

Copyright 2023 Ziang Xiao

SEEING US THROUGH MACHINES: DESIGNING AND BUILDING  
CONVERSATIONAL AI TO UNDERSTAND HUMANS

BY

ZIANG XIAO

DISSERTATION

Submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy in Computer Science  
in the Graduate College of the  
University of Illinois Urbana-Champaign, 2023

Urbana, Illinois

Doctoral Committee:

Professor Hari Sundaram, Chair  
Professor Karrie Karahalios, Co-chair  
Dr. Michelle X. Zhou, Juji, Inc.  
Professor Heng Ji  
Professor Brent W. Roberts

## ABSTRACT

Understanding humans at scale is essential for addressing some of the consequential challenges in human society. By gaining insights into why people act as they do, we can design informed interventions that have a positive societal impact, including improving public health, developing a sustainable economy, or advancing fair education. However, the complexity of human behavior necessitates novel and sophisticated tools and methods to capture cultural, social, environmental, and individual characteristics that heavily influence our behaviors. Further, understanding ethical concerns surrounding informed consent, privacy, and data collection requires interdisciplinary expertise and is essential to study human behavior responsibly. In this dissertation, we take up this challenge by exploring the use of Artificial Intelligence (AI) in the context of behavioral science studies, designing and building effective conversational AIs for information collection and informed consent.

This thesis starts by focusing on surveys, one of the most widely-used research methods in behavioral research. We studied conversational AIs to address today's survey research challenges: survey fatigue, inflexible survey structure, and lack of personalization. In an AI-driven conversational survey, a conversational agent asks questions, interprets a participant's responses, and probes answers whenever needed. We first studied an AI-driven conversational survey's response quality and participant engagement by comparing it with form-based surveys. After establishing the promise, We improved a conversational survey by equipping the AI agent with active listening skills through a human-in-the-loop framework. We further built a novel knowledge-driven language model to generate informative follow-up questions on the fly.

We then looked at the ethical practices in behavioral science research, informed consent procedure, in online studies. Due to the lack of a researcher's presence and guidance, online participants often failed to make informed participation decisions, putting them at unaware risks. In this study, we re-introduced interactivity to online informed consent using conversational AI. Our agent guided participants through the consent form step-by-step and answered their questions. Compared to the form-based interaction, we found the AI-powered chatbot improved consent form reading, promoted participants' feelings of agency, closed the power gap between the participant and the researcher, and ultimately benefited the study quality.

We ended this thesis with a series of empirical studies about how people interact with such conversational AIs. Drawing from the rich use of voice assistants, we considered

voice as another modality. We studied how voice assistants with different social metaphors influence people's reactions and perceptions of their information requests. We then deployed a conversational AI in the real world to collect students' team preferences and demonstrated how such an agent improves the student teaming experience.

This dissertation provides both empirical evidence of how to design effective conversational AI to understand human behavior at scale and technical frameworks to build such an agent. Most importantly, it contributes to design implications for future technologies to improve our understanding of how we interact with each other and our environment and push this research field forward.

*Dedicated to my beloved parents and families, who have been my isle through the tumultuous waves of life, guiding me with their unwavering love and support.*

## ACKNOWLEDGMENTS

During those countless drives through the cornfield, the haze cleared and obscure ideas started to glimmer with a strange new light. But more often than not, my thought scattered like leaves on the wind, carried away by the winter gusts that swept across the endless Midwest plains. The snow-swept streets of Urbana-Champaign beckoned me, as did the sight of cows and horses grazing near St. Mary Rd and the disappearing crimson line at the west edge of Windsor Rd. Memories of rainy drives along the Pacific coast, moist country roads by the Mississippi river, and desolate stretches of the Nevada desert flooded my mind, each a symbol of the many paths to the void. Despite this, amidst all this uncertainty, I am grateful for the companions who have accompanied me on these adventures. Their presence offered a comforting sense of stability in the midst of all this existential wandering.

First and foremost, I want to express my gratitude to my advisors, Prof. Hari Sundaram and Prof. Karrie Karahalios. They have played an instrumental role in my journey, providing invaluable guidance and support during many uncertain and rough times. Hari taught me the purpose of research and how to work with people. I treasure our wonderful conversations about causality, decision-making, bayesian inference, and many more. Karrie has always been the source of wisdom and support. She always pointed me to the light when I was lost in the labyrinth of science. Both are true role models whose impact on my life is beyond words. I aspire to follow in their footsteps and continue to seek the truth about science and humanity.

I sincerely appreciate my committee members, Dr. Michelle Zhou, Prof. Heng Ji, and Prof. Brent W. Robert. Their diverse experience and expertise broaden my views. Heng introduced me to natural language processing, and Brent guided me on the quest of understanding individual differences. From taking their classes, attending their office hours, and receiving their careful, generous feedback, I have gained invaluable insights into finding research ideas and tackling interdisciplinary problems. I'd especially like to thank Michelle for her tremendous support throughout my Ph.D. journey. Without Michelle, this dissertation may never have been written. I was fortunate to learn from her, work alongside her, and collaborate with her.

I had the great fortune of working with many fantastic mentors and collaborators. Thanks, Prof. Wai-Tat Fu, for taking me to the Cascade Lab, where everything begins. I want to thank Prof. Brian S. Woodard and the rest of the spatial visualization team for showing me the real-world impact I can make through my research. I would also like to thank Prof. Dov Cohen, who enlightened me on many important decisions since my undergrad. I thank Prof.

Brain Bailey, who significantly shaped my approach to teaching. I appreciate all the tips and conversations with Dr. Q. Vera Liao, who guides me and inspires me along this journey.

I met many of the characters in this dissertation during internships at Juji and Spotify Research. Thanks to Dr. Huahai Yang and Wenxi Chen for all the discussions on hot and sunny days about the future of artificial intelligence. Thanks to Dr. Sarah Mennicken and Dr. Jenn Thom for opening my eye to music recommendations and voice interactions and for the most memorable internship in an unusual time.

Deciding to do my Ph.D. in UIUC Computer Science was one of the best decisions I've ever made. I'm grateful for the support and wisdom from — and random chats with — Rick Barber, Patrick Crain, Ti-Chung Cheng, Robert Deloatch, Himel Dev, Jinda Han, Emily M. Hastings, Silas Hsu, Jennifer G. Kim, Aditya Karan, Adit Krishnan, Vinay Koshy, Sneha Krishna Kumaran, John Lee, Tiffany Wenting Li, Joon Sung Park, Sebastian Rodriguez, Wendy Shi, Kristen Vaccaro, Junting Wang, Wayne Y. Wu, and Yu-Chun Yen. Mostly, I want to thank Sanorita Dey, Hyo Jin Do, Mingkun Gao, Yubin Ge, Yi-Chieh Lee, Helen Wauck, Chi-Hsien Yen, Pingjing Yang, and Shiliang Zuo. Together, we have been through a rough but meaningful time.

Doing a Ph.D. is like driving on unexpected roads. I am grateful for every person and little creature living on the Tomaras Ave. I especially want to thank, Janice Ng, Terry Kwan, Yijun Lou, Jason Jinsong Cui, Muqi Zou, and Yu Xiong. You are my family and my navigators. I want to thank Lyosha Svinarski for providing the analog cave for me to escape from the digital world and for many hot summer random drives on the flat Midwest plains.

I'm infinitely grateful to my parents, who taught me how to perceive and approach the world with a peaceful mind. They nurtured an essential appreciation for learning. Thank you for letting me pursue the quest of uncovering the unknowns, even if it means I have to be thousands of miles away from you.

Special thanks to Yuqi Yao. Your presence, along with Bobo, has made this journey a more enjoyable one. Thank you for being brave and supporting every decision I made. All the things that make you the last person in these acknowledgments.

## TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION . . . . .	1
CHAPTER 2 RELATED WORK . . . . .	7
2.1 Survey Research . . . . .	7
2.2 Improving Survey Quality . . . . .	8
2.3 Consent Form Reading . . . . .	8
2.4 Improving Consent Form Reading . . . . .	9
2.5 Power Relation in Human Subject Research . . . . .	10
2.6 Feedback in Recommender Systems . . . . .	11
2.7 Information Collection In Situ . . . . .	12
2.8 Conversational AI and Natural Language Interface . . . . .	13
2.9 Conversational AI for Information Elicitation . . . . .	13
2.10 Evaluating Conversational AI . . . . .	15
2.11 Recent Advances in Conversational AI . . . . .	16
2.12 Designing the Social Aspects of a Conversational AI . . . . .	17
CHAPTER 3 CONVERSATIONAL AI FOR ONLINE SURVEY . . . . .	19
3.1 Overview . . . . .	19
3.2 Method . . . . .	22
3.3 Results . . . . .	35
3.4 Summary of Findings . . . . .	44
3.5 Discussion . . . . .	44
3.6 Limitations . . . . .	49
3.7 Design Implications . . . . .	54
3.8 Conclusions . . . . .	56
CHAPTER 4 CONVERSATIONAL AI WITH ACTIVE LISTENING SKILLS . . .	57
4.1 Overview . . . . .	57
4.2 Prototype Overview . . . . .	60
4.3 Prototype Design and Key Components . . . . .	61
4.4 Evaluation . . . . .	68
4.5 Limitations . . . . .	74
4.6 Design Implications . . . . .	76
4.7 Conclusion . . . . .	77

<b>CHAPTER 5 A KNOWLEDGE-DRIVEN APPROACH FOR FOLLOW-UP</b>	
<b>QUESTIONS GENERATION</b>	78
5.1 Overview	78
5.2 Dataset and Annotation	80
5.3 Models	81
5.4 Gricean Scores	83
5.5 Experiments	85
5.6 Error Analysis	89
5.7 Conclusion	90
<b>CHAPTER 6 CONVERSATIONAL AI FOR ONLINE INFORMED CONSENT</b>	91
6.1 Overview	91
6.2 Method	94
6.3 Results	104
6.4 Discussion	111
6.5 Limitations	115
6.6 Conclusion	116
<b>CHAPTER 7 USING A CONVERSATIONAL AI TO UNDERSTAND INDIVIDUALS AND HELP TEAMING</b>	118
7.1 Overview	118
7.2 INDIGO Field Deployment	121
7.3 Results	123
7.4 Discussion	134
7.5 Design Implications	135
7.6 Limitations	135
7.7 Conclusions	136
<b>CHAPTER 8 DESIGNING VOICE INTERFACE FOR COLLECTING EXPLICIT USER FEEDBACK</b>	137
8.1 Overview	137
8.2 Method Overview	139
8.3 Identifying Valuable Feedback Categories Through Expert Interviews	140
8.4 Scenario-Based Online Study	143
8.5 Results	150
8.6 Discussion	156
8.7 Conclusion	159
<b>CHAPTER 9 FUTURE DIRECTIONS</b>	160
9.1 Interactive Evaluation of Conversational AI	160
9.2 Power-Aware Interaction	161
9.3 Understanding Human Behavior through Everyday AI	161
9.4 AI for Social Science	163
<b>CHAPTER 10 CONCLUSION</b>	164

REFERENCES . . . . .	166
APPENDIX A APPENDIX FOR A KNOWLEDGE-DRIVEN APPROACH FOR FOLLOW-UP QUESTIONS GENERATION . . . . .	197
A.1 Implementation Details . . . . .	197
A.2 Data Annotation Process . . . . .	197
A.3 Dataset Statistics . . . . .	198
APPENDIX B APPENDIX FOR CONVERSATIONAL AI FOR ONLINE INFORMED CONSENT . . . . .	199
B.1 Manipulation Checks . . . . .	199
B.2 Bayesian Model Specifications . . . . .	199
B.3 Model Convergence . . . . .	202

## CHAPTER 1: INTRODUCTION

Understanding human behavior on a large scale is critical to addressing some of the most pressing challenges in today's society. By gaining insights into why people act the way they do, we can build explanatory theories about our behaviors, create artificial intelligence that can best work with humans, and design informed policies that can positively impact society as a whole in areas such as public health, sustainable economy, and fair education [1, 2, 3]. For instance, when it comes to public health, understanding human behavior can help doctors and policymakers design effective interventions to promote preventative behaviors, such as physical distancing or vaccination, that can potentially stop the spread of a global pandemic.

However, studying human behavior is a challenging task. Our unique behavior and decision-making process are shaped by various intertwined factors, including cultural, social, environmental, and individual characteristics [3, 4, 5, 6]. Such a multitude of factors makes it difficult to isolate specific causes and effects. Collecting and analyzing data on human behavior at scale is also challenging, especially when we want to draw conclusions that are generalizable to different groups of people, as it often requires diverse, representative, and sometimes longitudinal datasets. Such datasets are difficult to obtain and maintain, let alone data biases and measurement inaccuracies [7, 8, 9]. Moreover, research on human behavior at scale raises ethical concerns related to privacy, informed consent, and potential misuse of information [10, 11]. Balancing the benefits of understanding human behavior with these ethical considerations is a critical challenge.

How do we conduct ethical, robust, and generalizable studies about human behavior? Let's consider the following scenario. A doctor wants to study how people protect themselves in reacting to the novel pandemic. First, the doctor needs to find the right person to observe, in this scenario, people who are influenced by the disease in different situations and take different actions. Then, the doctor needs to record an individual's behavior while taking their actions and ask good questions to elicit why they are doing that, what affects their prevention behavior, and how other people influenced their actions. And the doctor needs to build a relationship with the person to gain their trust and collect the most authentic and unbiased answers while respecting their privacy. By understanding this individual's behavior, the doctor could quickly identify effective strategies to help that individual. A doctor can observe a few individuals' behaviors and help them, but what if there are millions of people from different locations, different cultures, and speaking different languages?

Therefore, sophisticated tools and methods are required to adequately capture the context and nuances and interact with study participants to collect authentic answers. And most

importantly, given the sensitivity of personal data and its implication for people's privacy, we need to understand the ethical concerns of such tools and methods deeply.

Motivated by the fundamental question of understanding humans at the scale, in this thesis, we explored the use of conversational AIs in behavioral science studies, designed and built effective and engaging conversational agents for information collection and informed consent, and studied how people interacted with such AIs in a real-world context.

Conversational AI enables computers to comprehend and respond to natural language through natural language interfaces [12, 13, 14]. It involves natural language processing (NLP), machine learning (ML), human-computer interaction (HCI), and other techniques to create chatbots, voice assistants, and other conversational interfaces that can naturally interact with people through text or speech. The conversational AI aims to make human-computer interactions more intuitive and natural, enabling people to communicate with machines as they speak to others. Some examples of conversational AI applications include chatbots that provide customer service, virtual assistants that help schedule appointments, and voice assistants that can recommend music and facilitate grocery shopping [13, 15, 16, 17].

Compared to traditional graphical interfaces, conversational AIs offers additional interactivity through dialogue. Interactivity is vital in human-computer interaction because it allows for a two-way exchange of information between people and the computer. This exchange enables the computer to respond to people's actions and provide feedback, improving the user experience and making it more engaging. Interactivity also allows for more natural and intuitive interactions with technology, as people can interact with computers in ways that are similar to how they interact with other people or objects in the real world. This can make technology more accessible and easier for a broader range of people.

However, building conversational AI that can truly converse is not a trivial task. It requires creating systems that can understand natural language and respond appropriately in real-time [14, 18, 19, 20]. Several challenges make this a complex task. First, it is challenging for a conversational AI to understand natural language. Language is full of ambiguity, including homonyms, synonyms, and idioms. Conversational AI must be able to disambiguate words and phrases to understand what the individual is trying to say [18]. Understanding language often requires understanding the context in which it is used. For example, the same word can mean different things depending on the situation or the speaker's tone. Moreover, people use language in various ways, such as using regional dialects. Conversational AI needs to handle this variability and understand what the individual means.

Second, it is difficult to build conversational AIs that can effectively handle complex conversation situations. Natural language conversations are nonlinear and often go back and forth [14, 20]. In a conversation, an individual may digress from a planned agenda for various

reasons. For example, some people may not understand a conversational agent’s utterance and want clarifications, while others might also be “uncooperative” and intentionally provide gibberish or digress conversation topics. Also, people may often refer to a topic in chat history and jump across different topics.

Third, it is challenging for conversational AI to respond to people in a natural, conversational way [19, 20]. A skillful conversationist should not only be able to express their own view clearly but also listen actively. In some cases, the conversational AI also needs to respond empathetically and convey understanding. Individuals may prefer different conversational styles and interpret the meaning based on their experience. An effective conversational AI should personalize the dialogue experience by tailoring responses to individual preferences. Current real-world conversational AIs mostly rely on conversation designers to create a preset of responses or templates, which limits the conversational AI’s ability to respond to various conversational contexts [14]. Although end-to-end dialogue systems show significant improvement, especially with large language models, the lack of control poses another challenge to power a conversational AI that can always converse responsibly [21].

The challenge exacerbates with the goal is to building conversational AIs to understand human behavior, where the conversational AI needs to be more proactive, drive the conversation, encourage authentic answers, and ask meaningful questions. Overcoming these challenges requires advanced techniques in natural language processing, machine learning, and human-computer interaction, as well as extensive testing and tuning to ensure that the conversational AI system works effectively in a wide range of situations.

In this thesis, we aimed to tackle those challenges in a sequence of studies about using conversational AIs to elicit information from people to understand their preferences and actions. We approached those challenges through two perspectives, *machine perspectives*, where we built systems and natural language models for accessible and skillful chatbots to conduct conversational surveys, and *human perspectives*, where we studied how people interact with, respond to, and perceive such conversational agents with a variety of designs in the real world.

**Machine perspectives.** How can we *build* engaging conversational agents to collect rich and truthful information through surveys? Building an effective interview conversational agent faces three major obstacles. First, the agent needs to effectively grasp and respond to participants’ highly diverse and often complex natural language input. Second, the agent needs to handle complex conversation situations to deliver interview skills such as prompting and probing and to ensure survey completion. Third, it is difficult for researchers who don’t have AI expertise or resources to take advantage of AI advances.

Chapter 4 examines the feasibility of using practical AI to build effective conversational

agents for survey interviews. we built a data-driven framework with human-in-the-loop that enables an interview chatbot to interpret diverse respondents' inputs and generate proper responses with active listening skills. The framework contains a human-AI collaborative annotation module that prepares high-quality training data from small, unstructured real-world conversations and a classifier to trigger rules at run time to guide the generation of proper responses, enabling active listening. The framework supported reproducibility, practicality, and scalability by focusing on publicly available techniques, low-resource models, and modular design. In the evaluation, the interview chatbot with active listening skills collected more informative and relevant responses while creating more engaging experiences. This work offers practical implementations to power chatbots with active listening skills, a hybrid chatbot design framework for developing progressive chatbot platforms, and design implications for building empathetic chatbots beyond interview tasks.

Having established the feasibility of using conversational AI to conduct surveys, in the following step, Chapter 6, we focused on the challenge of static survey structure and lack of personalization, e.g., how to enable conversational AI to generate follow-up questions to ask for more relevant and meaningful information. Interview chatbots today cannot parse the context of participants' responses and don't know what to ask next. This results in ineffective follow-up questions like "could you say a bit more?". We wield a knowledge-driven approach with a large language model to accomplish sophisticated follow-up question generation. We performed multi-task training to select salient knowledge entities and relations pairs in the context to teach the model what to ask. Subsequently, by identifying key contexts and knowledge, the model used prompt learning to generate follow-up questions that are coherent with the context and always aim for the most meaningful, deep, and interesting information. The model was deemed effective in both objective metrics and expert evaluations. With this work, future survey designers would create more dynamic and personalized surveys without writing complicated rules and tons of question templates.

**Human perspectives.** Building a conversational agent to ask questions is merely the first step; less known is how people will react to such an agent. We conducted a series of studies to answer the following questions,

- Could an agent drive an engaging survey while eliciting high-quality information? (see Chapter 3)
- Could an agent effectively deliver an informed consent process and improve study quality? (see Chapter 5)
- How does such an agent perform in the wild? (see Chapter 7)

- How could we better design an agent to request information through voice? (see Chapter 8)

A conversational agent can frame survey questions in a more personalized manner, provide human-like social interactions, and encourage information exchange in a natural way while bearing the risk of inaccurately interpreting and improperly responding to humans' diverse free-text input due to its limited natural language capability. To see if the agent could conduct an effective survey, we conducted a field experiment that compared the outcomes of an agent-driven survey and that of a typical form-based survey (see Chapter 3). To evaluate survey quality, we proposed a set of evaluation metrics based on Gricean Maxims, including the dimension of quantity, quality, relevance, and manner. Through an in-depth mixed-method analysis of over 5,200 free-text responses, we found the agent not only made the survey more engaging but also collected significantly more relevant, specific, and clear responses. This study indicates a promising but currently under-utilized way of creating more effective and engaging survey through conversational agents.

We then looked at another important step in online human subject research, informed consent (see Chapter 6). Through the informed consent process, participants learn about the study procedure, benefits, risks, and more to make an informed decision. However, recent studies showed that current practices might lead to uninformed decisions and expose participants to unknown risks, especially in online studies. Without the researcher's presence and guidance, online participants must read a lengthy form on their own with no answers to their questions. In this paper, we examined the role of an AI-powered chatbot in improving informed consent online. By comparing the conversational agent with form-based interaction, we found the chatbot improved consent form reading, promoted participants' feelings of agency, and closed the power gap between the participant and the researcher. Our exploratory analysis further revealed the altered power dynamic might eventually benefit study response quality.

How does such an agent perform in the wild? We built a conversational agent that can chat with students to learn teaming preferences (see Chapter 7). Our goal was to replace traditional teaming surveys to gather initial information from team members, such as team preferences and expectations. We found that students enjoyed conversing with the chatbot and offered open and honest input. The agent also helped instructors to understand student team preferences and perceptions better. This study indicated a viable path to create an AI research assistant who can follow a team and continuously interact with team members during their efforts.

Above studies in this thesis focus on text-based conversational agents. Given the recent

development of voice assistants, we believe that voice interaction could further open the interaction space with the new affordances. We examined three design dimensions of an agent that aims to collect people's feedback for recommender systems in-situ: who the agent is, how the agent should act, and when the agent should act (see Chapter 8). This study has intriguing results: people are more likely to respond to the agent's feedback request and perceive it as less disruptive; compared to direct questions, giving users the instruction to provide feedback creates less experience friction; and when the agent becomes more proactive, people prefer the instruction more. Our results allowed us to create more effective conversational agents to understand human behavior in situ.

