chatbot to 7Cups. We then invite the intended end-users – the listeners who need training – to evaluate the chatbot. However, we did not deploy the improved chatbot to the real world community because of the security risks and a lack of resources. In the future, we will carefully conduct iterations on the chatbot design and then deploy it to the community.

Additionally, because we recruited participants from the community platform by sending out advertisement emails and we provided limited payment, the number of experts we could involve in the design process was constrained. Different groups of experts may make different design decisions (Zhu et al. 2018). We potentially want to recruit a diverse range of experts, such as clinical psychologists or external experts who have fruitful experience with counselor training. Lastly, multiple cycles of development and deployment are needed to evaluate the utility of data work tasks as well as chatbot performance.

Our study highlights the intelligence beyond chatbot developers – from experts and novices on data work for chatbot design. Our study focuses on developing a MemberBot that simulated a user who needs help with his breakup issue. As chatbots are becoming increasingly widespread to support mental health, we could extend our design pipeline and empirical results to guide chatbot development in other contexts. For example, rather than working on a MemberBot, researchers could invite diverse experts, from community experts to clinical psychologists to chatbot developers to design a "MentorBot" which would guide listeners during live chats. Community experts and clinical psychologists can engage in the chatbot design by providing professional insights on how and when to give feedback to listeners. In addition, this design method can also be extended to the design of a chatbot that provides therapy as a "TherapyBot".

Conclusion

This paper contributes a comparative study to shed light on how experts and crowd workers engage, perform, and perceive the data work behind designing a chatbot for a mental health support scenario. Our work explores a series of data work activities where we involve participants to improve a "MemberBot". We find that both the skill and motivation of experts encourage them to hold longer conversations with the chatbot and to employ more effective mental health strategies compared to crowd workers. The analysis of participants' performance in design activities reveals that novices can be sufficient at correcting the AI's understanding of user intents, while experts prove to be better at authoring novel chatbot responses that offer emotional disclosure and new story lines. Future work could focus on creating data pipelines that carefully allocate different sources of intelligence, and on motivating experts to engage with different types of data work.

Acknowledgments

We would like to thank the 7Cups community team for their support in collecting the data, recruiting participants, and providing feedback. We thank Dr. Ken Holstein and Dr. Robert Kraut for providing feedback on the manuscript. We thank all our participants for their time and valuable insights. Finally, we thank all reviewers for their input. This work is supported by National Science Foundation(NSF) grants #2001851 and #2000782.

References

André, P.; Zhang, H.; Kim, J.; Chilton, L.; Dow, S. P.; and Miller, R. C. 2013. Community clustering: Leveraging an academic crowd to form coherent conference sessions. In *First AAAI conference on human computation and crowdsourcing*.

Bhardwaj, A.; Kim, J.; Dow, S.; Karger, D.; Madden, S.; Miller, R.; and Zhang, H. 2014. Attendee-sourcing: Exploring the design space of community-informed conference scheduling. In *Second AAAI conference on human computation and crowdsourcing*.

Bocklisch, T.; Faulkner, J.; Pawlowski, N.; and Nichol, A. 2017. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Cao, J.; Tanana, M.; Imel, Z. E.; Poitras, E.; Atkins, D. C.; and Srikumar, V. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. *arXiv preprint arXiv:1907.00326*.

Chan, J.; Dang, S.; and Dow, S. P. 2016. Improving crowd innovation with expert facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1223–1235.

Chen, J.; Wu, Y.; and Yang, D. 2020. Semi-Supervised Models via Data Augmentationfor Classifying Interactive Affective Responses. *arXiv preprint arXiv:2004.10972*.

Chi, M. T.; Glaser, R.; and Farr, M. J. 2014. *The nature of expertise*. Psychology Press.

Choi, Y.; Monserrat, T.-J. K. P.; Park, J.; Shin, H.; Lee, N.; and Kim, J. 2021. ProtoChat: Supporting the Conversation Design Process with Crowd Feedback. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3): 1–27.

Demasi, O.; Li, Y.; and Yu, Z. 2020. A Multi-Persona Chatbot for Hotline Counselor Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 3623–3636.

Dziri, N.; Kamalloo, E.; Mathewson, K. W.; and Zaiane, O. 2019. Evaluating coherence in dialogue systems using entailment. *arXiv preprint arXiv:1904.03371*.

Flores-Saviaga, C.; Granados, R.; Savage, L.; Escobedo, L.; and Savage, S. 2020. Understanding the complementary nature of paid and volunteer crowds for content creation. *Avances en Interacción Humano-Computadora*, (1): 37–44.

Halevy, A.; Norvig, P.; and Pereira, F. 2009. The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2): 8–12.