In summary, motivated by the challenge of understanding humans at scale, the thesis studied conversational AIs in behavioral science studies, focusing on building effective and engaging conversational agents for data collection and informed consent. The thesis tackled these challenges through both machine and human perspectives. The machine perspective focuses on building engaging conversational agents for survey interviews and generating informative follow-up questions. The human perspective examines the effectiveness of agent-driven surveys and informed consent processes, how people interact with such an agent in real-world contexts, and designing agents for voice interaction. The thesis demonstrated that conversational agents could be built to deliver engaging surveys, collect higher-quality information, and improve consent form reading in online human subject research. It also showed that conversational AI could be a valuable tool for collecting rich information in real-world contexts and provided insights into designing more effective voice interactions for in-situ human behavior understanding. This thesis further offered multiple design implications for future virtual research assistants that can engage with people responsibly and, most importantly, assist researchers in operationalizing their intuitions and curiosities about how humans think and behave.

The rest of the thesis organizes as follows. In Chapter 2, we will first give an overview of the landscape of existing literature and how this thesis fills the knowledge gap. We will then dive into a specific problem in human subject research online survey method and explore how conversational AI could improve today's survey method. In Chapter 3, 4, 6, we conducted a series of studies to build and evaluate skillful chatbots to conduct effective conversational surveys. In Chapter 5, we moved from information collection to another important yet underlooked procedure in online human-subject studies, the informed consent process. In Chapter 7, 8, we examined the utility of using the conversational agent in real-world settings and important design metaphors that change how individuals interact with such an agent. In Chapter 9, we conclude this thesis with future directions.

## CHAPTER 2: RELATED WORK

In this chapter, we will review related work that motivated my dissertation. We will start by presenting challenges faced by survey research and online informed consent in current behavioral science research. Then we will review conversational AI applications and gaps in building and evaluating effective conversational agents with different modalities. To motivate our empirical studies in real-world contexts, we will further review the literature regarding understanding individual differences and eliciting human feedback in situ.

### 2.1 SURVEY RESEARCH

Conducting surveys is a key method to collect data in behavioral science research [22]. In recent years, self-administered online surveys have replaced paper-and-pencil surveys and have become one of the most widely used methods to understand people's behavior, preferences, and decisions [23, 24].

Due to the extensive use of online surveys, survey fatigue is a challenge anyone who wishes to collect data faces. In [25], authors identified two typical types of survey fatigue, *survey response fatigue* and *survey-taking fatigue*. Studies show that when a survey grows in length, survey respondents may pay less attention to each question and stop responding to the increasing survey requests [26, 27].

Due to the added time and effort needed to formulate and type answers to such queries, surveys with open-ended questions make the issue worse [26, 28]. Open-ended questions are a crucial tool for gathering insightful information and are frequently used in self-administered online surveys [22]. Particularly when the answers cannot be pre-defined or the pre-defined answers can generate biases, open-ended questions allow respondents to formulate their answers freely [28, 29]. Moreover, open-ended questions enable the collection of deeper information, such as the context and justifications for the responses. Riiskjaer et al. [30] found the survey with open-ended questions elicits more useful information from patients in the hospital setting. However, studies showed open-ended questions, on the other hand, frequently lead to more cognitive overload and respondents are more likely to skip them or give mediocre or even irrelevant responses [26, 28, 31]. As a consequence, survey-taking fatigue adversely affects the response validity and reliability especially when open-ended questions are involved [28, 32, 33]

To combat survey-taking fatigue especially to motivate and guide survey participants to provide quality answers to open-ended questions, in this thesis, we first leveraged a novel

technology, conversational AI. We designed and built conversational agents to foster natural and engaging information exchange through dialogue and encourage deep and rich information with interview skills.

## 2.2 IMPROVING SURVEY QUALITY

This thesis is related to survey methods for improving survey quality. Researchers have put tremendous effort into improving sample validity and response quality. Heerwegh and Loosveldt [34] find that personalization can significantly increase web survey response rate by 4.4% while not necessarily leading to social desirability response bias. Behr et al. [35] have tested three probing variants and found that such probings lead to a higher number of meaningful answers in web surveys. In a field experiment with over 4000 participants, De Leeuw et al. [36] have shown that a polite probe can successfully reduce item non-response (“don’t know”) without sacrificing the reliability of the final answers. Conrad et al. [32] also show that interactive feedback can improve the quality of responses in web surveys. Additionally, Oudejans and Christian [33] propose using explanations and probings to improve the quality of user responses to open-ended questions. Smyth et al. [37] showed the effect of textbox size and instructions on improving people’s response quality to surveys with open-ended questions.

On the one hand, our work is built on top of existing findings. For example, we learned that interactive features, such as response feedback and probing, help improve participation and response quality. On the other hand, our thesis is a natural extension of existing efforts. In particular, we explore the use of conversational agents to offer various interactive features in a survey, hoping that such features would improve participant engagement and response quality.

## 2.3 CONSENT FORM READING

The concept of informed consent is embedded in the principles of many ethical guidelines including the Nuremberg Code, The Declaration of Helsinki, and The Belmont Report [10]. Four core elements ensemble the informed consent process, including disclosure, comprehension, voluntariness, and competency. Through the informed consent process, participants will learn about the study’s purpose, procedure, risks, and benefits to make an informed decision.

Despite its importance, people often sign the consent form without a thorough read, regardless of whether it is about a clinical trial that may risk their physical and mental health

or a study about their political opinions. For example, Lavelle-Jones et al. [38] found that 69% of patients preparing to undergo various surgical procedures signed the consent form without reading it carefully. Varnhagen et al. [39] showed that in a study regarding technology use, study participants could only recall less than 10% of the information contained in the form, and 35% of participants reported that they only skimmed the form or did not read it at all. Cummings et al. attributed the discrepancy between people's study participation decisions and their concerns regarding confidentiality, anonymity, data security, and study sensitivity in studies with open-data sharing practices to participants' inattentiveness to the consent form [40]. Consent form reading has become increasingly challenging as it moves online [41, 42, 43]. An online consent form is low-cost and easy to administrate, in addition to its broaden-reach across the internet. However, compared to the in-lab setting, no researcher will guide the participant through the consent form, explain the content, and clarify the participant's question [44, 45, 46]. Therefore, online informed consent often yields less informed participation decisions, especially when the study is more complicated and riskier.

Ineffective informed consent reading not only puts participants under unaware risks but also harms the study's validity and data quality [47, 48, 49]. Consent forms contain important information about study procedures and purposes, and comprehending such information may determine the success of later study manipulation, especially for deception studies with cover stories. Unaware confounding factors may further contribute to the replication crisis in social sciences.

Altogether, while the informed consent process plays a vital role for both participants and researchers in a research study, the current practice of conducting an informed consent process has many weaknesses, especially since more studies have started to collect participants' consent online. In this thesis, we aim to improve the online consent form reading by exploring novel interaction techniques, e.g., an AI-powered chatbot.

## 2.4 IMPROVING CONSENT FORM READING

While many prior studies have focused on improving consent form reading, the study results were not always consistent [44, 45]. One group of researchers focused on the design of the consent form, including text readability [50, 51, 52], length [41, 53, 54], layout [55], and media [56, 57, 58]. Dresden and Levitt [53] found study participants could retain more information from a consent form with less unnecessary information and simpler vocabulary. However, other studies found a concise consent form may not yield a higher comprehension score [52, 54]. Although study participants advocate for a shorter form with simpler language,

due to the regulation and the nature of a study, it is often difficult to achieve [42]. Researchers also explored converting text-based forms into multimedia. For example, Friedlander et al. found the utility of using video to deliver a consent form in increasing people’s engagement and comprehension [57]. However, there is no conclusive evidence of its effectiveness according to multiple meta-analyses [44, 59].

Another group of researchers brought interactivity into the informed consent process to create an engaging and personalized experience that facilitates consent form reading [46, 60, 61]. In an in-person setting, letting the researcher go through the consent form and answer the participant’s questions is deemed the most effective and desirable [46]. As more studies move online, researchers have started to explore new interactive features to improve online consent [60, 61]. One effective intervention is to test the participant’s knowledge about the consent form before signing [62]. Further, Balestra et al. used social annotations in online consent forms to facilitate consent form comprehension [61]. Bickmore et al. built an embodied agent that can explain a medical consent form to the reader and found people to be more satisfied with the agent that can tailor its explanation to the participant’s existing knowledge [60]. Although not all attempts were successful, many studies emphasized the importance of interaction with the researcher to ensure participants’ understanding of the consent information and to foster trust, especially for complex and risky studies [63].

We studied an AI-powered chatbot with a state-of-art hybrid chatbot framework. Compared to prior work [60, 64] where the agent largely relies on a rule-based system without taking natural language input, recent language technologies enabled our chatbot to answer a diverse set of questions in natural language and deliver engaging experiences with multiple conversational skills. Such a chatbot opens up a new opportunity to bring humanness back to the informed consent process and make the informed consent process more effective.

## 2.5 POWER RELATION IN HUMAN SUBJECT RESEARCH

In a power relation, the person with lower social-formative power is often constrained by their superior counterpart [65]. In the context of human subject research, the researcher has often been viewed as having total authority and being able to decide the resource distribution whereas the participant is often sitting at the lower end [66]. The power asymmetry between the researcher and the participant results from the researcher’s control over the participant’s recruitment, treatment, data, and compensation [11]. Such a power gap inhibits the participant’s autonomy, decreases study engagement, and deters authentic answers [11, 66].

Many researchers advocate for power redistribution to close the power gap in human subject

research for both ethical and data quality considerations [67, 68]. By reducing the power gap, the participants could be more engaged during the study, more comfortable disclosing their true thoughts, and more cooperative with the study procedure, which may ultimately benefit the study quality. For example, Chen analyzed the interviewer’s language use and found the reduced power gap encourages data richness [69]. However, some studies warn of the importance of maintaining the distance between the researcher and the participant for professional judgment [70, 71].

By sharing information about the study procedure, clarifying risks and benefits, and elaborating on the participant’s rights, the informed consent process is designed to close the information gap and ensure the participant’s autonomy [72]. Additionally, as one of the earliest interactions happens between the researcher and the participants, the informed consent process provides an excellent opportunity to redistribute power and establish trust. Kustatscher [73] used visual aids to improve the informed consent for children and found such an engaging process altered the power relation and created a more comfortable environment for information disclosure.

We further extended prior knowledge on the power relation in the research setting by showing the improved informed consent process with an AI-powered chatbot could close the power gap. Through our path analysis, we found the observed effect on study quality can be attributed to the altered power relation.

## 2.6 FEEDBACK IN RECOMMENDER SYSTEMS

Signals from the user, such as user behavior logs or explicit user feedback, are crucial in building and evaluating machine learning models. From various user signals, the system can learn and make inferences about a user’s interests and preferences [74, 75], and system builders can use those signals to understand when and how the model failed to deliver engaging experiences [76].

Generally, there are two types of user signals: implicit and explicit. Implicit user signals are often derived from user interaction logs, such as the engagement duration, the amount of scrolling on a webpage, skips, or likes. System builders need to make assumptions about implicit signals to understand user intention or preference, which often introduces a lot of noise [77]. In contrast, explicit user signals capture user preferences in a more direct and granular way. Often, such signals come from user ratings, survey questions, or other mechanisms that allow users to explicitly express their preferences or intentions [74, 78]. Taken together, implicit and explicit signals can provide a comprehensive view of the user experience.

In this study, we focused on one type of explicit user signal: user feedback. Feedback is crucial for developing recommender systems, from early collaborative filtering techniques to deep learning models. These systems learn user preferences through feedback, such as ratings and reviews [75, 79]. Machine learning practitioners also rely on user feedback to inspect system failures [80, 81] or discover user patterns [82].

However, eliciting explicit user feedback is challenging, as it can introduce more effort for the user, and self-reported data is not always reliable [83, 84]. Although people have studied how to elicit user feedback through traditional interfaces [85], using voice assistants to collect feedback spoken by the user to the assistant is underexplored.

## 2.7 INFORMATION COLLECTION IN SITU

Collecting data from individuals in situ, in the place and in the moment in which the interaction of interest occurs, has proven invaluable in many diverse fields, such as psychology [86], anthropology [87], and human computer interaction [88, 89]. Eliciting user feedback this way can mitigate memory biases during recall [90], ensure the collected data took place within the context of interest [88], and obtain richer contextual information [91].

There are several methods for collecting in situ feedback in research. For example, diary studies in which participants are instructed to note down self-observations that are of interest to the researchers; or the Experience Sampling Method [88, 92], where a system randomly probes and checks if participants are in the context of interest to then collect data in situ. Due to the increasing capabilities of mobile sensors, such systems can sometimes even detect the user context and proactively ask for the feedback. For example, Bachmann et al. propose a data collection tool that prompts users with context-aware triggers [93].

Since both implicit and explicit user signals are highly affected by the user's context [94], it is especially valuable to collect user feedback in situ. To do so, front-end developers and user interface designers often include feedback buttons, such as Like buttons, or text input boxes for suggestions on the interface [95]. However, such passive methods are often ignored by the user or result in data biased towards people with complaints [96]. To make the feedback collection more noticeable, sometimes these prompts for user feedback are triggered by certain user interactions [93]. However, such proactive methods can create a significant amount of unwanted friction to the user experience or unnecessary user burdens.

In the context of voice assistants, it is nearly impossible to passively collect user feedback, as we cannot just add a visual feedback button to a voice interface. Even if the voice assistant is on a device with a screen, we cannot be sure the user's focus of attention is on the screen without additional forms of measurement (e.g., eyetracking) [97]. But given the value of

explicit *in situ* user feedback, we set out to explore the potential design space for proactive information elicitation via voice assistants. For this, we evaluated different designs to identify design decisions that can alleviate some friction and create more willingness for people’s responses.

## 2.8 CONVERSATIONAL AI AND NATURAL LANGUAGE INTERFACE

Our work is related to a rich body of studies on the applications of conversational AI. For example, past studies have examined chatbots as a personal assistant [98], intelligent tutor [99], customer service agent [15, 100], job interviewer [101], and worker’s companion [102].

The HCI community has long been interested in the interaction benefits offered by conversational interfaces. The general consensus is that conversational interfaces offer several advantages over traditional WIMP (Windows, Icons, Menus, and Pointers) interfaces [12, 103]. First, conversational interfaces offer a natural and familiar way for people to express themselves, which in turn improves the usability of a system. Second, such interfaces are flexible and can accommodate diverse user requests without requiring people to follow a fixed path [20]. Third, such interfaces are often personified and their anthropomorphic features could help attract attention and gain trust [104]. Inspired by the potential advantages of conversational interfaces, our work investigates the use of conversational interfaces for survey methods. Differing from existing works on conversational interfaces, we are exploring a new application of conversational AI for conducting surveys, which has its own set of unique challenges. For example, survey participants would not be as motivated or cooperative as job candidates who interact with a conversational AI agent [101, 105]. Neither would survey participants be as tolerant or patient as a student or an employee who interacts with an AI companion [102].

However, conversational interfaces are far from perfect due to technical difficulties in processing natural language expressions and managing diverse and complex conversation situations [14, 20]. It is thus unknown how the shortcomings of conversational interfaces (e.g., failure to understand an individual’s input during a survey) might influence people’s experience in survey and consent form reading. Therefore, we hope to investigate whether and how conversational interfaces might bring practical value to traditional behavioral science research practices, which use WIMP-based interfaces to collect information.

## 2.9 CONVERSATIONAL AI FOR INFORMATION ELICITATION

Our work is directly relevant to the use of conversational AI for information elicitation. Researchers have built various AI agents to elicit information from an individual through a one-on-one, text-based conversation, such as eliciting information from a job candidate

[101] and gathering data from a study participant [106]. Information elicitation may serve various purposes. A common task is to elicit “parameters” for information retrieval or recommendation [107, 108, 109]. This system often supports task-oriented conversations, such as helping a customer book a flight reservation, finding a desired restaurant, and scheduling an event [110, 111, 112]. The main goal of such systems is for an agent to gather required information (e.g., travel dates) to perform a task [113]. More recently, conversational AI has been explored as a means to elicit information for additional purposes beyond fulfilling a specific task. For example, there have been agents that elicit information for recommending products, movies, and jobs [114, 115], group decision support [116], psychotherapy [16, 102], and voting [117]. An emerging area is using conversational AI to conduct in-depth interviews [101, 105] or longitudinal studies in the real world [102, 106]. Moreover, studies show that conversational AI may enhance user engagement and encourage self-disclosure. Prior work has demonstrated that the exhibited social behaviors of agents are effective in improving user engagements in various social settings by a set of metrics, such as interaction duration, breadth, and depth of self-disclosure (e.g., discussing personal subjects), and a positive attitude toward the agent and interaction outcome [116, 118, 119].

On the other hand, people’ positive attitude toward AI agents has raised concerns about privacy protection and encouraged studies to investigate the effect of chatbots on user trust and privacy in the context of information elicitation. For example, a recent study showed that people trusted a customer service chatbot more if they were well-informed in the conversation where the information was stored [117]. Sannon et al. found that people were less likely to share personal sensitive information (e.g., financial or health stress) with a personified chatbot [120]. However, none of the existing studies compared survey participants’ behaviors (e.g., self-disclosure and answer quality) influenced by a chatbot versus in a traditional online survey. We thus set out to gauge how conversational AI affects user engagement and the quality of survey results, hoping to discover new survey methods that may improve traditional online survey practices.

Different from our investigation of using a chatbot as a general surveying tool, prior studies tended to focus on examining the feasibility and effectiveness of a chatbot for a specific information elicitation task. However, unlike our work, none of the existing studies compared the effectiveness of a chatbot in information elicitation with that of a traditional approach. In particular, we want to quantitatively measure the holistic effect of chatbots on user engagement and response quality compared to that of a traditional online survey.

## 2.10 EVALUATING CONVERSATIONAL AI

With the advent of conversational interfaces, evaluating the effectiveness of such interfaces is a continuously evolving research topic. Traditionally, the evaluation criteria have been divided into objective metrics on the system performance (e.g., user input interpretation accuracy and user task completion rate) and subjective metrics based on user feedback (e.g., user satisfaction) [121, 122]. Objective metrics are directly computed from logs of the interaction and can be based on task or domain coverage, error rate, number of interaction issues, accuracy or other metrics compared to “ground truth” [121, 123, 124]. Subjective metrics are often based on user opinions around certain aspects, such as satisfaction, and intelligibility, (e.g., [125]). There are also comprehensive methodologies that consider both subjective user satisfaction and objective performance metrics including task success, dialog efficiency (e.g., duration, total turns), and dialog quality (e.g., latency) [122].

In addition to examining user satisfaction and usability of conversational AI agents, HCI researchers have investigated how agent behavior impacts users’ social perceptions, such as trust [126], rapport [127, 128], anthropomorphism, and likability [129]. Such user subjective feedback is often measured by questionnaires, i.e., Likert-scale ratings on self-reported statements. Additionally, automatic methods have been developed to predict user satisfaction based on users’ behavioral signals, such as dialogue acts, words, and user actions [98, 130].

Researchers from ML and NLP communities have shown that standard automatic language evaluation metrics (e.g., BLEU, METEOR) are ineffective for dialogue evaluation [124, 131, 132]. Hence, recent research has proposed various automatic metrics specifically for dialogue generation, which can be divided into two categories, reference-required and reference-free, based on whether a reference utterance is required.

Among reference-required metrics, a typical example is RUBER [133] which combines a referenced metric and a reference-free metric. The referenced metric computes the cosine similarity of word embeddings between a generated response and a human reference, while a reference-free RNN-based scorer measures the relatedness between a generated response and its context. Based on it, BERT-RUBER [134] replaces the RNN in RUBER with BERT [135] to employ contextualized word embeddings as an improvement.

As for reference-free metrics, Lan et al. [136] proposed PONE based on BERT-RUBER to distinguish positive and negative samples from a training set. Besides, MAUDE [137] adopts a similar training paradigm but with a different sampling strategy and uses the predicted score of a generated utterance given its context as the metric. Differently, FED [138] uses DialoGPT to compute the likelihood of manually designed follow-up utterances to measure

multiple qualities of dialog without any supervision.

Compared to the existing work, our study focuses on evaluating the outcomes of a conversational interface with a target goal—collecting high-quality survey data and keeping the survey taker engaged. We, therefore, have proposed several content-based metrics to measure response quality and participant engagement. To measure the quality of a follow-up question, we further proposed *Gricean Scores*, a set of reference-free metrics, with an additional perspective of knowledge graphs. We utilize different pre-trained language models for different measurements, which are aligned well with their pre-training tasks, so that the gap between downstream inference and pre-training may not be large.

## 2.11 RECENT ADVANCES IN CONVERSATIONAL AI

There are numerous computational approaches to building conversational agents, including both symbolic and data-driven approaches [20]. To cope with highly diverse user input, data-driven approaches have been used extensively to handle open-domain conversations. A number of data-driven approaches are used to train retrieval models that find the most probable machine response from a repository of pre-defined responses for a given user input (e.g., [139, 140, 141, 142, 143]). Additionally, generative approaches have been explored to synthesize machine responses that do not exist before (e.g., [141, 144, 145]). However, the quality of generated responses may be erroneous or incoherent, not yet ready for practical applications.

Neither retrieval-based nor generative models alone are practical for building interview chatbots since they require large amounts of training data—often millions or billions of conversation exchanges [141, 142, 143]. It is difficult to obtain interview data let alone large amounts due to the private or sensitive nature of many interviews. Moreover, a lack of interpretability and control of data-driven results would put an interview chatbot at risk especially in high-stakes contexts, such as customer interviews [146].

To improve interpretability, recently researchers have explored hybrid approaches. For example, Hu et al. propose to incorporate rules as the weights of neural networks to improve interpretability and performance [147]. Sundararajan et al. propose an approach to identify which input features contribute to the prediction of a deep network [148]. Their approach can extract rules from the networks to help interpret the prediction results and debug the networks. These hybrid approaches have inspired us in developing our prototype, which is perhaps the first of exploring a hybrid framework for building interview chatbots.

We further advanced our hybrid framework with more generative models. Previous attempts have explored follow-up question generation in more specific domains and scenarios.

For example, several attempts focused on graduate school admission interviews [149, 150] and built a small corpus by simulating interviews between participants. They have tried template filling-based and sentence retrieval-based methods for generating follow-up questions. Additionally, job interview is another main domain of the task. To this end, S.B. et al, [151] adopted pre-trained GPT-2 and fine-tuned it on their own collected corpus. By contrast, Inoue et al. [152] pre-defined a set of follow-up questions under different categories and turned the generation task as a question selection problem. Another line of related research lies in question generation in conversational systems. For example, Wang et al. [153] devised typed decoders to model a type distribution over *interrogatives*, *topic words* and *ordinary words*, and used it to modulate the final generation distribution. Motivated by the difficulty in formulating users' complex information needs, Aliannejadi et al. [154] collected a dataset and used a BERT-based retrieval model to retrieve all relevant clarifying questions from a large pool of questions in open-domain information-seeking conversational systems. Generally speaking, most previous works built retrieval models which heavily rely on a sentence corpus for selection as templates or final outputs, and these methods may be hard to be generalized to other datasets and fail to generate diverse questions. By contrast, we focused on the scenario of conversational surveys where the information collection goal is more diverse. We also exploit a knowledge base to guide the generative model to generate more focused and informative questions.

## 2.12 DESIGNING THE SOCIAL ASPECTS OF A CONVERSATIONAL AI

From a technical implementation standpoint, voice assistant interactions consist of voice input from a person and synthesized speech output from the device. However, the Computers are Social Actors (CASA) paradigm suggests that people's relationships with computers are fundamentally social, which means humans engage with computers in a manner similar to how they engage with each other [155]. Nowadays, voice assistants have been deeply embedded in our daily life, and the interaction design of a voice assistant should be accountable and appropriate for various social situations [156]. Moreover, designers need to make numerous decisions about the assistant's voice and behavior to design a voice assistant that will resonate with users [157]. These decisions are especially crucial when designing an experience where users and a voice assistant engage in social interactions, such as prompting and providing feedback.

One such decision is how to frame or introduce the voice assistant to the user. Designers can frame a voice assistant using metaphors with social characteristics. For instance, framing a conversational assistant by presenting dimensions of warmth and competence can influence

people's willingness to interact and cooperate with agents, even if they signal low competence [158]. Braun et al. also explore in-car voice assistant designs on the dimensions of Dominant – Submissive and Hostile – Friendly and suggests if the voice assistant character matches the user's personality it will encourage trust and engagement [159]. The right framing for the voice assistant is important since social features for non-living actors can deceive users into overestimating the assistant's capabilities, causing frustration when encountering limitations in actual use [160, 161].

Another set of design decisions involves how the assistant will act and react to users through specific conversational and elicitation strategies. Users will explore the limits of their voice assistants by issuing playful commands [162]. In autonomous vehicles, small talk to build rapport can foster trust between the voice assistant and the user [163]. Voice assistants can also shape work practices by highlighting social information elicited from users (e.g., where people deviated from their routine) [164]. Whether the interaction is led by the assistant or the user can also affect user perceptions of the voice assistant [165].

Most current interactions with voice assistants are rather transactional, where the user issues a request and the assistant provides a response [17]. However, more recent work has started to explore opportunities for voice assistants to play a more proactive role, e.g., by trying to predict opportune moments to initiate voice interactions when driving [166], or by empowering users to have meaningful interactions with their own usage data [167].

In our work, we address these different approaches and design choices, extending research by focusing on voice assistant framing (e.g., Assistant, Collaborator, or Learner), options for different elicitation strategies, and the effects that different proactivity levels have on the perception of feedback elicitation via voice.

## CHAPTER 3: CONVERSATIONAL AI FOR ONLINE SURVEY

### 3.1 OVERVIEW

In many domains of behavioral science, including HCI research [22], conducting surveys is a key method to collect data. With the widespread use of the internet, self-administered online surveys have replaced old-fashioned paper-and-pencil surveys and have become one of the most widely used methods to collect information from a target audience [23, 24].

Due to the extensive use of online surveys, survey fatigue is a challenge faced by anyone who wishes to collect data. Research indicates two typical types of survey fatigue [25]. One is *survey response fatigue*. Since people are inundated with survey requests, they are unwilling to take any surveys [168]. The other is *survey-taking fatigue*. Evidence shows that as a survey grows in length, participants spend less time on each question and the completion rate also drops significantly [26, 27]. For example, one of the biggest survey platforms, SurveyMonkey, shows that on average, participants spend 5 minutes to complete a 10-question survey but 10 minutes to finish a 30-question survey<sup>1</sup>.

The problem is exacerbated with open-ended questions because of the extra time and effort required for formulating and typing responses to such questions [26, 28]. Open-ended questions are an important method to collect valuable data and are widely used in self-administered online surveys [22]. In particular, open-ended questions allow respondents to phrase their answers freely when the options of responses cannot be pre-defined or the pre-defined responses may introduce biases [28, 29]. Moreover, open-ended questions help collect deeper insights, such as the background and rationales behind the answers [28, 169]. However, open-ended questions often induce more cognitive burdens and survey-taking fatigue, and respondents are more likely to skip such questions or provide low-quality or even irrelevant answers [26, 28, 31]. Consequently, survey-taking fatigue adversely affects the quality and reliability of the data collected especially when open-ended questions are involved [28, 32, 33]

To combat survey-taking fatigue especially to motivate and guide survey participants to provide quality answers to open-ended questions, several approaches have been proposed. One set of proposals is to inject interactive features into a static online survey, such as providing response feedback [32] and probing responses [33], to improve response quality and encourage participant engagement. However, no existing survey platforms support such interactive features nor do they automatically motivate and guide survey participants to

---

<sup>1</sup>[www.surveymonkey.com/cURIosity/5-best-ways-to-get-survey-data/](http://www.surveymonkey.com/cURIosity/5-best-ways-to-get-survey-data/)

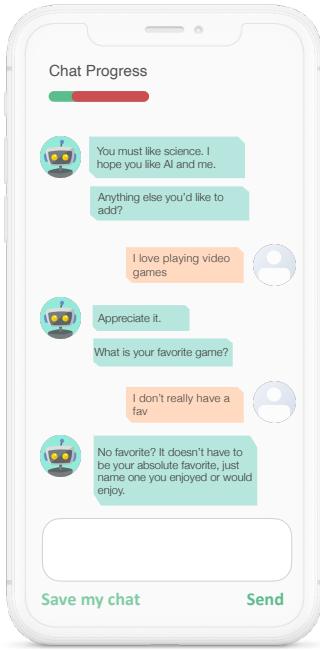


Figure 3.1: A screenshot of an AI-powered chatbot conducting a conversational survey about video game experiences. In a conversational survey, the agent asks open-ended questions, probes answers, and handles social dialogues.

provide quality answers to open-ended questions during a survey.

A lack of support for such interactive features on existing platforms may be due to two main reasons. First, it is difficult to automatically interpret participants’ natural language responses to an open-ended question due to the diversity and complexity of such responses [28]. For example, when asked “*What do you think of the product*”, participants’ responses could be: “*N/A*”, “*I don’t know*”, or “*Although I’ve heard of the product, I’ve never used it so I don’t know what to say.*” Interpreting such highly diverse or complex free-text input requires sophisticated natural language processing algorithms, which is a non-trivial task [14]. Second, even if a system can interpret participants’ free-text responses to open-ended questions, it is difficult to manage potentially complex interactions based on participant responses. Using the above example, a participant may be unwilling to answer the open-ended question and may even provide a gibberish response such as “*afasf asfasf afyiasfaf asf*” [170]. In another example, a participant is willing to answer the question, but provides a very terse answer such as “*not bad*” as opposed to detailed, rich information. Yet in another example, instead of answering the question, a participant asks a clarification question “*Which aspects of the product do you want me to comment on*”. Handling all these situations or their compositions requires that a system not only understands a participant’s input but also automatically handles diverse interaction situations, which is very challenging to implement [14].

On the other hand, the advent of chatbots with their increasingly more powerful conversational capabilities can offer an alternative approach to static online surveys. Specifically, an AI-powered chatbot can conduct a *conversational survey*. As shown in Fig. 3.1, in a conversational survey, a chatbot asks open-ended questions, probes answers, and handles social dialogues.

Intuitively, a chatbot-powered conversational survey retains the advantages of online surveys and offers several additional benefits especially facilitating gathering participant responses to open-ended questions. First, a chatbot can frame survey questions in more personalized, conversational messages, which might help improve participant engagement and response quality [28, 34, 171]. Second, the perceived anthropomorphic characteristics of a chatbot can potentially deliver human-like social interactions that encourage survey participants to reveal personal insights [106]. Third, it is natural for a chatbot to interactively encourage information exchange in the course of a survey, such as providing response feedback and probing responses, which in turn helps reduce survey-taking fatigue and improve response quality. Moreover, it is the inherent functions of chatbots that interpret diverse user natural language input and handle complex conversations. As a result, chatbots can potentially serve as a moderator and proactively manage a survey process, such as dealing with “uncooperative” participants, clarifying the meaning of a question per a participant’s request, and guiding a participant to provide richer and more authentic responses [101, 119].

Despite their benefits, chatbots bear several risks for their use in conducting surveys. First, a turn-by-turn chat requires participants to take extra time and effort to complete a survey. It is unclear whether people would be willing to take the time to chat and complete a survey, let alone providing quality responses. The risk is even higher for surveys with paid participants, who would not be rewarded for taking a longer survey. Second, current chatbots are far from perfect and their limited conversation capabilities may lead to user disappointment and frustration [14]. It is unknown whether the limited capabilities would deter participants from offering quality responses or completing a survey. Moreover, it is difficult for a chatbot to accurately interpret and properly respond to humans’ diverse free-text input to open-ended questions [105]. Once participants realize that a chatbot cannot fully understand or assess their responses, it is unknown whether they would do mischief by intentionally feeding the chatbot with bogus responses, which would adversely affect the overall response quality. Finally, the use of a personified conversational system may lead to user behaviors that affect survey quality. For example, studies show that people have positivity bias when giving opinions to an agent [172], producing potentially biased survey results.

To our knowledge, there have not been any in-depth studies examining the effectiveness and limitations of AI-powered chatbot surveys in contrast to typical online surveys. We,

therefore, ask two research questions:

- **RQ3.1:** How would user response quality differ, especially the quality of user free-text responses to open-ended questions in an AI-powered chatbot-driven survey vs. a traditional online survey?
- **RQ3.2:** How would participant engagement differ in an AI-powered chatbot-driven survey vs. a traditional online survey?

This chapter systematically compared the holistic effect of an AI-powered conversational survey with that of a typical online survey on response quality and participant engagement <sup>2</sup>. As a result, this chapter offers three unique contributions.

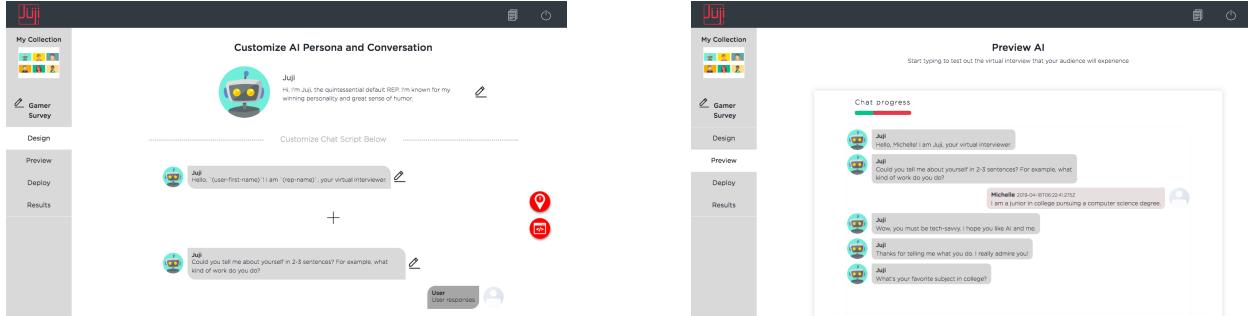
- **An understanding of the holistic effect of AI-powered chatbots on users in surveys with open-ended questions.** Our findings revealed the practical value of AI-powered chatbot surveys especially in eliciting higher-quality responses and increasing respondents' engagement.
- **Design implications of AI-powered chatbots for survey success.** Our work discusses design considerations, such as enabling active listening and early intervention, for creating effective chatbots for conducting conversational surveys, especially facilitating the collection of quality survey responses and improving participant engagement.
- **New opportunities of conducting AI-enabled, human-subject research.** The demonstrated effectiveness of an AI-powered chatbot survey and the simplicity of creating such a chatbot-driven survey open up opportunities of employing AI-powered chatbots to aid in human-subject research, including AI-powered semi-structured interviews and chatbot-driven longitudinal field studies.

## 3.2 METHOD

To answer our two research questions, we designed and conducted a between-subjects field study that compared the outcomes of two survey methods, an AI-powered chatbot survey and a typical form-based survey, on the quality of collected information and participant engagement.

---

<sup>2</sup>This Chapter has been published at ACM Transactions on Computer-Human Interaction (TOCHI) [173]



(a) Juji’s interface for editing questions in a conversational survey.

(b) Juji’s interface for previewing the chatbot built for a conversational survey.

Figure 3.2: Juji’s graphical user interface (GUI) for creating a chatbot survey. The survey designer could directly edit and test a conversational survey.

### 3.2.1 Study Background

To ensure that our findings have ecological validity and practical value, we teamed up with a global leading market research firm that specializes in discovering customer insights for the entertainment industry, including game companies and movie studios. Per the request of the firm, we set up the field study to accomplish two goals. First, the firm wanted to gauge gamers’ opinions of two video game trailers recently released at the Electronic Entertainment Expo (E3) 2018, the premier trade event for the video game industry. Second, they wanted to compare the effect of a chatbot survey with that of a typical online survey which they frequently use to collect such information.

### 3.2.2 Study Platform

To compare the effect of a chatbot survey with that of a typical form-based survey, our study was set up on two platforms.

**Qualtrics** Qualtrics <sup>3</sup> is one of the most popular online survey platforms. Since our collaborator uses Qualtrics frequently for market research surveys, they set up the form-based survey used in this study on Qualtrics. In the Qualtrics survey, an open-ended question is presented with a text box where a participant enters his/her answer. Participants can view and answer only one question at a time and must submit their answers to a presented question before moving on to the next question. A web URL was generated to distribute the survey.

---

<sup>3</sup>[www.qualtrics.com](http://www.qualtrics.com)

Technique	Synopsis	Example
<b>Response Feedback</b>	Acknowledge a user's responses and provide positive feedback to encourage information disclosure	User: 'I am a student majoring in Computer Science.' Chatbot: 'Wow, you must love science. I hope you like AI and me.'
<b>Response Prompting</b>	Ask for more information	Chatbot: 'I'd love to hear more. Go on...'
<b>Response Probing</b>	Probe answers to elicit deeper information	User: 'I love to read books and cook.' Chatbot: 'Why do you like to read?'
<b>Social Acknowledgement</b>	Acknowledge a user's input to convey understanding and verbalize emotions.	User: "'You are so smart' Chatbot: 'Thanks, I'm glad you are happy with me'
<b>Handle Conversation Flow</b>	Handle a user's request for controlling an interview flow and encourage continuation.	User: 'Wait. Give me a second' Chatbot: 'Take your time. We'll continue whenever you are ready'
<b>Handle User Excuses</b>	Handle a user's excuses not to answer a question and encourage continuation.	User: 'I don't know.' Chatbot: 'No worries, just share what's on your mind.'
<b>Handle User Questions</b>	Handle a user's question to encourage continuation.	User: 'Why are you repeating the question?' Chatbot: 'Why not? :) I am pretty sure this is for your own good since I am a responsible A.I.'

Table 3.1: Juji chatbot's key conversational skills. Juji's chatbot support seven key conversation skills that are important in a conversation survey, including response feedback, response prompting, response probing, social acknowledgment, handle conventional flow, handle user excuses, and handle user questions.

**Juji Chatbot** Juji <sup>4</sup> offers an AI-powered chatbot platform where users can create and deploy their own customized chatbots for various tasks. For example, Juji was used to develop chatbots for interviewing job candidates [101] or interviewing college students for teaming purposes [174].

Specifically, a survey creator uses Juji's graphical user interface (GUI) to input a set of

<sup>4</sup>[juji.io](http://juji.io)

survey questions and the order of the questions to be asked. A chatbot is then automatically built with a set of default conversation capabilities as described below, such as handling a conversation around an open-ended question as well as managing several types of user digressions or side-talking dialogues. Fig 3.2 are screenshots of the Juji creator GUI with which a survey creator can design, preview, and deploy a chatbot to conduct a conversational survey. The survey creator can add, delete, and modify a survey question (Fig 3.2a). Juji supports several types of questions, including choice-based questions and open-ended questions. Just like using Qualtrics, the creator can preview a conversation with the created chatbot before deploying it (Fig 3.2b). Similar to a Qualtrics survey, the chatbot is also distributed via a URL.

We chose to use Juji in our study for three reasons. First, the customization and deployment of a Juji chatbot is very similar to creating a survey on Qualtrics. This allows survey creators to easily design, test, and deploy their own conversational surveys, especially if such chatbot-driven surveys are shown to be valuable. Second, the Juji platform is publicly accessible, which allows our study to be easily replicated by others. Third, compared to simple chatbots with very limited conversational capabilities, the Juji chatbot is equipped with a rich set of conversational skills (Table 3.1). Since such conversation capabilities are considered critical for effective human-agent interactions [113], we are interested in investigating how such an AI-powered chatbot would affect survey participants and ultimately the quality of their responses.

Nonetheless, a recent chatbot study finds that it is still quite challenging to create capable chatbots for handling tasks like meeting scheduling [14]. It is unclear how well the Juji chatbot would perform in handling more complex tasks such as surveying users with open-ended questions, where user responses to such questions are much more diverse and complex to anticipate and handle than those in meeting scheduling [28, 105]. To minimize potential adverse effects that the Juji chatbot might have on survey participants and survey results due to its limitations, in this study we decided to employ the Juji's *all* available interaction features to maximize its conversation capabilities (Table 3.1). Accordingly, this study focused on investigating the holistic effect of the chatbot on survey participants and survey results, instead of the contribution of individual features, which could be examined in the future with more controlled studies.

We hypothesize that the Juji chatbot's conversational capabilities would encourage participant engagement and improve the quality of their responses along several aspects. First, survey research shows that interactive feedback improves the quality of survey responses [32, 175, 176, 177]. Since the Juji chatbot can interactively acknowledge user input including user free-text input during a survey, this ability may help elicit quality user responses. Below

is a conversation excerpt between the Juji chatbot and a participant at the beginning of the survey.

Chatbot: First, let me tell you a bit about myself. I'm an Artificial Intelligence (AI), and I am here to chat with you to learn your thoughts on new games. Now could you tell me about yourself in 2 to 3 sentences? I am curious to know what kind of game player you are.

User: I'm an artist and designer. I put a lot of time and effort into my work. I am extremely driven and competitive, I have lots of goals in life as well.

Chatbot: It's a great pleasure to meet someone who is creative. I am very impressed by what you do.

As shown above, the chatbot's response may make the user feel heard and thus would be more willing to continue the conversation. However, in a conversational survey, open-ended questions may elicit arbitrarily complex or diverse user responses. Below are several example user responses to the same self-introduction question asked above by the Juji chatbot.

*"I am an artist, I paint, draw, (and) sculpt. I Love to cook. my favorite food is pizza. I like playing video games. I wouldn't consider myself to be a hardcore gamer or anything. I love casino games."*

*"I am a 29 year old single mother that enjoys playing games with my child. It is something we both enjoy!"*

*"I am a young adult who plays video games in my spare time. I enjoy video games with compelling story lines that get me invested in characters."*

While the Juji chatbot has a built-in rich dialog library that can handle a wide range of user input, it may not always be able to interpret a user input and respond well. Our study thus is also intended to examine how such a capable but imperfect chatbot would interact with survey participants and influence survey results.

On the one hand, studies show that participants may not always be cooperative during a survey, which may result in shallow, irrelevant, or even bogus responses [22, 178]. On the other hand, research shows that proper "interventions", such as interactive probing and nudging, could help improve response quality [33, 35, 36]. Juji chatbot's built-in conversational skills provide natural "interventions". For example, the Juji chatbot could detect certain gibberish and irrelevant user responses to an open-ended question.

Chatbot:	What do you think of the trailer you just watched?
User:	no opinion
Chatbot:	I'm pretty sure you have some valuable input. Please do share.
Chatbot:	What is your favorite game?
User:	I don't really have a favorite.
Chatbot:	It doesn't have to be your absolute favorite, just name one you enjoyed or would enjoy.

Some participants might dodge a question by responding with “*I don’t know.*” or “*I don’t want to answer.*” In such cases, the Juji chatbot also has a set of built-in conversation prompts to handle various user excuses or digressions. Below are two examples.

Instead of answering a question, a participant may even ask a chatbot the same question reciprocally. For example, when a participant is asked “*What do you think of the game trailer you just watched?*”, the participant might respond: “*what do you think*”. Again, the Juji chatbot has built-in conversation capabilities to handle various user questions, including answering a participant’s reciprocal or clarification questions.

To the best of our knowledge, no existing study has examined the use and effect of chatbots with rich conversational capabilities in the context of conducting surveys with dominantly open-ended questions. This is also why we chose the Juji chatbot: it has a rich set of conversation capabilities that can handle diverse and complex user free-text input, as well as manage diverse, nonlinear conversation situations [179]. Moreover, it can be easily customized by feeding it with different survey questions. Equally important, since Juji is a publicly available platform, it should be easy for other researchers and practitioners to replicate the study presented here to further validate and explore the values of such a chatbot in facilitating human-subject research.

### 3.2.3 Survey Questions

Collaborating with the market research firm, we designed a survey that consisted of mostly open-ended questions with a few choice-based questions as described below. The survey contained three major parts.