Hancock, B.; Bordes, A.; Mazare, P.-E.; and Weston, J. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415*.

Hashimoto, C.; and Sassano, M. 2018. Detecting absurd conversations from intelligent assistant logs by exploiting user feedback utterances. In *Proceedings of the 2018 World Wide Web Conference*, 147–156.

Heimerl, K.; Gawalt, B.; Chen, K.; Parikh, T.; and Hartmann, B. 2012. CommunitySourcing: engaging local crowds to perform expert work via physical kiosks. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1539–1548.

Hettmansperger, T. P.; and McKean, J. W. 2010. *Robust nonparametric statistical methods*. CRC Press.

Holstein, K.; Harpstead, E.; Gulotta, R.; and Forlizzi, J. 2020. Replay Enactments: Exploring Possible Futures through Historical Data. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 1607–1618.

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.

Jonell, P.; Bystedt, M.; Dogan, F. I.; Fallgren, P.; Ivarsson, J.; Slukova, M.; Wennberg, U.; Lopes, J.; Boye, J.; and Skantze, G. 2018. Fantom: A crowdsourced social chatbot using an evolving dialog graph. *Proc. Alexa Prize*.

Kandel, S.; Paepcke, A.; Hellerstein, J. M.; and Heer, J. 2012. Enterprise data analysis and visualization: An interview study. *IEEE transactions on visualization and computer graphics*, 18(12): 2917–2926.

Kim, J.; and Monroy-Hernandez, A. 2016. Storia: Summarizing social media content based on narrative theory using crowdsourcing. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1018–1027.

Li, J.; Miller, A. H.; Chopra, S.; Ranzato, M.; and Weston, J. 2016. Dialogue learning with human-in-the-loop. *arXiv preprint arXiv:1611.09823*.

Mason, W.; and Watts, D. J. 2009. Financial incentives and the" performance of crowds". In *Proceedings of the ACM SIGKDD workshop on human computation*, 77–85.

Miller, W. R.; Moyers, T. B.; Ernst, D.; and Amrhein, P. 2003. Manual for the motivational interviewing skill code (MISC). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*.

Muller, M.; Lange, I.; Wang, D.; Piorkowski, D.; Tsay, J.; Liao, Q. V.; Dugan, C.; and Erickson, T. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–15.

Nie, Y.; Williamson, M.; Bansal, M.; Kiela, D.; and Weston, J. 2020. I like fish, especially dolphins: Addressing Contradictions in Dialogue Modelling. *arXiv preprint arXiv:2012.13391*.

Poetz, M. K.; and Schreier, M. 2012. The value of crowdsourcing: can users really compete with professionals in generating new product ideas? *Journal of product innovation management*, 29(2): 245–256.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

Sambasivan, N.; and Veeraraghavan, R. 2022. The Deskilling of Domain Expertise in AI Development. In *CHI Conference on Human Factors in Computing Systems*, 1–14.

See, L.; Comber, A.; Salk, C.; Fritz, S.; Van Der Velde, M.; Perger, C.; Schill, C.; McCallum, I.; Kraxner, F.; and Obersteiner, M. 2013. Comparing the quality of crowdsourced data contributed by expert and non-experts. *PloS one*, 8(7): e69958.

Tamir, D. I.; and Mitchell, J. P. 2012. Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences*, 109(21): 8038–8043.

Vaughan, J. W. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.*, 18(1): 7026–7071.

Wang, A.; Hoang, C. D. V.; and Kan, M.-Y. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1): 9–31.

Wauck, H.; Yen, Y.-C.; Fu, W.-T.; Gerber, E.; Dow, S. P.; and Bailey, B. P. 2017. From in the class or in the wild? Peers provide better design feedback than external crowds. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 5580–5591.

Weston, J. E. 2016. Dialog-based language learning. In *Advances in Neural Information Processing Systems*, 829–837.

Yang, D.; Yao, Z.; and Kraut, R. 2017. Self-disclosure and channel difference in online health support groups. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Yao, Z.; Yang, D.; Levine, J. M.; Low, C. A.; Smith, T.; Zhu, H.; and Kraut, R. E. 2021. Join, Stay or Go? A Closer Look at Members' Life Cycles in Online Health Communities. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–22.

Yu, Z.; Xu, Z.; Black, A. W.; and Rudnicky, A. 2016. Chatbot evaluation and database expansion via crowdsourcing. In *Proceedings of the chatbot workshop of LREC*, volume 63, 102.

Yuan, A.; Luther, K.; Krause, M.; Vennix, S. I.; Dow, S. P.; and Hartmann, B. 2016. Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1005–1017.

Zhu, H.; Yu, B.; Halfaker, A.; and Terveen, L. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–23.