- *Warm up.* Each survey started out with 3 open-ended questions. A participant was first asked to introduce him/herself in 2 to 3 sentences. S/he was then asked to talk about his/her favorite games, and what new games s/he is most looking forward to playing in the next three months.
- *Game Trailer Assessment.* Each participant was asked to watch two game trailers,

one at a time. After watching a trailer, the participant was asked to describe his/her thoughts and feelings by answering a set of questions:

- What is your immediate reaction to this trailer?
- What do you like about it?
- What do you not like about it?
- How interested are you in purchasing the game you just saw in the trailer? Please rate your level of interest in purchasing the game on a scale of 1-5, 1 being no interest, and 5 being very interested.
- Why did you give this score?
- How has the trailer influenced your interest?
- What would influence your buying decision the most?

In this part, all the questions were open-ended except the rating question. To avoid potential biases, the order of showing the two-game trailers was randomly decided for each participant.

- *Additional Information.* Each participant was also asked to provide additional information, such as what game platforms they use the most, where they look for information about games, and their basic demographics including gender, age, and level of education.

Both the Juji chatbot and Qualtrics surveys used the same set of questions shown above in the same wording and order. To ensure consistency, in both conditions participants can take a survey on a desktop machine, a mobile device, or switch between the two. At the end of the chatbot survey, the chatbot also asked the participants for optional comments about their survey experience <sup>5</sup>.

### 3.2.4 Participants

Our collaborator—the market research firm—hired a panel company to recruit target participants for the study. The panel company is the world’s second largest company that specializes in recruiting and managing survey participants for a number of industries. It maintains a large database of hundreds of millions of survey participants across all demographics around the world. In our study, the market research firm requested the target

---

<sup>5</sup>Participants’ additional comments were not counted when measuring *Response Length*

audience to be US video gamers who are 18 years or older and must play video games at least one hour per week. Based on these criteria, the panel company queried its database and found a large pool of candidates whose profiles matched the two criteria. The pool was randomly divided into two groups where the Qualtrics link was sent to one group and the chatbot link to another group.

### 3.2.5 Measures

To answer our two research questions, we wanted to compare the quality of collected information (RQ3.1) and participants' engagement level (RQ3.2) between the use of the chatbot survey and the Qualtrics survey. The survey results were stored in two CSV files, respectively. Each CSV file contained only question-response pairs. The side talking in a chatbot survey was not in the CSV file. Instead, such information was captured in the chat transcripts. Each completed survey was also stamped with a start and finish time. Most of our analyses shown below were based on the content captured in the two CSV files.

**Assessing Information Quality** Collecting quality information is often the most important goal that a survey is set out to achieve. Although our surveys contained both open-ended and choice-based questions, in this study we focused on assessing and comparing the quality of free-text responses to open-ended questions for three reasons. First, open-ended questions were intended to elicit richer and more in-depth input from the participants, which would enable our collaborator (the marketing research firm) to better understand gamers' thoughts and feelings and inform business decisions (e.g., product development and marketing messages). Second, eliciting quality responses to open-ended questions has been a major challenge in traditional online surveys because participants are often not motivated and unwilling to provide in-depth, quality input. [33]. In this study, we wish to investigate whether the Juji chatbot's conversational skills, such as probing and prompting, could help alleviate such a challenge. Third, it is difficult to determine let alone compare the quality of user responses to choice-based questions, since it is hard to tell whether a participant has made a sensible or just a random choice to such a question during a survey.

To the best of our knowledge, there is no effective tool that can assess the quality of free-text responses to open-ended questions *automatically*. We thus had to manually assess the quality of each free-text user response collected in the surveys. To guide us to assess the quality of user responses systematically, we developed a set of content-based metrics based on Gricean Maxims [180]. The Gricean Maxims, proposed by H.P. Grice in 1975, are a set of communication principles to which both speaker and listener should adhere

Gricean Maxims	Definition	Quality Metrics	Definition
<b>Quantity</b>	One should be as informative as possible	<i>Informativeness</i> <i>Specificity</i>	A participant’s response should be as informative as possible A response should give as much information as needed.
<b>Quality</b>	One should communicate truthfully		A participant’s response should be authentic. Since it is difficult to assess the truthfulness of a user input, we didn’t measure this aspect directly. Our another measure, the level of self-disclosure might signal a level of authenticity.
<b>Relevance</b>	One should provide relevant information	<i>Relevance</i>	A participant’s response should be relevant to a question asked
<b>Manner</b>	One should communicate in a clear and orderly manner	<i>Clarity</i>	A participant’s response should be clear

Table 3.2: Gricean Maxims used to guide the development of information quality metrics. We measure a participant’s response to an open-ended question from four aspects: informativeness, specificity, relevance, and clarity.

to ensure effective communication. Gricean Maxims are often considered “cooperative principles to guide effective communications” [181]. In the context of conducting surveys, a “cooperative” participant would obey all the maxims to produce quality responses. For example, a participant’s relevant answer to an open-ended question complies with the Gricean relevance maxim, while a participant’s clear response to a question satisfies the Gricean clarity maxim. For our purpose, we use the maxims to guide us in defining a set of metrics that quantitatively measure the quality of information communicated by survey participants. As shown in Table 3.2, we measure the quality of information—a participant’s response to an open-ended question from four aspects: *informativeness*, *specificity*, *relevance*, and *clarity*.

Guided by these quality metrics, two researchers independently went through participants’ free-text responses to open-ended questions and manually assessed the quality of each of the response by the three aspects (i.e., *relevance*, *specificity*, and *clarity*). More details about our coding protocol are presented in 3.2.5.

**Informativeness** By the Gricean Maxim of quantity, an effective communication should be informative. To measure the *informativeness* of a text response, we computed the amount of information conveyed in the response by *bits* (shannons) based on information theory [182]. More precisely, the informativeness of a text response is the sum of each of its word’s *surprisal*, the inverse of its expected frequency appearing in modern English (Formula 4.1). In other words, the more frequently a word (e.g., the common word “the”) appears in modern English communications, the less information it conveys.

$$I(\text{Response}) = \sum \log_2 \frac{1}{F(\text{word}_n)} \quad (3.1)$$

To obtain an accurate estimate of a word’s frequency in modern English, we averaged a word’s frequencies in four text corpora, the British National Corpus [183], The Brown Corpus [184], Webtext [185], and the NPS Chat Corpus [186]. For each participant, we computed a total *informativeness* based on his/her free-text responses to all open-ended questions.

**Specificity** Although our informativeness metric mentioned above computes the amount of information conveyed by a user’s text response, it does not assess how specific the response is. Specific responses often provide sufficient details, which not only help information collectors better understand and utilize the responses, but also enable them to acquire more valuable, in-depth insights. For a given open-ended question, text responses could be very diverse, complex or even ambiguous. Since we could not find a reliable natural language processing tool to assess the specificity of diverse text responses to a given question automatically, we manually assessed the *specificity* of each text response on three levels: 0 - generic description only, 1 - specific concepts, 2 - specific concepts with detailed examples.

By our specificity metric, a response would obtain a level-0 specificity if it provides only a shallow or abstract description. For example, when asked “*What is your immediate reaction to this trailer?*”, a typical shallow response with level-0 specificity was

“*I love it, it looks interesting.*”

In contrast, a response with level-1 specificity conveys more specific information, such as the following statement:

“*I am interested in the game and I really like the graphic*”.

The most specific responses with a specificity score of 2 normally offer detailed descriptions. For example, one such response stated

*“Completely blown away! It is unlike any game I have ever seen! There are so many different scenes, so many different characters who look different from each other, a lot of cool weapons and gadgets, so many different ways of fighting, so many different places you can go and it is so action-packed”*

**Relevance** By the Gricean Maxim of relevance, a quality communication should be relevant to the communication context. In a survey context, a quality response should be relevant to the survey question asked. Not only do irrelevant responses provide no value, but they also burden the analysis process. For a given open-ended question, text responses could be very diverse and complex. Similar to assessing the *specificity* of a text response, we manually assessed the *relevance* of each text response on three levels: 0 - Irrelevant, 1 - Somewhat Relevant, and 2 - Relevant.

A response was considered irrelevant if it did not relate to the question asked at all. For example, a gibberish response like “Yhhchxbxb” was considered irrelevant and received a relevance rating 0. Certain responses were considered partially relevant as they did not answer an asked question directly but still provided useful input. Here is an example response that was assigned a relevance score 1, when asked “*What do you like about it (the game trailer)?*”, a participant responded:

*“I don’t like it. I don’t purchase these types of games. But for people who enjoy these types, they might enjoy it.”*

Responses that directly and clearly answered an asked question were considered relevant and assigned a relevance score 2.

For each participant, a total relevance score was also computed by adding up the relevance scores of each response.

**Clarity** By the Gricean Maxim of clarity, an effective communication should be the *clarity* of each text response by how easily the response could be understood by humans without ambiguity, regardless of its topical focus, on three levels: 0 - illegible text, 1 - incomplete sentences, 2-clearly articulated response. Again, given today’s natural language processing capabilities, automatically and reliably scoring the *clarity* of a text response is difficult. We decided to manually score the clarify of each text response.

Gibberish or nonsense responses were marked with 0. We marked responses as partially legible with a score of 1 if they contained incomplete sentences or grammatical errors that impeded a reader’s ability to interpret the responses. For example, when asked “*What is your immediate reaction to the trailer?*”, a response “*very good*” was marked with a clarity

score of 1. Responses obtained a clarity score of 2, if they were articulated with completed sentences with no serious grammatical issues.

**Measuring Level of Engagement** In a typical online survey setting, evidence shows that participants do not tolerate long surveys. As a survey grows in length, the time spent on each question dramatically decreases, and the completion rate also drops significantly. The tolerance for lengthier surveys is even lower for customer-related surveys like the one in our study.

Because a chatbot survey is still a novelty but with flaws (e.g., unable to handle certain user input), we wanted to find out how it would impact participant engagement. In particular, we measured the level of *participant engagement* from three aspects: *engagement duration*, *response length*, and *self-disclosure*.

**Engagement Duration** *Engagement duration* measures how long a participant takes to complete a survey. A longer engagement duration suggests that a participant more willingly stays engaged longer. This was especially true in our case since each participant was rewarded by completing a survey, and not how much time s/he took. For each participant, the engagement duration was automatically logged by each platform.

**Response Length** *Response length* is the word count of each participant’s free-text responses. Similar to engagement duration, the response length also signals participants’ willingness to stay engaged if they are willing to write longer responses.

**Self-Disclosure** Self-disclosure is often used as an indicator for measuring human-agent engagement, as reflected by the breadth and depth of topics exchanged in human-computer conversations [116, 118], based on the social penetration theory [187]. Self-disclosure is particularly important for survey research that aims to elicit personal thoughts and feelings. To measure *self-disclosure*, we manually analyzed each participant’s response to the self-introduction question, and manually counted the number of attributes or topics mentioned (e.g., age, gender, and hobbies). Below lists several example participant responses that were coded with varied level of self-disclosure.

Participants who were most willing to disclose about themselves often offered detailed descriptions about themselves in their responses. For example, the following response mentions the participant’s *age*, *gender*, *marital status*, *favorite game type*, *favorite game*, *game playing history* and even *living condition*.

*"I'm a 29 year old single guy living alone adn i love RPG I really like games like fortnite and call of duty...those have always been my most favorite games and I continue to like them into adult hood...i lose track of time playing them"*

In contrast, the response below said much less about the participant except the video game s/he plays.

*"The only video games I play is mariokart."*

Since Gricean Maxims serve merely as theoretical guidelines for us to measure the quality of user free-text responses to open-ended questions, our current metrics are just one of many ways to *estimate* the quality of user free-text responses. By no means are these metrics unique. Other similar metrics may be defined based on specific situations. For example, if a survey cares more about getting relevant responses than the specificity of the responses, different weights might be associated with each aspect to compute a weighted *response quality index* (RQI). In general, Gricean Maxims can be used as a framework to guide the definition of computational metrics that measure communication quality. The framework allows researchers who are interested in measuring survey response quality to use a systematic method to do so.

**Coding Protocol** Many of the metrics mentioned above require human coding effort. Since we collected over 11,000 free-text responses to nineteen open-ended questions, manually coding each response on all the metrics would require tremendous effort<sup>6</sup>. We thus worked with our collaborator—the market research firm to first identify the analysis requirements based on their business needs. Per the purpose of the study, they selected the top nine most important questions of which answers would help them derive the desired market insights. We manually analyzed a total of 5238 text responses to these nine questions and coded each response on its *relevance*, *clarity*, and *specificity*. We manually assessed 582 participant responses to the self-introduction question and coded each response on its level of *self-disclosure*.

In general, we used a 3-step process to manually code the text responses. First, two human coders independently rated about 10% of randomly sampled responses of each selected question. Second, they reconciled differences and came up with a set of more consistent coding criteria. Third, they used the established coding criteria to code the rest of the responses independently. To avoid potential biases, the coders were blind to the source

---

<sup>6</sup>We did consider the use of Amazon Mechanical Turkers to help code the data. But we could not do so due to the confidentiality of the results.

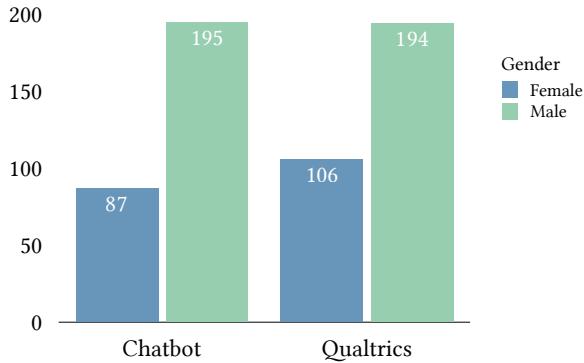


Figure 3.3: Distribution of participants by gender. Among the 282 chatbot survey takers, 87 (30.85%) were female, and 195 (69.15%) were male; while 106 (35.33%) of 300 Qualtrics participants were female, and 194 (64.67%) were male

of responses. After all responses were coded, a Krippendorff’s alpha was used to measure the inter-rater reliability of the coded results [188]. If the Krippendorff’s alpha was above 0.8, indicating a high level of agreement between the coders, the coders then moved on to reconcile the remaining differences if there were any. Otherwise, the coders discussed the differences and re-iterated on the coding process. The Krippendorff’s alpha ranged from 0.80 to 0.99 for each set of coding.

### 3.3 RESULTS

#### 3.3.1 Overview

We received a total of 582 completed surveys: 282 chatbot surveys and 300 Qualtrics surveys. As shown in Fig 3.3, among the 282 chatbot survey takers, 87 (30.85%) were female, and 195 (69.15%) were male; while 106 (35.33%) of 300 Qualtrics participants were female, and 194 (64.67%) were male. Fig 3.4 and Fig 3.5 show the distribution of participants by their age group and level of education. The participants’ ages ranged from 18-50 years old, where the majority (61.86%) of them were between 18-34 years old. Among the 582 people who completed their survey, 50% received at least a college degree. The average self-reported weekly gaming time is 16.90 hours ( $SD = 13.50$  hours).

To measure survey success, two key metrics are often used: *response rate* (Equation 3.2) and *completion rate* (Equation 3.3) [189].

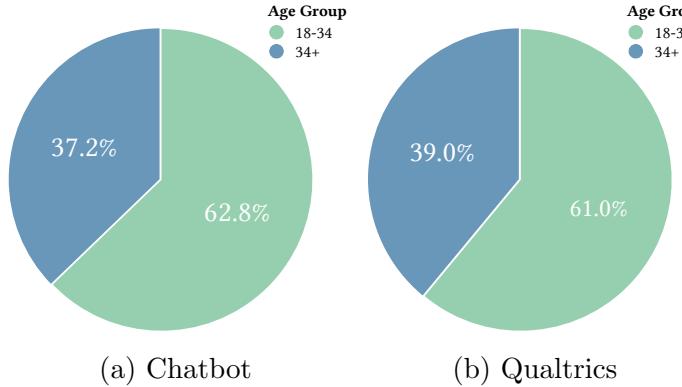


Figure 3.4: Age distribution of participants. The participants' ages ranged from 18-50 years old, where the majority (61.86%) of them were between 18-34 years old.

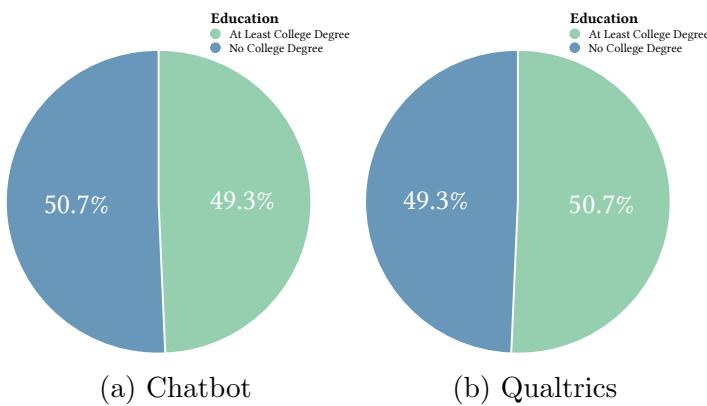


Figure 3.5: Distribution of participants by education level. Among the 582 people who completed their survey, 50% received at least a college degree.

$$\text{response rate} = \frac{\text{number of participants clicked on a survey link}}{\text{number of the participants invited}} \quad (3.2)$$

$$\text{completion rate} = \frac{\text{number of participants completed a survey}}{\text{number of participants clicked on a survey link and qualified}} \quad (3.3)$$

Based on the data provided by the panel company, Fig 3.6 shows the *response rate* and *completion rate* for the chatbot and Qualtrics surveys, respectively. The *response rate* of the chatbot group (84.6%) was lower than that of the Qualtrics group (91.2%). A Two Proportion Z-test shows that the difference is significant:  $z = 4.35$ ,  $p < 0.01^{**}$ . The difference may be due to participant's unfamiliarity to the conversational survey (see Sec 3.6.1 for further discussion). However, the *completion rate* of the chatbot group (54%) was more than twice (2.2 times) the *completion rate* of the Qualtrics group (24.2%). A Two Proportion Z-test

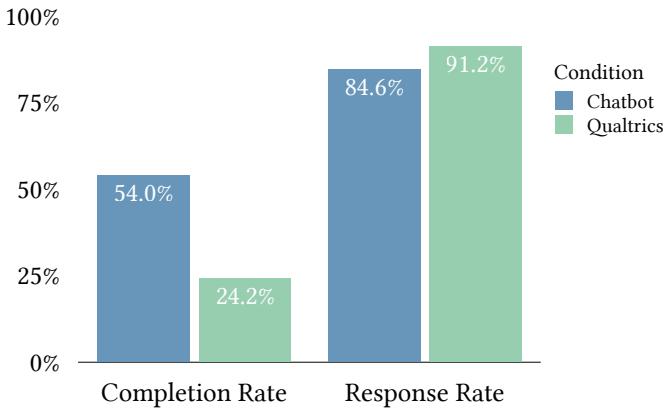


Figure 3.6: The response rate and completion rate of the Chatbot survey and the Qualtrics survey. The response rate of the chatbot group (84.6%) was lower than that of the Qualtrics group (91.2%). A Two Proportion Z-test shows that the difference is significant. The completion rate of the chatbot group (54%) was more than twice the completion rate of the Qualtrics group (24.2%). A Two Proportion Z-test shows that the difference is significant.

shows that the difference is significant:  $z = -12.16$ ,  $p < 0.01^{**}$ . The higher completion rate suggests the better “stickiness” of the chatbot survey.

Below we present the results to answer our two research questions. Since the goal of our study is to compare the outcomes of two survey methods, chatbot versus Qualtrics, on two sets of measures (response quality metrics and participant engagement metrics), we chose to run ANCOVA analyses. ANCOVA is a general linear model blending ANOVA and regression, which helps us examine the true effect of the survey method [190]. In each ANCOVA analysis shown below, the independent variable was the survey method used, and the dependent variable was a computed response quality metric or an engagement metric. Since research suggests that demographics influence people’s behavior with new technology [191, 192], all analyses were controlled for participants’ age, gender, education level and weekly gaming time. Each analysis was also controlled for by participants’ weekly gaming time, as research shows that game playing experience impacts people’s beliefs and attitude toward technology [193, 194]. Furthermore, the analyses involving *response quality*, *response length*, and *self-disclosure* were also controlled for *engagement duration*, since we wished to assess whether the chatbot led to improved information quality, response length, and self-disclosure even when controlled for the amount of time that participants spent on completing a survey.

Before running ANCOVA analyses, we first examined the correlations among all our dependent variables (i.e., response quality and participant engagement metrics) to see how they may be related to each other (Table 3.3). It is interesting to note that most of the

Outcome Measures	1	2	3	4	5	6	7
Response Quality							
1. Informativeness	–						
2. Relevance	0.36**	–					
3. Clarity	0.45**	0.92**	–				
4. Specificity	0.60**	0.75**	0.80**	–			
Participant Engagement							
5. Engagement Duration	0.06	0.04	0.02	0.03	–		
6. Response Length	0.84**	0.31**	0.41**	0.55**	0.10**	–	
7. Self-Disclosure	0.11**	0.18**	0.19**	0.25**	0.03	0.13**	–

<sup>a</sup> N = 582.

<sup>b</sup>\*p < 0.05, \*\* p < 0.01. \*\*\* p < 0.001

Table 3.3: Correlations between dependent measures. Most of the variables were correlated except engagement duration, which did not significantly correlate with any other metrics except response length. No interaction effects were found.

variables were correlated except *engagement duration*, which did not significantly correlate with any other metrics except *response length*. No interaction effects were found. Intuitively, this result seems sensible since most of the metrics were assessing the *content* of participants' responses (e.g., *specificity* and *self-disclosure*). This also implies that *engagement duration alone* would not signal the quality of participant responses.

Table 3.4 summarizes the ANCOVA analysis results for each measure, of which details will be discussed below.

### 3.3.2 RQ3.1: How Would the Quality of Responses Differ?

As mentioned in Section 3.2, we have developed a set of metrics to measure the quality of a user response from four aspects in Table 3.2. Using the responses collected by the two survey methods, we compared their quality by each metric.

**Informativeness** By Formula 4.1, we computed an *informativeness* score of each completed survey based on the participant responses given in that survey. Our results showed that on average the chatbot surveys collected 39% more information than the Qualtrics surveys. With the survey method as its independent variable and controlling for demographics (i.e., gender, age, and education level), weekly game-playing time, and engagement duration, an ANCOVA analysis Table 3.4 showed that the chatbot surveys collected significantly richer information than the Qualtrics surveys, and the survey method was a significant factor contributing to such differences. In addition, among the control variables, the level of education, was the only

Measures	Chatbot		Qualtrics		F	p	$\eta_p^2$
	M	SD	M	SD			
<b>Responses Quality</b>							
Informativeness (bits)	283.33	152.90	203.53	184.19	F(1, 576)=38.55	<0.01**	0.06
Relevance	15.72	4.16	14.05	5.55	F(1, 576)=17.63	<0.01**	0.03
Response Quality Index	27.28	10.20	21.70	10.31	F(1, 576)=48.72	<0.01**	0.08
<b>Participant Engagement</b>							
Engagement Duration (mins)	24.38	13.42	17.90	17.20	F(1, 576)=24.60	<0.01**	0.03
Response Length (words)	90.11	46.23	63.98	54.17	F(1, 576)=57.92	<0.01**	0.09
Self-Disclosure	5.16	2.26	3.57	2.45	F(1, 576)=34.82	<0.01**	0.06

<sup>a</sup> All results were controlled for participant's demographics, including gender, age, education level, and weekly gaming time.

<sup>b</sup> Results for Responses Quality (including Informativeness, Relevance, and Response Quality Index), Response Length and Self-Disclosure were additionally controlled for participant's Engagement Duration.

<sup>c</sup> Results for Engagement Duration have additional control for Response Length.

Table 3.4: Results summary including ANCOVA analysis results on individual measures. In general, the participants who completed a chatbot survey offered much more relevant, specific, and clear responses to open-ended questions than their counterparts did in a Qualtrics survey. And the participants were willingly spending more time with the Juji chatbot, writing longer responses, and disclosing more information about themselves in depth and scope.

factor shown to be significant, although there was no interaction effect between the survey method and the educational level. Specifically, participants with at least a college degree ( $M = 259.93$  bits,  $SD = 166.14$  bits) offered richer responses than those without a college degree ( $M = 224.46$  bits,  $SD = 180.53$  bits):  $F(1, 576) = 6.81$ ,  $p < 0.01^{**}$ ,  $\eta_p^2 = 0.01$ . There was no evidence suggesting the effect of age, gender, engagement duration, or game-playing time.

**Relevance** Next, we examined the *relevance* of collected responses. As mentioned in Section 3.5, we manually assessed the *relevance* of participants' free-text responses to a selected set of nine open-ended questions. For each completed survey, we created a *relevance index* by combining all its responses' relevance scores additively. The results showed that on average the chatbot surveys collected 12% more relevant responses than the Qualtrics surveys did.

With the survey method as the independent variable and controlling for demographics, game-playing time, and engagement duration, an ANCOVA analysis Table 3.4 revealed that the survey method contributed to the differences in *relevance* significantly. In other words, the participants who completed a chatbot survey provided more relevant responses than those who finished a Qualtrics survey. Also results showed that people who played more games per week tended to provide more relevant responses in a survey ( $\beta = 0.04$ ,  $p < 0.05^*$ ). This result suggests that enthusiastic gamers perhaps are more receptive to chatbots and more

willing to offer quality information during their interaction with the chatbots. No interaction effects were found.

To help us better understand the differences in response *relevance*, we further examined the surveys with a *relevance index* value of zero (0), which implied none of their responses was relevant. We found that 27 (9.00%) out of 300 completed Qualtrics surveys contained all gibberish (e.g., "fdlfdbdffdh" or its variants) or bogus statements (e.g., "Funding from a state Itsdhzxoy" given as a self-intro). In contrast, only 7 (2.48%) out of 282 completed chatbot surveys contained completely irrelevant responses<sup>7</sup>. A Two Proportion Z-test showed a significant difference in the proportion of gibberish responses between two conditions ( $z = 3.35$ ,  $p < 0.01^{**}$ ). This also implies that participants were less likely to "cheat" when interacting with a chatbot in a survey. Due to inadequate data collected, it is however unclear which chatbot behavior contributed to the result (e.g., probing, prompting, and social commenting). It might be the combination of the perceived anthropomorphic characteristics of the chatbot and the novelty factor. From the participants' own comments left at the end of a chatbot survey, it is the fact that most of the participants have not experienced a chatbot-driven conversational survey and they were not fully aware of the chatbot's capabilities yet.

**Response Quality Index** For each relevant response, we further examined its quality by two more metrics: *specificity* and *clarity*. To do so, we created an overall *response quality index (RQI)* by aggregating the three quality metrics:

$$RQI = \sum_{n=1}^N relevance[i] \times clarity[i] \times specificity[i] \quad (3.4)$$

(N is the number of responses in a completed survey)

By the above formula, we computed the *overall response quality* of each completed survey. The results showed that on average the completed chatbot surveys produced 25.7% better quality responses than the Qualtrics surveys did. Using the *RQI* as an independent variable and controlling for demographics, game-playing time, and engagement duration, an ANCOVA analysis again showed that the survey method was a significant factor that contributed to the quality differences. Not only did the participants who completed a chatbot survey offer more relevant responses, but their responses were also more specific and clearer than those collected by the Qualtrics surveys.

---

<sup>7</sup>Although the Juji chatbot can detect certain gibberish, its gibberish detection was turned off for certain questions. For example, it was turned off for questions asking about one's favorite game or game platform. This is because many legit game names or game platforms might be considered gibberish as they don't exist in the generic natural language corpora used for Juji's gibberish detection algorithm.

In particular, the participants in the chatbot survey often offered more specific details per the chatbot's question. Here is an example response to the question "*what's your immediate reaction to the trailer*":

*"amazing and very fluid. I like the pace of the game, the weapons, the soldiers gears as well as the setting or the terrains where the action take place."*

Even a short response in the chatbot survey contained specifics. Below is such a short response to the same question above:

*"I like the scenery. It shows old, broken down building."*

In contrast, the responses collected by the Qualtrics surveys were more terse and abstract. Here are some example responses to the same question above

*"It looks interesting."*

*"very good"*

By the response quality index (*RQI*), participants with at least a college degree ( $M = 23.46$ ,  $SD = 11.79$ ) also provided higher quality responses than those without ( $M = 19.89$ ,  $SD = 12.90$ );  $F(1, 576) = 14.31$ ,  $p < 0.01^{**}$ ,  $\eta_p^2 = 0.02$ . Intuitively, this result seems sensible since the level of education would influence one's knowledge and communication.

### 3.3.3 RQ3.2: How Would a Chatbot Impact Participant Engagement?

To compare participant engagement exhibited in the two survey methods, we examined three measures: *engagement duration*, *response length*, and the level of *self-disclosure*.

**Engagement Duration** For a completed survey, *engagement duration* recorded how much time a participant took to finish the survey. Our result showed that participants took seven more minutes on average to complete a chatbot survey than finishing a Qualtrics survey. With the survey method as an independent variable and controlling for demographics, game-playing time and response length, an ANCOVA analysis showed that the duration differences were significant, and the survey method was the only significant factor contributing to such differences.

Coupling with the survey completion rate (chatbot surveys 54% vs. Qualtrics 24%), the result implied that the participants were willingly engaged with the Juji chatbot longer. This is especially true considering that the participants were paid just for completing a survey not for the amount of time spent.

**Response Length** A longer engagement duration does not necessarily mean that a participant is willing to contribute more content. We thus measured *response length* to estimate a participant’s level of engagement by his/her willingness to make content contributions. We counted the number of words that participants contributed to each completed survey. On average, the participants contributed 30 more words in a chatbot survey than in a Qualtrics survey. Using *response length* as an independent variable controlling for demographics, gaming time, and engagement duration, an ANCOVA analysis showed that such differences were significant and the survey method was the only significant factor contributing to the differences (Table 3.4).

This result implies that the participants were also willing to contribute more content in a chatbot survey regardless of their demographics, game-playing time, and the time spent with a chatbot.

**Self-disclosure** Research shows that one’s willingness to disclose him/herself in a human-to-human or human-to-machine interaction indicates one’s level of engagement in the interaction [118, 195, 196]. In our study, since each participant was asked to introduce him/herself (“*Could you describe yourself in 2-3 sentences*”), we examined how much information the participant disclosed about him/herself. Our coded results showed that on average the participants revealed 1.6 more types of information about themselves (e.g., age, occupation, pet, and game preferences) in a chatbot survey than in a Qualtrics survey. An ANCOVA analysis, controlled for demographics, game-playing time, and engagement duration, showed that such differences were significant and the survey method was the only significant factor contributing to such differences.

We further examined the types of information disclosed by the participants. We found that 32.62% of the participants disclosed personal information (e.g., age, gender, and marital status) to the chatbot, while only 15.67% of the participants offered similar information in the Qualtrics survey. Specifically, the participants disclosed various types of detailed information about themselves in a chat, such as personal facts, daily activities, and personality. Below are three example responses from the chatbot surveys.

*“Well I am a student working towards my masters at the same time working part-time at Starbucks on campus. I study/work in the morning and game at night Spending 8 hours studying/working , 5 napping, and the rest gaming. Snacking through out the whole time but I try to get at least one hour every day of my game time exercising.”*

*“I am 44 years old, married to my soulmate with a dog named Killer. I like to play guitar and make my own music. I enjoy computers and anything techie. I work part time and make money online doing different things.”*

*“I am a chill laid back person. I like sports and music.”*

In contrast, much fewer participants did so in their Qualtrics surveys. Their responses were mostly about the type of games they like. Below is an example.

*“I like to play games like creating kingdoms and restaurants.”*

Asking a participant to make a self-introduction was the very first question in the chatbot survey after the chatbot said hello and introduced itself<sup>8</sup>. We thus conjecture that participants' willingness to self disclose could be attributed to a tendency of reciprocity, which was found to deepen users' self-disclosure in previous studies of human-agent interaction [196, 197].

**Participants' Feedback** Quantitatively, our analyses show the significant differences in both response quality and participant engagement resulted from two types of surveys and in most cases the survey method is the only significant factor contributing to such differences. To better understand how the participants felt about their experience with the chatbot, we also examined their voluntary comments left at the end of each chatbot survey.

Among the 282 participants who finished a chatbot survey, 70% (193 out of 277<sup>9</sup>) left optional comments at the end of their survey. 95% (183 out of 193) were positive, 2% (3 out 193) were negative, and the remainder were neutral<sup>10</sup>. The positive comments roughly fell into four categories. Among the positive comments, 30% (54 out of 183 positive comments) indicated personal connections with the Juji chatbot, and 43% (79 out of 183) positive comments expressed how much the participants enjoyed the chat and found the survey entertaining. In addition, 22% (40 out of 183) praised that chatting with Juji was the best survey format that they had experienced, while the remaining 5% commented how smart the chatbot was. In short, based on their comments 67% of the participants enjoyed their engagement with the chatbot for one of the main reasons identified above. Below shows a list of example comments.

One third of participants seemed to have made a personal connection with the chatbot.

*“the whole time i was doing this survey it felt like i was talking to a friend and sharing the same common ground. i loved that i wish it didnt have to end”*

43% of participants simply enjoyed their chat with Juji and had a good time.

*“very dynamic and very fluid conversation you have great quality thanks”*

---

<sup>8</sup>The Juji chatbot has a default conversation opening that can be customized by a creator

<sup>9</sup>5 participants left gibberish comments and were discounted

<sup>10</sup>The most of the neutral comments simply said “thank you”

About a quarter of participants considered the chatbot survey was fun and the best survey experience they ever had.

*“This was the most fun I’ve had ever taking a survey, I absolutely loved it. Please do more in the future, it was a fresh experience!”*

*“This was one of the most entertaining surveys ive done. Great Job at making it not boring like others.”*

### 3.4 SUMMARY OF FINDINGS

As captured in part in Table 3.4, our study results revealed three key findings as summarized below.

- **The chatbot elicited significantly higher quality responses.** The participants who completed a chatbot survey offered much more relevant, specific, and clear responses to *open-ended* questions than their counterparts did in a Qualtrics survey.
- **The chatbot encouraged significantly more participant engagement.** The participants were willingly spending more time with the Juji chatbot, writing longer responses, and disclosing more information about themselves in depth and scope.
- **The chatbot pleased a majority of participants.** The participants’ comments show that they enjoyed chatting with Juji and preferred taking this kind of conversational survey in the future. Even considering the novelty factor and the potential positivity tendency toward a humanized machine [172], these comments were still encouraging especially considering that the comments were from seasoned, paid survey takers and there was no additional reward for them to spend more time on a survey or leave optional comments.

### 3.5 DISCUSSION

The main purpose of surveys is to elicit quality information from a target audience to inform decisions. One of the most effective ways to elicit quality information is through an engaging conversation [176]. However, having a conversation, especially one with a human interviewer, may be time-consuming or induce potential biases. In addition to the findings presented in Sec. 3.3, here we discuss additional benefits and risks of chatbot-driven conversational surveys.

### 3.5.1 Quality Responses without Positivity Bias

From Sec. 3.3, our study results clearly indicate that conversational surveys draw out more relevant and richer user responses as well as foster better user engagement. This is one of the obvious benefits of using a chatbot for conducting surveys. However, existing research shows that in a human-computer conversation, users might give more positive responses because of their affections for humanized machines [172]. Any biased survey results, however, would prevent researchers from discovering the truth and hinder decision making. We thus examined whether our chatbot surveys caused any potential positivity biases in key responses that would impact business decisions.

In our study, our collaborator cared very much about the “end results”—participants’ answers to a Likert scale question for each game trailer on a scale of 1-5 *“How interested are you in purchasing the game you just saw in this trailer”*. We compared the ratings in the chatbot surveys (Trailer 1:  $M = 3.76$ ,  $SD = 1.62$ ; Trailer 2:  $M = 3.75$ ,  $SD = 1.63$ ) and those in the Qualtrics surveys (Trailer 1:  $M = 3.73$ ,  $SD = 1.47$ ; Trailer 2:  $M = 3.75$ ,  $SD = 1.49$ ). Controlling for demographics, game-playing time, and engagement duration, an ANCOVA analysis showed that the rating differences were not significant: Trailer 1:  $F(576) = 3.62$ ,  $p = 0.06$ ,  $\eta_p^2 = 0.01$ ; Trailer 2:  $F(576) = 2.18$ ,  $p = 0.14$ ,  $\eta_p^2 < 0.01$ . In other words, the use of a chatbot did not influence the participants to provide more positive ratings.

Although the participants’ ratings for the trailers did not seem to be influenced by their affection for the chatbot, it is unclear whether their overwhelmingly positive comments (Sec. 3.3) about their perception of or attitude toward the chatbot was biased.

Nonetheless, our study reveals a key benefit of a chatbot survey: eliciting richer and deeper participant responses while not causing unwanted positivity biases.

### 3.5.2 Coping with Survey Fatigue

On average the participants spent over 20 minutes with the Juji chatbot, which is considered extraordinarily long in an online survey context<sup>11</sup>. Since our results show that the participants were willingly engaged with the chatbot longer and still very positive about their experience, another benefit of a chatbot survey seems to be in combating survey fatigue. To verify this benefit, we examined the participants’ response quality overtime, since survey taking fatigue would negatively affect response quality. We did not observe any quality degradation over time with a survey that lasted for about 20 minutes. On the opposite, we found a question *“Why do you give this score?”* appeared very late in the survey even elicited better (i.e., more

---

<sup>11</sup>[www.surveymonkey.com/curiosity/5-best-ways-to-get-survey-data/](http://www.surveymonkey.com/curiosity/5-best-ways-to-get-survey-data/)

relevant, specific, and clear) responses than most of the *what* questions asked in the middle of the survey. This might be attributed to the question prompts used. Previous work shows that *why* questions encourage people to think deeper and offer quality responses [198].

Moreover, the interactive nature of the Juji chatbot appeared to help overcome survey-taking fatigue. Based on their comments, 42.2% (119 out of 282) of participants explicitly mentioned that they really enjoyed their chat with Juji and thought the experience “cool”, “entertaining”, and “amazing”. Despite the potential novelty effect as discussed below, the participants explicitly mentioned that they liked their interaction with the Juji chatbot, which made the survey not boring like typical online surveys. Because of the interaction, certain participants even felt that they were talking to Juji the chatbot as if they were chatting with a “friend”, “nice guy”, or a “brother”. It seems that such bonding encouraged the participants to stay engaged and alleviated the survey-taking fatigue often experienced in a traditional, static survey.

### 3.5.3 Texting with a Chatbot on Mobile Devices

With the widespread use of mobile devices, more and more people take surveys on their mobile phones. Statistics provided by Survey Monkey showed that about 15% of people nowadays take surveys on their mobile devices<sup>12</sup>. However, studies showed when using mobile devices, people’s attention is limited and people are not willing to type long sentences which is critical to elicit high-quality responses especially to open-ended questions [199, 200, 201]. On the other hand, a Gallup poll<sup>13</sup> shows that Americans under 50 use texting on their mobile phones as the dominant communication method. Thus, another benefit of using a chatbot survey seems to provide survey participants with a natural communication form that they are most familiar with.

On the other hand, previous studies suggested the use of mobile devices may prevent people from entering quality answers to open-ended questions [199, 200, 201]. We thus investigated the relationships between the use of mobile devices and participants’ behavior in our study. Since our collaborator did not collect participant’s device information in the Qualtrics surveys, we had only the device information for the chatbot survey participants. Out of 282 chatbot survey takers, 95 (33.7%) of them used mobile devices, including cell phones and tablets. We analyzed the relationships between participant’s device use (mobile vs. non-mobile) and various response quality metrics and participant engagement metrics (Section 3.2.5). For each metric, we constructed an ANCOVA analysis controlling for demographics, game-playing time,

---

<sup>12</sup>[www.surveymonkey.com/curiosity/are-people-completing-surveys-on-mobile-devices/](http://www.surveymonkey.com/curiosity/are-people-completing-surveys-on-mobile-devices/)

<sup>13</sup>[news.gallup.com/poll/179288/new-era-communication-americans.aspx](http://news.gallup.com/poll/179288/new-era-communication-americans.aspx)

and engagement duration. We found no significant associations between the usage of mobile devices and the participant’s response quality nor engagement. This is very encouraging, since our study results show no evidence that the use of mobile devices would adversely influence participants’ behavior in a chatbot survey like ours.

### 3.5.4 Understanding User Characteristics

One of the purposes of conducting surveys is to understand the characteristics of target participants [22]. Understanding participant characteristics has multiple benefits. For example, they can be used to help explain survey results. In addition, they can be used to effectively guide a conversation (e.g., persuading a user based on his/her personal characteristics [202, 203]). Unlike an ordinary chatbot, another unique feature of the Juji chatbot is its ability to analyze a user’s text input on the fly and infer the user’s characteristics [179]. In the current study, we explored this feature of the Juji chatbot preliminarily. Near the end of each chatbot survey, the chatbot analyzed a participant’s text responses given in the survey and automatically inferred his/her key gamer characteristics.

For example, one participant was told that she had the characteristics of a “*social gamer*”, who is extroverted, friendly and enjoys playing games with friends. In contrast, another participant was informed that he possessed the characteristics of a “*complete gamer*”, who is very driven to achieve all game milestones. To verify the accuracy of its inference, the chatbot also asked each participant to rate the accuracy of their inferred gamer characteristics on a scale of 1-5, 1 being completely off and 5 being very accurate. The results showed that  $M = 4.55$ ,  $SD = 0.65$ , which was very encouraging.

To better explain the participants’ ratings about their interest in purchasing a game, our collaborator wanted to know how participants’ inferred gamer characteristics were related to their “interest to purchase” rating of the two respective game trailers. To do so, we performed a regression analysis on the inferred participants’ gamer characteristics and their game purchase interest. Our results showed that “*social gamers*” preferred game 1—a shooting strategic game with many action elements, while “*passionate gamers*” favored the second game—an action-oriented strategic game with a compelling storyline.

Although in this study our investigation of the relationships between participants’ inferred characteristics and their purchase interest is very preliminary, our study suggests another benefit of chatbot-driven surveys. Specifically, a chatbot can potentially achieve a “two birds with one stone” outcome: eliciting information from survey participants and using the elicited information to infer participants’ characteristics at the same time. This would reduce survey-taking time since separate surveys (e.g., gamer type survey) intended to understand

participants' characteristics may no longer be needed. Moreover, the inferred participant characteristics would help gain deeper insights into the collected information.

### 3.5.5 Chatbot Addiction

Numerous studies show that people may become addicted to powerful and omnipresent technologies. For example, research shows that the omnipresence of mobile phones has caused mobile phone use addiction [204]. Users' own characteristics, such as personality traits, could even predict addiction behavior [205]. Likewise, as chatbots' capabilities become more advanced and their uses become more ubiquitous, the bonding between humans and machines may grow stronger and potentially lead to certain unwanted effects, such as chatbot addiction. Compared to other technology addictions, chatbot addictions would be more likely to occur and harder to overcome due to the anthropomorphic and personal nature of human-chatbot interactions. It is thus important for chatbot designers and developers to be aware of potential usage behaviors including addiction behavior, and consider proper behavior detection and prevention mechanisms as part of the chatbot design. For example, research shows that certain user characteristics, such as extroversion and self-esteem, could predict problem uses of technology including technology addiction [205]. Chatbot designers and developers may borrow such findings to detect potential addiction behavior and deter users from overengaging with a chatbot.

### 3.5.6 User Privacy and Control

Our study results reported in the last section along with other studies suggest that users are willing to disclose sensitive, personal information to a chatbot (e.g., [119]). In addition, a recent study found that a chatbot could gain a user's trust if it informs users that their data will be securely stored [117]. While gaining user trust helps a chatbot elicit authentic information in a survey context, these findings reveal potential risks of the malicious uses of chatbots. Such chatbots can manipulate users to gain their trust and steal their sensitive, personal information. Moreover, even benevolent chatbots might elicit unnecessary sensitive information since users tend to disclose more to a chatbot than to a human [206]. Those unnecessary self-disclosures may expose users to privacy leakage or identity theft risks if data breaches occur.

Therefore, proper chatbot design and evaluation guidelines should be in place to allow a chatbot to accomplish its task while protecting user privacy. For example, in our study certain participants disclosed where they live and what kind of job they do when they were asked to

introduce themselves. In such a case, the chatbot could be designed to warn the participants not to disclose personally identifiable information. In addition, the chatbot could show a participant what information has been gathered and provide the participant the options to control the use of the information (e.g., obfuscating parts of the information as needed [207]). Not only do these privacy controls help guard user privacy, but they also improve user engagement and satisfaction [208]. However, such user controls may interfere with the collection of authentic information (e.g., authentic patient information for determining proper treatment) as certain parts of data might be obfuscated or removed. Therefore, future studies are needed to investigate how to reach a balanced design of chatbot-driven surveys that can protect user privacy while guard information validity.

### 3.6 LIMITATIONS

Our current study has several limitations, including flaws in the study operations and scope of the results.

#### 3.6.1 Study Controls

Just like any field studies, our study was limited by practical constraints imposed by real-world operations. One constraint was that we had to use a panel company to find the qualified participants (over 18 years old, hard-core gamers) for the purpose of the study. We had no control over how the two target audience groups were selected, how recruiting messages were sent out, or how the reward was determined. For example, the initial recruiting messages sent to the two target groups were the same except one containing a link to Qualtrics and the other to the chatbot. During the first couple of days, such a message however caused a 30% abandon rate in the chatbot survey. Our log data indicated that a large number of participants abandoned the survey as soon as the chat screen appeared. Our guess was that they thought this was a regular survey per the recruiting message but the chat screen did not look like a typical online survey they were used to, which made them abandon the survey immediately. We thus asked the panel company to revise the recruiting message, which explicitly informed the participants that they would chat with an AI-based chatbot in this survey. The later revised message may adversely impact the response rate. The participants may choose not to take the survey because of their familiarity with the novel chatbot-driven survey. Although the revised message reduced the abandon rate dramatically, the overall completion rate given by the panel company (Equation 3.3) was affected by this incident.

### 3.6.2 Study Audience and Scope

Since our study aimed at understanding gamers’ thoughts and feelings about newly released game trailers, this gamer-focused study might limit the applications of our results to other populations. Although research shows that the newer generation grows up playing video games, gamers often form their own beliefs and perhaps are more open and receptive to new technologies such as chatbots [194]. As mentioned in Section 4, our analyses showed that game-playing time indeed contributed to the difference in response relevance. It is thus unclear whether our results would hold for non-gamer populations. Additionally, our study is on participants’ opinions of game trailers of popular games, it is unclear whether our results would hold for other types of surveys, for example, employee engagement surveys or market research surveys for more “mundane” products (e.g., household products) or services (e.g., banking service). Although none of the analyses showed that participants’ *age* contributed to the differences in two survey methods, two-thirds (2/3) of the participants were between the age of 18-34 (Fig 3.4). Therefore, it is also unclear whether our results would hold for populations in other age groups.

Additionally, our study focused on investigating the use of chatbots for collecting user free-text responses to open-ended questions. As mentioned in the Introduction, open-ended questions are an important way to elicit important user insights and are widely used in web-surveys [31]. However, eliciting quality responses to open-ended questions is very challenging in a typical online survey since participants are often not motivated to provide detailed, rich input. Moreover, no existing survey platforms provide tools to facilitate the collection of user responses to open-ended questions. Therefore, our study focused on examining the use of an AI-powered chatbot and its effect on surveys with open-ended questions, hoping to find new ways to aid in such surveys. However, it is unknown how the use of an AI-powered chatbot would aid in surveys with choice-based questions. Although one recent study shows that a chatbot survey exhibited less satisficing behavior on choice-based questions [177], it is unclear whether such a chatbot handles any user digressions (e.g., a user does not give an answer by selecting a choice) as in our study. Further studies are definitely needed to examine the effectiveness of chatbots in eliciting information through all types of questions.

### 3.6.3 Novelty Effect

In recent years, chatbots or intelligent agents have been widely adopted in our daily lives [209, 210]. In a business context, chatbots have also been used in a wide range of applications, from job interviewing [101] to serving as a workplace companion [102]. Nonetheless, it is

still uncommon to use chatbots for conducting lengthy, conversational surveys that mainly consisted of open-ended questions as we did in our study. Furthermore, few chatbots used in a survey context [106, 177] have any conversation skills capable of handling diverse and complex user interactions as the Juji chatbot did.

As the first study of its kind, our study setting was a novelty to the majority of the participants. In particular, two novelty factors presented in our study might have affected participant behavior and biased our study results: (1) the form of the conversational survey itself with mainly open-ended questions; and (2) the rich conversation skills of the Juji chatbot. Since we could not control for the novelty effect in our current study design, we do not know the contributions of the novelty factors. This is certainly one of the limitations of the study.

While we are planning longitudinal studies to examine the influence of the novelty effect, here we briefly discuss the potential effect of the two novelty factors mentioned above. Just as any novel technology, the novelty effect may wear off as chatbots become a norm. In our case, as machine-driven conversational surveys become more common, the effect caused by the first novelty factor is most likely to wear off, similar to the fact that online surveys are now a norm in lieu of pencil and paper surveys. However, as chatbots' conversation capabilities become more powerful, the second novelty factor would continue influencing user behavior beyond their novelty. This is because chatbots' inherent interactive features, such as acknowledging, prompting, and probing, directly influence human-machine interactions [113].

In our view, the influence of chatbot technologies on user behavior would be similar to that of mobile technologies. Initially, user adoption of mobile devices might be attributed to the novelty of the technology. However, as the mobile technologies advance and can offer users a wide range of functions, their influence on user behavior is beyond their novelty. Analogously, as AI agents advance their capabilities, e.g., their better understanding of user responses and initiating more adaptive questioning (see “Going beyond surveys”), chatbots will greatly influence user behavior beyond their novelty. Just like mobile technologies, they may even induce unwanted effects (e.g., chatbot addiction, as discussed above).

### 3.6.4 Effect of Individual Chatbot Skills

Our study was specifically designed to investigate the use of a chatbot *as a whole* for conducting surveys and compare its outcomes (i.e., resulted survey quality and user engagement) with that of a traditional online survey. From analyzing the chat transcripts, we observed that the Juji chatbot used a rich set of conversation skills to interact with a participant

in a conversational survey, such as information prompting, answer probing, and handling side-talking (Table 3.1).

Below is a chat transcript showing two of the features, prompting and probing.

Chatbot:	What is your favorite type of games?
User:	sports and adventure strategies
Chatbot:	Got it, I'm interested in hearing a bit more. Please? <Prompting>
User:	I love video games since I was a very fanatical child and I am currently following the latest trends
Chatbot:	Ok. Thanks. Why do you like it? <Probing>
User:	it's my favorite pastime and latest technology

In fact, from the 282 completed chat transcripts, on average, each participant was prompted by the chatbot 2.9 times for more information. The participants complied with the prompting request and offered more information 71% of the time. In addition to information prompting, the chatbot asked follow-up probing questions as shown above. By the transcripts, the chatbot asked at least one follow-up question to 173 out of 282 (61%) participants who completed a conversation survey. The participants answered a follow-up question 89% of the time.

Moreover, the chatbot engaged the participants in side-conversations in 32 out of 282 (11%) conversations, such as answering the participants' questions or dealing with their unwillingness to provide answers. Below is a chat excerpt, which shows how the chatbot handled the side conversations and was persistent in getting an answer from the participant.

Chatbot:	What do you not like about it?
User:	Why are you repeating the question
Chatbot:	Why not? :) I am pretty sure this is for your own good since I am a responsible A.I. <Side Talking> Well, let's just get going. What do you not like about it?
User:	Woah you understood that smart ai

The participants' behavior observed in our study suggests that conducting a conversational survey is more than merely asking survey questions one by one especially when open-ended survey questions are involved. As shown in the chat transcripts, the chatbot's various conversation skills, such as the abilities to prompt for more information and probe answers

whenever appropriate, contributed to the survey success. In addition, the chatbot's abilities to handle diverse conversation situations, such as user excuses, questions, and bogus responses, also contributed to the survey success. Intuitively, each of the chatbot conversation skill might have contributed to the survey success differently.

However, the data collected in this study is inadequate to tease apart and quantify the contribution of prompting or any other chatbot interaction feature due to the highly adaptive chatbot behavior, which was intended to offer the best possible user experience in this study. For example, in the study, the triggering of prompting depended on the nature of user input and the context (e.g., the type of question asked). For example, a participant who gave an irrelevant response such as gibberish input was prompted to provide a relevant answer, while a participant who gave a relevant but terse response was prompted to give more detailed, specific input. The question context also determined the triggering of chatbot's interaction features. For example, in the question "*what's your reaction to the trailer you just watched*", the chatbot would prompt a participant for more information if the participant's response is very short. In contrast, when asked "*what's your favorite game*", a participant's short response might be sufficient and would not trigger the chatbot prompting.

In a conversation, different open-ended questions require different relevant answers. Additionally, to balance the amount of information elicited and the time required to complete a survey, the chatbot could not afford to prompt a participant for more information in every open-ended question. Moreover, the multiple interaction features of the chatbot often worked in combinations to nudge users to provide quality input. As shown in the chat transcript below, one user's input triggered a prompting, a probing, and another prompting. Since a chatbot's interaction features have both benefits and risks, rigorously quantifying the contribution of each chatbot conversation skill would be valuable. However, this would require additional, fully controlled experiments beyond the scope of the current study.

Chatbot:	What is your favorite type of games (e.g., first-person shooter, multi-player, open world, strategy, sports/racing, battle royale, etc.)? Please give an example of such games.
User:	sports are my favorite
Chatbot:	Oh, go on... <Prompting>
User:	i like fighting and racing too
Chatbot:	Thanks. Why do you like it? <Probing>
User:	its fun to play
Chatbot:	Well, anything else? <Prompting>
User:	mario kart is one of my favorites

### 3.7 DESIGN IMPLICATIONS

Here we present several design considerations for creating effective chatbots for conducting conversational surveys.

#### 3.7.1 Active Listening

Quantitatively, our results showed that the Juji chatbot elicited significantly higher quality responses and significantly more participant engagement. Qualitatively, over half of the participants (67.4%) expressed their highly positive experience with the chatbot. Our qualitative analysis of the chat transcripts also revealed that the chatbot’s rich conversation skills play a critical role in the success of a conversational survey. These skills enable the chatbot to *actively listen* to its users and make the users feel heard, which not only delighted the users, but also nudged the users to contribute more quality answers. Our findings are consistent with other study findings that active listening improves communication effectiveness in text-based communication [175] as well as in information elicitation [176]. In other words, adopting a chatbot with active listening skills helps achieve survey effectiveness.

#### 3.7.2 Intervening Early

In our initial analysis, we noticed that the quality of a participant’s response to the first survey question (self-intro) seems a barometer of the quality of his/her overall responses. For example, the participants who gave a gibberish response to the first question also gave gibberish responses to all other questions. We thus analyzed the correlation between the quality of the first responses to that of overall responses. We found a *significant* correlation between the *relevance* scores of participants’ first responses and that of their overall responses:  $r = 0.78$ ,  $N = 582$ ,  $p < 0.01^{**}$ . Similarly, a significant correlation exists between the response quality index of the first responses and that of overall responses:  $r = 0.70$ ,  $N = 582$ ,  $p < 0.01^{**}$ .

These results imply that if a participant diligently answers the first question, it is highly likely that s/he would do so for the entire survey. In this light, the first open-ended question could be used as an effective screener to assess a participant’s willingness to do a survey. Since prior work shows that machine ”intervention” could be effective (e.g., preventing fake answers in an interview [211]), a chatbot could even intervene if it detects a participant’s unwillingness to answer the first question. Such early intervention benefits information collectors as well as participants since the participants can be reminded of their responsibilities and expectations.

### 3.7.3 Mixing Chatbot with Qualtrics

Our collaborator, the market research firm, was satisfied with the study results and wishes to use a Juji-like chatbot in their future studies especially for eliciting qualitative responses. They, however, want to explore a hybrid use of a chatbot with a typical online survey where they can use a platform like Qualtrics to ask complex quantitative questions (e.g., Matrix rating) while leveraging a Juji-like chatbot for open-ended questions. We see two ways of integrating the two methods: embedding the chatbot into a Qualtrics survey or vice versa. On the one hand, it might be more effective to start with a chatbot and then move to a Qualtrics survey because of the chatbot's interactive features including its potential ability to intervene. On the other hand, it may be better to start with quantitative questions on Qualtrics and then transition to the chatbot, which would help battle survey taking fatigue. It would be interesting to study different hybrid models and their effect on survey results and participant experience.

### 3.7.4 Creating Empathetic and Responsible AI Agents Beyond Surveys

From this study, we observed a chatbot's abilities to conduct surveys or structured interviews based on a set of pre-defined questions and the order of the questions. The conversation capabilities of the chatbot demonstrated in our study suggest the natural next step to advance its capabilities for wider applications. In particular, we see the opportunity to use a chatbot for replacing structured phone interviews. It is also possible to develop a chatbot for semi-structured interviewing. Starting with an interview guide, a chatbot will ask open-ended questions, interpret user answers, and *automatically* come up with follow-up questions to drill down on interesting ideas that emerged in the conversation and uncover hidden insights. To achieve this goal, several key advances still need to be made, including automatically formulating critical questions based on a participant's response, as suggested by the Socratic questioning method [198].

Moreover, our results including participants' comments also indicate the possibility of creating a new generation of AI agents, which can deeply understand users (e.g., one user is open-minded and easy going while another is analytical and cautious) and proactively guide users based on the conversation and the unique characteristics of the users (e.g., customizing survey questions for analytical participants). Not only can such AI agents exhibit empathy during their interactions with users and gain user trust, but they can also be made responsible for their actions (e.g., guiding users to provide authentic information while guarding user privacy).

In our study, although the average engagement duration was only 20 minutes, certain participants already felt a personal connection with the Juji chatbot. We envision that empathetic and responsible agents can better bond with users and accomplish a wide range of tasks beyond conducting surveys. For example, they can serve as personal well-being coaches, career counselors, and personal caretakers. Such agents will also push the boundary of relational agents [118] and help achieve the goal of creating true human-machine symbiosis [212]. It would be exciting to investigate the uses of and effect of such AI agents on people’s daily lives.

### 3.8 CONCLUSIONS

We reported a field study that compared the outcomes of a chatbot-driven survey and that of a typical online survey. The study involved about 600 participants, half of them took a chatbot survey on Juji ([juji.io](http://juji.io)) and the other half filled out a form-based online survey on Qualtrics ([qualtrics.com](http://qualtrics.com)). Compared to the form-based Qualtrics survey, the AI-powered chatbot survey was a conversational survey during which the chatbot provided interactive feedback to free-text responses, prompted for information, probed answers, and handled various social dialogues whenever appropriate. Through an in-depth analysis of over 5200 free-text responses collected from the study, our results showed that the participants who completed a chatbot survey provided significantly more relevant, specific, and clear free-text responses than their counterparts did in a Qualtrics survey. They were also more willing to spend time with the chatbot, provide longer responses, and disclose more information about themselves. 190 (67.4%) of the participants who engaged with the chatbot also expressed their positive experience and willingness to take surveys in a chat format.

Given our study results and the simplicity of creating and deploying a chatbot survey like the one used in our study, our work suggests a new and promising method for conducting effective surveys especially for the purpose of collecting free-text responses to open-ended questions and overcoming survey taking fatigue. With the increasing use of chatbots, our results also present important design implications for creating and employing chatbots for survey success. In particular, chatbots should be equipped with active listening skills to guide participants in a conversational survey, which will elicit higher quality responses and deliver better engagement experience. Chatbots can also be used to deliver early interventions at the beginning of a survey, which would encourage quality responses and prevent cheating. Moreover, it is desirable to mix the use of a chatbot and a typical online survey to accommodate different survey goals and maximize survey success.

## CHAPTER 4: CONVERSATIONAL AI WITH ACTIVE LISTENING SKILLS

### 4.1 OVERVIEW

As we demonstrated in the last chapter, information collection with conversational AI could collect higher-quality information and drive a more engaging experience. In this chapter, our goal is to build more effective conversational AI for information collection<sup>14</sup>. To conduct an effective conversational survey, the agent should have skills similar to that of effective human interviewers [101, 214]. One such important skill is active listening — the abilities to understand and respond to a conversation partner properly [215, 216]. Active listening is shown to facilitate interviews, e.g., eliciting higher quality responses [215, 216, 217] and making an interviewer more socially attractive [218]. In addition, studies find that active listening helps not only oral communication, but also online text communication, including text messaging [219, 220]. Inspired by these findings, we hypothesize that interview chatbots with active listening would be more effective at conducting interviews and engaging interviewees. Fig 4.1 shows an example of such a chatbot, which can understand the user’s input and summarize it in its response, making the user feel heard.

Despite recent advances in Artificial Intelligence (AI), it is still challenging to build capable chatbots [14], let alone create chatbots with active listening skills. Below we highlight three main challenges specific to building effective interview chatbots with active listening skills.

First, it is challenging to build interview chatbots that can effectively grasp and respond to user input to open-ended interview questions, which is the core of active listening.

For example, in one of our user surveys, a chatbot asked an open-ended question “*what’s the top challenge you’re facing*”. One user responded:

*“The biggest challenge I’ve faced is finding a since [sense] of purpose. Being around like minded individuals who are constantly wanting more out of life through countless jobs I’ve never found something I was proud of as much as...”*

Another user answered the same question differently:

*“With a new baby I have a lot of additional expenses. So I have to try to obtain additional income. I try to earn extra income by working on mturk, but the pay is low and I don’t like the additional time taken away from my...”*

---

<sup>14</sup>This Chapter has been published at ACM CHI Conference on Human Factors in Computing Systems (CHI 20) [213]

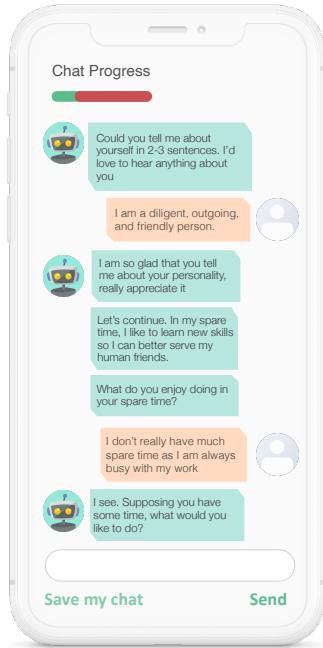


Figure 4.1: A screenshot of an example interview conducted by a chatbot (AI Minion) and a user (Sara).

Given such user input, an effective chatbot should respond to each person empathetically to make them feel heard. Few chatbot platforms however enable chatbots to handle such complex and diverse user input. For example, popular chatbot platforms like Chatfuel<sup>15</sup> and Manychat<sup>16</sup> hardly handle user free-text input. More advanced platforms like Google Dialogflow<sup>17</sup> and IBM Watson Assistant<sup>18</sup> support Natural Language Processing (NLP), but they require that a chatbot designer enumerate all user intents to be handled. With such a method, it would be very challenging to build an interview chatbot, since it is difficult to anticipate diverse responses to open-ended questions and enumerate all possible user intents.

Second, it is difficult to build interview chatbots that can effectively handle complex conversation situations to complete an interview task. As indicated by a recent report, natural language conversations are nonlinear and often go back and forth [14]. In an interview, an individual may digress from a planned agenda for various reasons. For example, some people may not understand an interview question and want clarifications, while others might dodge a question by responding with “*Why do you want to know?*” or “*I don’t know.*” People might misunderstand a question or simply do not know how to answer it. For example,

---

<sup>15</sup>[chatfuel.com](http://chatfuel.com)

<sup>16</sup>[manychat.com](http://manychat.com)

<sup>17</sup>[cloud.google.com/dialogflow](http://cloud.google.com/dialogflow)

<sup>18</sup>[www.ibm.com/products/watson-assistant](http://www.ibm.com/products/watson-assistant)

one person offered an ambiguous response to the question mentioned above:

*“Most challenges are met as an opportunity to grow. Hardest part is losing friends.”*

People may also be “uncooperative” and intentionally provide gibberish or irrelevant responses, such as those observed in crowd-sourced studies [170].

To complete an interview task, a chatbot must “remember” and stick to an interview agenda no matter how many times or how far a conversation has digressed from the agenda. However, most existing chatbot systems support scripted dialog trees instead of dynamic, graph-like conversations required by effective interview chatbots.

Third, it is difficult for chatbot designers to take advantage of AI advances due to a lack of AI expertise or resources. For example, deep learning has enabled powerful conversational AI [139, 140, 141, 142, 143] and might help address the first challenge mentioned above. However, these models require large amounts of training data (i.e., interview data), which are hard to acquire.

Given the three challenges mentioned above, we explore new ways to build effective interview chatbots. As the first step, we are investigating the feasibility and effectiveness of using existing AI technologies to build effective interview chatbots with active listening skills.

Our investigation aims at answering two research questions:

- **RQ4.1:** Whether and how can we employ publicly available AI technologies to build effective interview chatbots with active listening skills?
- **RQ4.2:** How effective can such interview chatbots be at handling complex and diverse user input and affecting user experience and interview quality?

To answer the above questions, in this chapter, we have developed a prototype system for building chatbots with active listening skills. We used our prototype to create two chatbots with and without active listening skills, respectively. We evaluated both chatbots live with 206 Amazon Mechanical Turkers to compare their performance by a set of metrics, including the quality of user responses and user perception.

As will be seen, our answers to the two research questions demonstrate the feasibility and effectiveness of using publicly available AI technologies to build effective interview chatbots. As a result, our work offers three contributions:

- *Practical approaches to effective interview chatbots.* Our work presents practical implementations to power chatbots with a specific set of active listening skills.

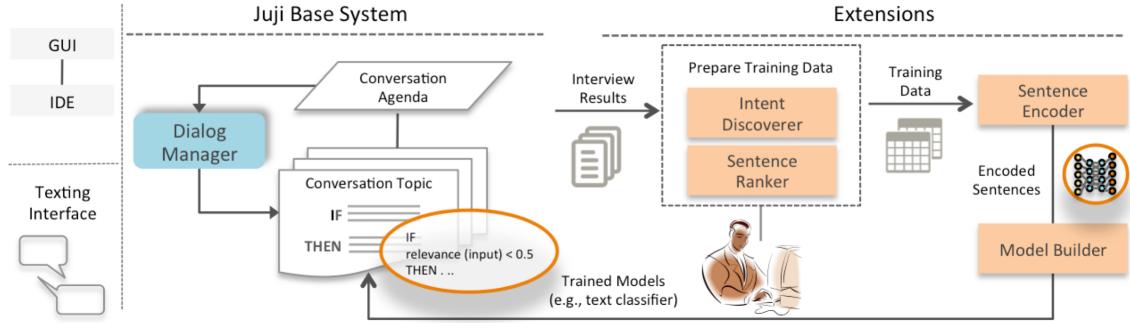


Figure 4.2: Overview of our prototype system for building an effective interview conversational agent with active listening skills. The prototype system extends Juji’s base system by bootstrapping conversation history and identifying more fine-grained intents to enable active listening skills.

- *A hybrid framework for developing progressive chatbot platforms.* Our work demonstrates a hybrid chatbot design framework, where rules can be used to bootstrap a chatbot and data-driven models can then be used to improve the chatbot’s conversation capabilities.
- *Design implications for building empathetic chatbots beyond interview tasks.* Since active listening aids effective communications beyond interviews, our work presents design considerations for building empathetic chatbots for a wide variety of applications, including counseling and training, beyond interview tasks.

## 4.2 PROTOTYPE OVERVIEW

Our prototype is intended to demonstrate the feasibility of building effective interview chatbots with existing AI technologies. To achieve this goal, we have investigated a number of approaches to conversational agents, which fall into two broad categories: rule-based and data-driven [20].

On the one hand, a rule-based approach uses explicitly coded knowledge (e.g., grammar) to handle user input [221, 222]. Rules are easy to understand and can also be used to bootstrap a chatbot quickly. However, it requires expertise and manual efforts to code the knowledge. Moreover, it is difficult to code complex, implicit knowledge as in the complex user input shown earlier.

On the other hand, data-driven approaches, such as deep learning, have shown their promises for handling complex and diverse conversations (e.g., [102, 139, 143]). Such approaches, however, require intensive computational resource and large training data sets. Moreover, chatbot designers who are not AI experts may not know how to fine-tune the approaches or

control the use of the results to prevent inappropriate chatbot behavior.

Based on our investigation, a hybrid system that combines rule-based and data-driven approaches seems most promising. We thus have developed our prototype with two main parts: (1) a rule-based system and (2) extensions in support of data-driven models.

As shown in Fig. 4.2, Juji, a publicly available chatbot platform ([juji.io](http://juji.io)), serves as our rule-based system. As detailed later in Sec. 4.3.3, Juji was chosen for several reasons. On Juji, chatbot designers can create, customize, and deploy a chatbot with a graphical user interface (GUI) or an interactive development environment (IDE). Authoring a Juji chatbot is to define a *conversation agenda* (interview agenda) that consists of one or more conversation topics (interview topics) and the temporal order of the topics to be discussed. To drive a conversation, the *dialog manager* activates and manages each topic on a conversation agenda by their temporal order [19]. End users (e.g., interviewees) can interact with a Juji chatbot via a texting interface on a website or via Facebook Messenger.

Our prototype also includes a set of extensions that support data-driven approaches to improve chatbots. In particular, these extensions help chatbot designers incorporate data models from two aspects: (a) preparing training data and (b) training models to handle user free-text input. To prepare training data, the *intent discoverer* and the *sentence ranker* work together to *automatically* identify user intents and label training data. Since automated analyses are often imperfect, a human can examine and rectify mislabeled data.

The *sentence encoder* encodes labeled training data into fixed-length dense vectors to capture rich linguistic features. The *model builder* then uses the encoded data to train various models, such as text classification models for user intent prediction. The trained models can then be used to control chatbot behavior. For example, one model may predict how semantically relevant a user input is to an open-ended interview question. Based on the prediction, a chatbot can respond properly during an interview.

## 4.3 PROTOTYPE DESIGN AND KEY COMPONENTS

Here we present the scope and technical focus of our prototype along with our design criteria. We then describe the key components of our prototype.

### 4.3.1 Prototype Scope and Technical Focus

In human conversations, active learning refers to one's ability to understand his/her conversation partner and respond properly [216, 223]. Active listening is shown to facilitate interviews, e.g., eliciting higher-quality responses [139] and making an interviewer more

Technique	Synopsis	Example
<b>Paraphrasing</b>	Restate a user input to convey understanding.	<i>“I see you love to hang out with your friends.”</i>
<b>Verbalizing Emotions</b>	Reflect a user’s emotions in words to show empathy.	<i>“I can tell intellectual activities make you happy. Just keep doing what you love.”</i>
<b>Summarizing</b>	Summarize the key ideas stated by a user to convey understanding.	<i>“If I hear you right, you care about others and have great leadership potential.”</i>
<b>Encouraging</b>	Offer ideas and suggestions to encourage conversation.	<i>“You’ve made an interesting point, could you elaborate? ”</i>

Table 4.1: Active listening skills supported by our prototype. In this study, our prototype supports four key active listening skills: paraphrasing, verbalizing emotions, summarizing, and encouraging.

socially attractive [224]. In addition, studies find that active listening helps not only oral communication, but also online text communication, such as text messaging and email [219, 220]. Inspired by these findings, we hypothesize that interview chatbots that can actively listen would be more effective at conducting interviews and engaging interviewees.

Active listening can be carried out through a number of communication techniques, such as paraphrasing, verbalizing emotions, and summarizing [219, 225]. As a start, we focus on supporting a subset of them (Table 4.1). For example, we exclude impromptu questioning techniques, such as *balancing* [219], since currently we build chatbots for only structured interviewing with pre-defined questions.

Although active listening techniques vary in forms, there is a minimal technical requirement for an interview chatbot to apply any technique: it must comprehend the semantics of a user response. To *encourage* a user to provide a better response, for example, a chatbot must at least understand how semantically relevant the user response is to its question. Similarly, a chatbot must understand the underlying semantics of a user response to *summarize* the response.

To enable active listening, our technical focus thus is on interpreting the semantics of user free-text responses.

### 4.3.2 Prototype Design Criteria

When designing our prototype and selecting component technologies, we faced many choices. In general, our decisions were guided by three criteria:

- **Reproducibility.** We select only publicly available technologies so other researchers and practitioners can easily reproduce our work.
- **Practicality.** We prefer technologies that require low resource and little training or customization.
- **Adoption.** We favor technologies that non-AI experts can use to build custom interview chatbots.

### 4.3.3 The Juji Base System

Among many chatbot platforms, we chose Juji as the rule-based system of our prototype for several reasons. First, Juji meets all three design criteria described above. Specifically, Juji is publicly available, and its rule-based approach allows a chatbot designer to bootstrap a chatbot without any training data. Second, Juji has demonstrated its success in support of interview chatbots. Several recent studies show various interview applications of Juji chatbots, such as job candidate interviews [101], market research interviews [173], and student interviews [226, 227, 228]. In addition to these two reasons, we chose Juji for another two important reasons.

**Ability to Manage Non-linear Conversations** Conducting interviews is a complex task, which must handle a task-oriented (e.g., [19]) and an open-domain social dialog (e.g., [139, 140]). Specifically, an interview chatbot has a task (interview) to complete. An interview by nature is also a social dialog during which an interviewee may “wander off” from an interview topic. Here we use *side talking* to refer to any conversations outside the pre-defined interview topics on an interview agenda.

To ensure interview quality, an effective interview chatbot must recognize and handle side talking properly, no matter where and how many times it occurs during an interview. Juji is the only chatbot platform that we could find and can handle complex conversation flows automatically. Specifically, Juji keeps track of a conversation context automatically and always brings a conversation back on track from side talking [19, 173]. Table ?? is an example showing how Juji handles multiple user interruptions and still manages to bring the user back to the original question.

---

Chatbot:	I'd love to learn what kind of reader you are. What types of books do you like to read?
User:	I don't know. What about you?
Chatbot:	Sorry, I cannot read yet. Could we go back to my question?
User:	What was your question?
Chatbot:	I was asking: What types of books do you like to read?
User:	It's really hard to say since I read a lot.
Chatbot:	No worries, just share what's on your mind.
User:	I guess my favorite kind, would be sci-fis.

---

Table 4.2: An example of how our chatbot handles a non-linear conversation. In this example, the chatbot handled the user’s question and stuck to the information goal.

**Extensibility** We selected Juji also for its extensibility as its rules take third-party functions or API [84]. For example, an external gibberish-detection function can be embedded in a rule to handle user gibberish input. This feature allows us to extend Juji easily. Moreover, Juji topic-based conversation model allows us to make targeted, topic-specific improvements. Compared to improving end-to-end dialogs (e.g., [139, 140]), supporting topic-specific improvements has two benefits. First, it requires much less training data to achieve good performance (see Sec. 4.4.1). Second, it avoids propagation of modifications or errors to other topics. In contrast, in an end-to-end approach, training data may improve certain parts of a dialog but adversely affect other parts. It is often difficult to control such effects.

#### 4.3.4 The Extensions: Data-Driven User Intent Prediction

To power interview chatbots with active listening skills (Table 4.1), the key focus of our prototype is to interpret the meaning of complex and diverse user input. As Juji pattern-based rules cannot handle such user input, we have extended Juji to support data-driven approaches to semantic interpretation. However, interpreting fine-grained semantics of natural language expressions is still very challenging even with data-driven approaches. We thus scaled down the challenge to focus on identifying the *semantic gist* of user input—the high-level intent that users imply.

Recent advances show the effectiveness of using text classification for identifying implied user intent (e.g., [139, 140, 143, 229]). Additionally, text classification can process text at scale, which is important in our application since an interview chatbot may converse with thousands of users at the same time. Rich public resources are also available for building

text classification models (e.g., Google developers resource). Our extensions thus focus on supporting text classification models to identify the intent of user input.

To facilitate the construction of text classification models especially for non-AI experts, we have built a set of components to support a three-step model construction: (a) preparing training data, (b) encoding training data, and (c) training text classification models.

**Preparing Training Data: Machine-Aided Labeling** A key step for building text classification models is to obtain labeled training data. User responses related to an interview question (topic) can serve as the training data for building topic-focused classification models. For example, on the interview topic “*What’s your top challenge*”, user responses to this question can be used as training data to build classification models for improving the conversation around this topic. However, labeling user responses is non-trivial. First, these responses are highly diverse. It is difficult for humans to consistently code the intent for a large number of responses. Second, user responses may be complex and convey multiple intents simultaneously, which makes human coding even harder.

We support a 3-step process to facilitate data labeling: intent discovery, auto-labeling and human validation.

**Step 4.1. LDA-based Intent Discovery** To classify user intents, we first identify the intents conveyed by training data. We chose to use Latent Dirichlet Allocation (LDA) model for this task because LDA is effective at extracting hidden intents (topics) [230]. It is an unsupervised technique and requires no training. Moreover, LDA implementations are publicly available.

In our prototype, we implemented LDA with Gensim [231], an open-source library. Given a set of training data (user responses), LDA automatically derives a set of intents (topics) from summarizing these responses (documents). Since these extracted intents are unordered, we enhanced the LDA results by ranking the intents by their coverage [232]. For example, the LDA analysis of 2680 user responses to “*what’s your top challenge*” identified top-3 intents by coverage: coping with changes (23.15% of user responses), people problems (21.27%), and time management issues (15.03%).

**Step 4.2. Centroid-based Sentence Ranking** The LDA-extracted intents (topics) are typically summarized by a set of words. However, it is difficult to label the intents by the associated words due to missing context [233]. On the other hand, a user response often conveys multiple intents, it is difficult to identify representative user responses (positive examples) of a given intent.

We thus further enhance the LDA results to automatically rank user responses by their semantic proximity to an intent. In particular, we implemented a centroid-based approach. Given an identified intent, we first clustered user responses whose probability distribution over the intent exceeds a threshold. We then used LexRank to rank responses based on their lexical centrality and semantic proximity to the cluster centroid [234]. The higher-ranked responses can be considered positive examples for identifying the intent. The lower-ranked ones can then be used as negative examples. The identified positive responses also help a human understand and label an associated intent easily.

**Step 4.3. Human Validation** Given a set of user responses, the two steps described above auto-label these responses as positive and negative examples representing user intent. The results however may not always be correct. A human should always validate the auto-labeled data and decide whether there is enough data to train a classification model. A human could always adjust various parameters used in steps 1-2 (e.g., adjusting the number of user intents to be identified by LDA) to redo the analysis and obtain new labeled data.

**Encoding Training Data: Sentence Embedding** To train text classification models, we need to represent the training data uniformly. We have experimented with several encoders and we chose to use the publicly available Universal Sentence Encoder (USE) in the Google TensorFlow library to encode each example as a 512-dimension vector [235]. We made this choice for several reasons. First, the encoder is trained and optimized for representing longer text like ours. It automatically captures rich, latent features in the text. Second, it is trained on a wide variety of data sources and generalized for diverse NLP tasks. Third, the TensorFlow library publishes its internal variables, and we can fine-tune the model with our own data as needed.

Moreover, USE meets our design criteria better than other encoders, especially by balancing performance and resource requirements. For example, we experimented with the last layer of pre-trained BERT-Large model for sentence embedding [135]. It produced comparative performance (e.g., F1 scores) but required twice the size of the USE to store the smallest encoding. Although we can always fine-tune en-coders to improve their performance [135], we did not do so in our current prototype, since our goal is to examine the effectiveness of the technologies as is—“lowest-hanging fruits” before exploring alternatives (Sec. 4.3.2).

**Training Text Classification Models** Given the encoded training data, training text classification models are straightforward. For their interpretability, performance, and requirement on training data sets, currently, we focus on training binary classification models that

output a probability score [22, 37].

#### 4.3.5 Enabling Active Listening

Since our prototype uses Juji conversation topics and the rules defined in these topics to drive chatbot behavior, we incorporate the trained classification models into the rules associated with specific topics. The rules will be triggered by the prediction results at run time to guide the generation of proper system responses, enabling active listening. Below we use a concrete example to demonstrate the use of such prediction models.

Consider the interview topic “*what’s your top challenge*”. Four classification models are trained to process user input and handle four high-level user intents on this topic. The first model (*Relevance*) predicts whether a user response is relevant to the topic, and another three predict whether a user response implies one of three intents, respectively: *time management issue* (C1), *people problems* (C2), and *coping with changes* (C3). To incorporate these models, we add a rule attached to the topic:

```
IF      Relevance(?u) >threshold1 && (?model ← (max  
          (C1(?u), C2(?u), C3(?u)) >threshold2))  
THEN  generate-response(?model)
```

The above rule states that if a user input (?u) is relevant and implies one or more of the three intents, the chatbot generates a response based on the best-detected intent. Assume a user input:

“*I think the main challenge will be starting a new job, . . . I will have to learn new ways of doing things, and starting over is not always easy.*”

In this case, model C3 predicts the user input as “*coping with changes*” with the highest probability. The chatbot then uses templates to generate a set of system response candidates based on the identified semantic intent and active listening techniques (Table 4.1). Below is a set of responses generated using the summarizing technique on the “*coping with changes*” intent:

“*Your description really resonates with me as I also struggle coping with changes or new settings.*” “*I would feel the same in your situation since handling new things is always challenging. Thanks for sharing.*” “*Coping with changes is always hard and I wish I could help you in such situations once I become smarter.*”

Currently, a response is randomly selected [101]. Note that Juji always offers a default response if no user intent can be predicted with a certain level of confidence.

## 4.4 EVALUATION

To evaluate the performance of our prototype, we have conducted extensive experiments. Here we report two sets of results: (1) effectiveness of predicting user implied input, and (2) impact on user response and experience.

### 4.4.1 Evaluating User Intent Prediction Models

To test the feasibility of our prototype for powering interview chatbots with active listening skills, we conducted a set of experiments to measure how well text classification models can predict the implied intents of user responses. The accuracy of these models directly determines how well an interview chatbot can listen actively during an interview.

We have been using the Juji base system to build chatbots for various real-world applications, such as student surveys. In these applications, we used Juji built-in topics as it provides over 60 topics for building chatbots [19]. From these applications, we have accumulated a number of user responses on various topics. We decided to test our prototype on improving four most popular interview topics:

- **Q4.1:** *Could you tell me about yourself in 2-3 sentences?*
- **Q4.2:** *What do you enjoy doing in your spare time?*
- **Q4.3:** *What is the best thing about you?*
- **Q4.4:** *What is the biggest challenge you face now?*

We chose the above four topics for the following reasons. First, those topics often appear in an interview to help establish rapport [236] and usually elicit diverse user responses. Handling those commonly used topics well can benefit many real-world applications. Second, handling more specific interview topics (e.g., asking about one’s work experience) requires deeper domain knowledge and in turn more advanced AI/NLP. Since we didn’t know if off-the-shelf AI would work, we focused on general interview topics first. Third, our extensions require training data. Since these four topics are the most common topics in actual interviews, they produced most of the training data.

To test the effectiveness of our approach with small training sets, we created a training data set for each topic by randomly selecting about 4000 user responses on that topic.

For each interview topic, our enhanced LDA model first analyzed its training set and identified 4-5 implied intents, each of which covered at least 10% training data. For each

	<b>Precision</b>	<b>Recall</b>	<b>F1</b>	<b>Accuracy</b>
<b>Q4.1 (self intro)</b>	0.7661	0.8433	0.8001	0.8506
<b>Q4.2 (hobbies)</b>	0.9003	0.7995	0.8332	0.9004
<b>Q4.3 (best about u)</b>	0.8061	0.8367	0.8199	0.8098
<b>Q4.4 (top challenge)</b>	0.8653	0.5728	0.6543	0.9215

Table 4.3: Stratified 10-fold cross validation for four topics. Logistic Regression for Q4.1, and SVM for Q4.2, Q4.3, and Q4.4.

identified intent, the centroid-based analysis then produced a set of ranked responses by their semantic proximity to the intent and auto-labeled the top 20% as positive examples and the bottom 20% as negative examples. A human then verified and amended the labels if needed. As a result, a total of 17 user intents were identified across four interview topics, with about 1000 labeled training samples per intent.

We then trained a total of 68 text classification models for 17 user intents, each with four popular classification methods: logistic regression, linear SVM, Adaboost, and Naïve Bayes. For each of 68 models, we performed stratified 10-fold cross validations and examined four standard performance metrics: *precision*, *recall*, *F1*, and *accuracy*. Since we cannot fit all 68 sets of results in the paper, here we report the overall results and two representative sets.

Table 4.3 shows the averaged performance of best models per topic. Overall, logistic regression and linear SVM performed the best across all data sets. Logistic regression performed the best over more heterogeneous data sets, while linear SVM performed the best on more homogeneous data sets. Table 4.4 shows two such example results. At the top of the table, logistic regression performed the best over heterogeneous expressions on leadership—user responses to “*what’s your top talent*”. These responses are syntactically diverse and semantically complex. Here are two responses:

“*I always think through things and try to do everything I care for the people around me and those I care about. I am quick to give advice and emotional support, and I will always give my time when I feel that I have it*”

“*I am an encouraging person and I know that I can help others and unite a group that is starting with myself. . . . I will give my all 100 % of the time. I believe not only in myself but in others and I think that is something really impactful when it comes to moving forward...*”

On the bottom of Table 4.4, linear SVM performed the best over more homogeneous expressions on hanging out with friends—user responses to “*what do you enjoy doing in your*

<b><i>Leadership</i></b> (positive examples: 567, negative examples: 479)				
	Precision	Recall	F1	Accuracy
<b>Logistic Regression</b>	<b>0.7795</b>	0.7654	0.7724	<b>0.7534</b>
<b>Linear SVM</b>	0.7624	<b>0.7865</b>	<b>0.7743</b>	0.7505
<b>AdaBoost</b>	0.7463	0.7462	0.7462	0.7228
<b>Naïve Bayes</b>	0.7402	0.7457	0.7429	0.7188
<b><i>Hangout w/ friends</i></b> (positive examples: 529, negative examples: 492)				
	Precision	Recall	F1	Accuracy
<b>Logistic Regression</b>	0.9118	0.8865	0.8990	0.8962
<b>Linear SVM</b>	<b>0.9167</b>	<b>0.894</b>	<b>0.9052</b>	<b>0.9020</b>
<b>AdaBoost</b>	0.8593	0.8506	0.8549	0.8502
<b>Naïve Bayes</b>	0.8527	0.8657	0.8452	0.8345

Table 4.4: Stratified 10-fold cross-validations for two data sets. Overall, logistic regression and linear SVM performed the best across all data sets. Logistic regression performed the best over more heterogeneous data sets, while linear SVM performed the best on more homogeneous data sets.

*spare time*”. Here are two examples:

*“I like to spend time with my friends—we talk or do fun outdoor activities together.” “I enjoy spending time with friends on the weekends. ... surfing for me was a great way to make friends.”*

#### 4.4.2 Live Chatbot Evaluation

To evaluate our prototype in the field, we designed and conducted a between-subject study that compared the performance of chatbots with and without active listening.

**Study Design** We designed a 10-minute interview on six topics, including the four popular topics mentioned in Sec. 4.4.1. After chatting on each topic, users were asked to rate how well the chatbot understood them. In addition to the four topics, users were asked of their opinion about the chatbot and what the chatbot could do for them. Near the end of an interview, users were also asked to rate the chatbot on two more aspects: their interest of chatting with the chatbot in the future and their overall chat experience. All the ratings were on a 5-point Likert scale, 1 being poor and 5 being excellent. We also collected user basic demographics such as gender and age group.

We built two chatbots for this study. One was built with only the Juji base system and served as the baseline (Baseline). The other was built with our full prototype, first bootstrapped by the Juji based system and then improved by the text classification models mentioned in Sec. 4.4.1. This chatbot (Full Version) demonstrated active listening on the four popular interview topics. Both chatbots asked the same interview questions in the same order.

We randomly assigned the participants to either chat with the baseline or the full version of chatbot. In both conditions, participants were given the same message, “interviewing by an AI chatbot”, and a randomly assigned chatbot URL to join the chat.

**Participants** We recruited participants on Amazon Mechanical Turk with an approval rate equal to or greater than 99% and located in the U.S. or Canada. We paid each participant \$12.5/hr.

**Measures** Active listening is supposed to improve interviewee experience and the quality of an interview [173]. We thus compared the effects of the two chatbots on user experience and interview quality by a set of metrics.

**Engagement duration** This measures how long a user engaged with a chatbot, which implies a user’s willingness to engage and the quality of interaction [217].

**Response length.** This counts the total number of words in a user’s text responses, which also indicates a user’s willingness to engage especially in an interview [217].

**Response informativeness** To estimate the quality of an interview, we computed the *informativeness* of each user’s text responses ( $R$ ) based on [173] to measure the richness of information contained in the responses by “bits”:

$$Informativeness(R) = \sum \log_2 \frac{1}{F(word_n)} \quad (4.1)$$

By this formula, the more frequently a word (e.g., the common word “the”) appears in modern English communications, the less information it conveys.

**Response Quality Index (RQI)** To measure the response quality of each participant, we created a *Response Quality Index (RQI)* as in our previous chapter (Sec. 3.2.5). It measures

the overall response quality of N responses given by a participant on three dimensions (specificity, relevance, and clarity):

$$RQI = \sum_{n=1}^N relevance[i] \times clarity[i] \times specificity[i] \quad (4.2)$$

(N is the number of responses in a completed survey)

**User ratings** For each participant, we computed three 5-point Likert scale. One rated the comprehension of a chatbot and an index (*agentC*) was created by adding up the participant’s ratings on each of the four topics. Another (*interestR*) rated a participant’s interest of chatting with a chatbot, while the third (*chatR*) rated the participant’s chat experience.

**Results** We received a total of 206 completed interviews: 108 engaged with the full version of chatbot (56% female, 44% male), while 98 interacted with the baseline (37% female and 63% male). The participants were all above 18, and 134 (65%) of them were between 18-34 years old. On average, the participants spent 9.33 minutes with the full version of chatbot and 7.82 minutes with the baseline. To compute RQI for each participant, we manually coded each response following a similar process described in [173]. A total of 824 responses were coded, each with three dimensions on a 3-point Likert scale: 0 (bad), 1 (ok), and 2 (good).

To compare the true effect of two chatbots, we chose to use ANCOVA analyses, a general linear model blending ANOVA and regression [237]. In each ANCOVA analysis, the independent variable was the version of chatbot used, and the dependent variable was one of the measures described above. Each analysis was controlled for demographics, which may influence the results, according to previous studies [155].

Table 4.5 summarizes the results. The chatbot with active listening skills (full version) outperformed the baseline significantly across all measures. The full version scored higher on agent comprehension (*agentC*), user interest (*interestR*), and user chat experience (*chatR*). It also chatted with the participants longer (*engagement duration*), and elicited more words (*response length*), richer information (*informativeness*), and higher-quality responses (RQI).

Additionally, we evaluated the performance of predicting user intents from the participants’ input. Since there is no space to list individual prediction results for all 68 models, Table 4.6 shows the averaged performance of the best prediction models for each topic. The models performed worse in the real world than the cross-validations for Q4.3 and Q4.4, while showing comparable results in Q4.1 and Q4.2 (Table 4.3).

Overall, the models used for predicting user intents in Q4.2 (hobbies) performed the best, while the models for Q4.4 (top challenge) performed the worst. In Q4.4, the participants’

	Full Version		Baseline		<i>F</i>	<i>p</i>	$\eta_p^2$
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
<b>Engagement Duration</b> (mins)	9.33	4.47	7.82	4.85	$F(1,201)=5.79$	$<0.05^*$	0.03
<b>Response Length</b> (words)	125.70	56.96	100.65	48.57	$F(1,201)=11.72$	$<0.01^{**}$	0.05
<b>Informativeness</b> (bits)	109.66	52.32	90.21	43.78	$F(1,201)=8.29$	$<0.01^{**}$	0.05
<b>Response Quality Index (RQI)</b>	21.28	6.00	16.83	6.28	$F(1, 201)=26.98$	$<0.01^{**}$	0.12
<b>Agent Comprehension (agentC)</b>	13.22	3.74	11.13	4.28	$F(1, 202)=13.87$	$<0.01^{**}$	0.06
<b>User Interest (interestR)</b>	3.56	1.24	3.07	1.29	$F(1, 202)=7.79$	$<0.01^{**}$	0.04
<b>User Chat Experience (chatR)</b>	3.81	1.03	3.19	1.20	$F(1, 202)=15.97$	$<0.01^{**}$	0.07

- a. All results were controlled for participant demographics, including gender and age group.
- b. Results for Informativeness, RQI, and Response Length were also controlled for Engagement Duration.
- c. Results for Engagement Duration were also controlled for Response Length.

Table 4.5: Comparison results of two versions of chatbots with ANCOVA analyses. Compared to the baseline, the chatbot created by our prototype delivered more engaging experiences and collected higher-quality information.

	Precision	Recall	F1	Accuracy
<b>Q4.1 (self intro)</b>	0.7673	0.6784	0.6810	0.7993
<b>Q4.2 (hobbies)</b>	0.8634	0.8536	0.8372	0.9267
<b>Q4.3 (best about u)</b>	0.6065	0.8064	0.6246	0.8007
<b>Q4.4 (top challenge)</b>	0.6357	0.3605	0.4268	0.9237

Table 4.6: Intent prediction for four interview topics. Logistic Regression for Q4.1 and Q4.3, and SVM for Q4.2 and Q4.4. The accuracy metric was biased due to many true negatives.

responses were *very different* from the training data, which was collected mostly from student interviews. For example, many participants talked about their financial challenges, which were not covered by our training data. While it is difficult to anticipate user responses to an open-ended interview topic, our prototype could be used to make fast incremental improvements. For example, new user responses can be analyzed and labeled using our prototype extensions and then used to train classification models for predicting new user intents and improving chatbot capabilities.

#### 4.4.3 Summary of Findings

Our extensive evaluation results helped answer our two research questions.

**RQ4.1:** It is feasible to use existing and practical AI technologies to build effective interview chatbots — chatbots with active listening skills.

**RQ4.2:** Chatbots with active listening skills are more effective at engaging users and eliciting quality user responses, compared to those without such skills.

## 4.5 LIMITATIONS

### 4.5.1 Effect of Individual Active Listening Techniques

Our work compared the performance of interview chatbots with or without active listening skills. We however did not examine how each skill impacts chatbot performance. Consider a user input “*I like outdoor activities, such as hiking and running*”. A chatbot can demonstrate active listening in multiple ways, e.g., one *verbalizing emotions* (VE) and the other *summarizing* the content (S):

*I can tell physical activities make you happy* (VE)

*It seems you are physically active* (S)

Without knowing which one would be more effective, now a chatbot randomly selects one. To make more effective use of different active listening techniques, we need to investigate their individual effect on different users (e.g., emotional vs. calm person) and on different interview topics (e.g., a sensitive topic discussing one’s hardship versus a generic topic on one’s hobbies). Since such investigations require carefully designed, fully controlled experiments, we leave them to our future work.

### 4.5.2 Interpreting Deeper User Intents

Our prototype supports the interpretation of high-level, hidden user intents—the semantic gist of user input. Several participants in our study voiced that certain chatbot responses are still vague and shallow. To generate more meaningful responses, a chatbot must extract deeper user intents, such as the conveyed semantic concepts and relationships among the concepts. Recent work on knowledge embedding that incorporates knowledge graphs with neural networks may offer a potential solution [238]. These approaches support diverse NLP tasks, such as relation extraction, which can be used to improve chatbots.

### 4.5.3 Interrelating Interview Topics

Based on Juji’s topic model, our chatbots treat each interview topic independently. In reality, interview topics may be semantically related. For example, in our field study, a participant made his self-introduction as follows:

*“I’m currently a student. I like to watch football in my spare time, and study I guess.”*

Later, the participant was annoyed by the hobby question, since he thought he already told the chatbot during the self-intro. Currently, the self-intro and the hobby are considered two separate topics. It does not use a user’s input given in one topic to influence the discussion on another. Nonetheless, this made the participant feel that the chatbot did not pay attention to her input or could not remember it.

To improve this situation, an interview chatbot needs to remember a user’s input, and use it to guide follow-on conversations. One potential solution is to build a knowledge graph to relate different interview topics. This would however require that user input be parsed into a knowledge graph that can be retrieved and reasoned [239]. The main challenge is to determine what knowledge entities (e.g., hobby) should be extracted since user free-text responses given in an interview are complex and highly diverse.

### 4.5.4 Active Listening by Asking

One key active listening technique not supported by our prototype is to ask impromptu questions based on a user response to deepening a conversation [217]. This requires that a chatbot automatically come up with follow-up questions based on a user’s input. Although there is research on question generation, it is often done in a static context [240, 241] or without a conversation goal [153]. We are exploring how to generate effective questions in a highly dynamic yet goal-oriented context like interviewing. This will enable an interview chatbot to follow on interesting ideas emerged during a conversation and discover unexpected, new insights.

### 4.5.5 Experimenting with Alternatives

When building our prototype, we made careful technical choices by our design criteria. Nevertheless, we recognize there are many alternatives that may be equally or more effective. For example, instead of using LDA to discover user intents, we could explore correlation explanation that requires less domain knowledge of the data [242]. We could also try fine-tuning sentence encodings [135] or training sentence embedding and text classification jointly

[243] to see if we could achieve better performance. As all these explorations require deeper expertise and more resources, we will do so in the near future and compare their effect on the performance of interview chatbots.

#### 4.5.6 Evaluating Prototype Usability

Although our prototype is intended to help non-AI experts build effective interview chatbots, we have not yet evaluated the usability of the prototype for two reasons. First, we want to verify the feasibility and effect of the prototype before evaluating its usability for chatbot designers. Second, our current prototype reuses the Juji GUI for customizing and deploying a chatbot, while providing a command-line interface for data labeling and model training. We feel this UI combination is cumbersome and a more integrated UI is needed as we advance our prototype.

#### 4.5.7 Evaluating in the Real World

While the four interview topics in our evaluation are the common topics which can benefit many real-world interviews, interview in the real world often contains more specific questions, such as one's previous working experience. Evaluating our prototype on more specific interview questions may bring different insights however it is also challenging. Compared to commonly used interview topics, specific topics appear less which leads to less training data to train a robust model for our prototype. Meanwhile, more specific requires deeper domain knowledge and in turn more advanced AI/NLP models.

Additionally, although chatbot interviews often happen online with participants similar to participants on Amazon Mechanical Turk, evaluating our prototype in the real world with actual interviewee may further benefit the generalizability of our evaluation.

### 4.6 DESIGN IMPLICATIONS

Our work demonstrates the feasibility of building interview chatbots with active listening skills and the effectiveness of such chatbots. It thus presents several design implications for building better chatbots and chatbot platforms.

#### 4.6.1 Practical Approaches to Effective Interview Chatbots

Our evaluation shows the effectiveness of an interview chatbot with active listening skills better at engaging users and eliciting quality user input. It implies that data-driven approaches such as text classification used in our prototype can power interview chatbots with active listening skills. Although our extension is based on Juji’s rule-based system, our extensions are platform agnostic since it can be added to any chatbot platforms to enhance chatbot capabilities. Moreover, data-driven models can be effectively trained for each topic with small training data sets, more practical than training end-to-end dialogs with large data sets.

#### 4.6.2 Hybrid and Progressive Chatbot Platforms

Our prototype shows how to support incremental chatbot improvements. Specifically, its hybrid chatbot design framework enables designers to bootstrap a chatbot with rules, use it to collect training data, and train models to improve it. Moreover, chatbot platforms should offer utilities like LDA-based intent discovery to help designers label training data and facilitate the improvement cycle.

#### 4.6.3 Building Empathetic Chatbots Beyond Interviewing

Active listening is used widely in situations like counseling and training [215, 216] beyond interviewing. An AI counselor or coach can also be powered with active listening skills to be more effective in their tasks. So far few systems support the easy creation and customization of empathetic AI agents with active listening skills. Our work is a stepping stone toward this direction to enable non-AI experts to create effective chatbots with active listening skills for a wide variety of applications beyond interview tasks

### 4.7 CONCLUSION

To investigate the feasibility of using publicly available technologies for building effective interview chatbots and the effect of such chatbots, we have presented a prototype that combines a rule-based chatbot builder with data-driven models to power interview chatbots with active listening skills. These skills enable a chatbot to better handle complex and diverse user responses to open-ended interview questions. As a result, such a chatbot delivers more engaging user experience and elicit higher-quality user responses.

## CHAPTER 5: A KNOWLEDGE-DRIVEN APPROACH FOR FOLLOW-UP QUESTIONS GENERATION

### 5.1 OVERVIEW

In this chapter, we further looked into the challenge of inflexible survey structure and aimed to generate informative follow-up questions for conversational surveys. Compared to structured form-based surveys, an open-ended conversational survey enables a more dynamic structure and personalized experience through follow-up questions. Although a good follow-up question could probe and prompt more information based on people’s responses, the current practice of delivering follow-up questions in conversational surveys is largely rule-based [14]. As a result, we explored an automatic way of generating follow-up questions for conversational surveys.

Compared to the follow-up questions in open-domain conversation, a good follow-up question in a conversation survey considers the goals and constraints of the survey objective. For example, the follow-up question should be relevant to the survey topic, should not be redundant or irrelevant, and should be phrased in a clear manner. It requires the system to understand the context deeply, adhere to the survey objective, and generate questions that are easy to understand.

To enable the automatic generation of follow-up questions in an open-domain conversational survey, we aim to tackle three specific challenges. First, there is no dataset for follow-up question generation in the context of open-domain conversational surveys. Related datasets are small or with a specific focus, such as job interviews [151] or graduate school admission interviews [149, 150]. Additionally, no prior dataset considers the background knowledge beyond the dialogue history, which limits the model’s ability to deeply understand the survey objective and context.

Second, existing methods for follow-up question generation are either template filling [150, 152] or seq2seq [149, 151, 153]. However, both methods lead to unsatisfying results. Template filling limits the diversity of question types and often fails to personalize based on the participant’s response, especially in dynamic and open-ended conversational surveys. As for standard seq2seq methods, they cannot generate questions that adhere to the overall survey objective and are relevant to the context. How to improve the diversity of generated follow-up questions while not limited by templates while utilizing relevant background knowledge remains an unresolved problem.

Third, no established metrics could effectively evaluate the generated follow-up questions for conversational surveys. The same dialogue history could inspire various valid follow-up

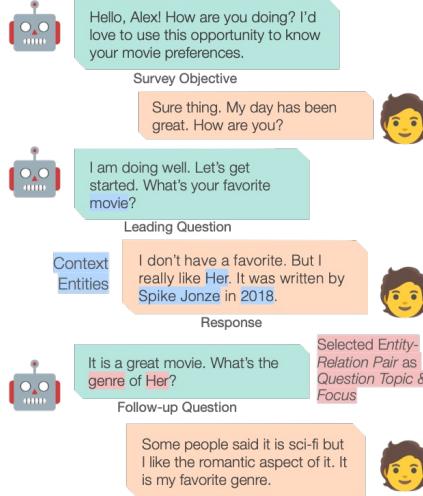


Figure 5.1: A conversational survey snippet that aims to understand people’s movie preferences where the agent generates a follow-up question based on the dialogue history and a selected entity-relation pair.

questions, and the same question could be phrased differently. Therefore, common text generation metrics depending on only one ground truth usually underestimate question quality. And human evaluation is hard to scale and compare [151, 244].

In this study, we proposed a novel task of knowledge-driven follow-up question generation in conversational surveys, as shown in Figure 5.1. To explore the feasibility and effectiveness of the task, we collected a human-annotated dataset including background knowledge, proposed a two-staged knowledge-driven baseline model and designed a set of reference-free Gricean-inspired evaluation metrics.

Specifically, we first collected a dataset with knowledge annotation and a human-written follow-up question based on the dialogue history and background knowledge to collect relevant and valuable information in a coherent and clear manner. To demonstrate feasibility, we then proposed a baseline model that leverages knowledge to steer the generative model. We imitated the human question-generation process by first using a knowledge selector to identify the question topic and focus [245] based on the dialogue history and a background knowledge base. We then combined the dialogue history and selected question topic and focus as a prompt for a generative model to generate follow-up questions that could lead to valuable and diverse information that contributes to the survey objective. To systematically evaluate the quality of generated follow-up questions, we designed a new set of reference-free evaluation metrics, *Gricean Scores*, based on Gricean Maxims [180]. Gricean Maxims is a collection of communication principles to which both speaker and listener should adhere to engage in effective communication. Our *Gricean Scores* (see Sec. 5.4) measured the quality of a follow-up question based on the following aspects: *Relevance*, *Informativeness*, *Truthfulness*,

*Clarity*, and *Coherence*. *Gricean Scores* align with human evaluation well and provide more insights than traditional reference-based metrics.

Our contributions are as follows:

- A new problem and a dataset for follow-up question generation in conversational surveys, which has background annotation and human-written follow-up questions.
- An effective baseline model that leverages knowledge as a control for large language models to generate informative, coherent, and clear follow-up questions.
- A set of reference-free evaluation metrics based on Gricean Maxims, i.e., *Gricean Scores*, that evaluate the quality of open-ended text from various perspectives.

## 5.2 DATASET AND ANNOTATION

Three principles guide our dataset construction: first, the topic needs to be diverse as conversational surveys can cover many topics. Second, the follow-up question should adhere to the conversation context, and aims to retrieve novel and specific information from the respondent. Third, the knowledge should be explicit, especially salient knowledge that guides the follow-up question generation.

### 5.2.1 Question-Answer Pairs Construction

We built our dataset based on OpenDialKG [246], which contains open-domain conversations between two human agents about a given topic such as movies, books, sports, and music. Those topics often appear in conversational surveys to build rapport and collect various responses. Each utterance in the original dataset has been manually annotated with entities from Freebase [247]. We first extracted question-answer pairs from OpenDialKG and then manually selected question-answer pairs that fit our goal.

### 5.2.2 Follow-up Questions Annotation

We invited workers on Amazon Mechanical Turk to create follow-up questions for each question-answer pair. We first presented a dialogue history, e.g., a question-answer pair, with the overarching survey objective and instructed people to imagine themselves as an interviewer who aims to collect informative and truthful information that contributes to the overall survey objective from their interviewees through a follow-up question. To facilitate

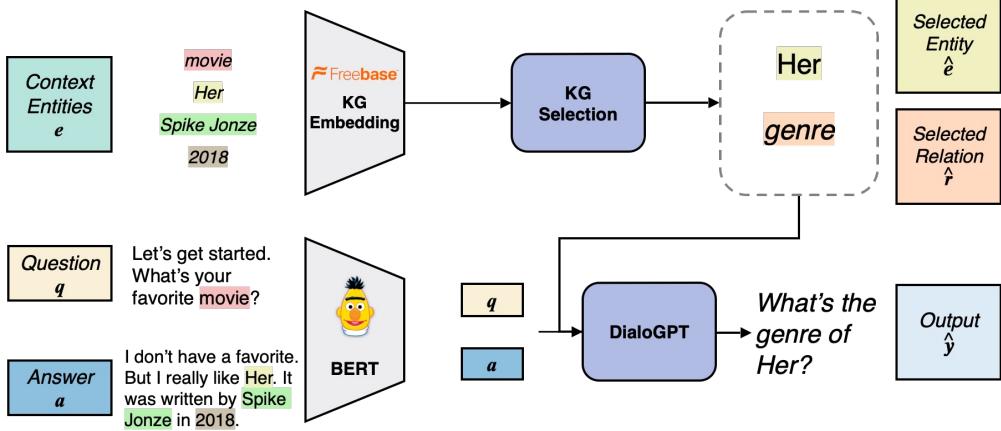


Figure 5.2: The overall architecture of the proposed framework. We propose a two-phase framework that can (i) select an entity-relation pair as the question topic and focus that is conditioned on the dialogue history to steer the generation relevant and adhere to the context (Sec. 5.3.2), and then (ii) generate a follow-up question that is triggered by the selected entity-relation pair (Sec. 5.3.3).

the creation of high-quality follow-up questions, we first asked annotators to select the most interesting and meaningful topic that they want to ask about in the follow-up question from mentioned knowledge entities in the dialogue history. Then, annotators needed to specify the relation of the selected entity as the follow-up question focus. As the last step, we instructed annotators to write a follow-up question in a clear and coherent manner based on the above-mentioned criteria. We manually went through all annotations and made minor edits to ensure quality. We compensated our workers at the rate of \$12 per hour.

## 5.3 MODELS

### 5.3.1 Overview

We formalized the task of follow-up question generation as follows: given a dialogue history  $X_i$  of the last two turns, consisting of a QA pair  $(Q_i, A_i)$ , with a set of context entities  $\mathcal{E}_i = \{e_1, e_2, \dots, e_n\}$  and their relations  $\mathcal{R}$ , a system is supposed to generate a follow-up question  $Y_i$  that is specific to the selected entity-relation pair  $(e_j, r_k)$ . Here  $e_j \in \mathcal{E}_i$ ,  $r_k \in \mathcal{R}(e_j)$ , and we denote  $\mathcal{R}(e_j)$  as the set containing all relations connected to the entity  $e_j$ .

Inspired by how human experts generate questions in semi-structured interviews [248], we propose a two-phase framework that can (i) select an entity-relation pair as the question topic and focus that conditioned on the dialogue history to steer the generation relevant and adhere to the context (Sec. 5.3.2), and then (ii) generate a follow-up question that is triggered by the selected entity-relation pair (Sec. 5.3.3). The overall structure of the proposed framework is shown in Figure 5.2.

### 5.3.2 Knowledge Selection

Our knowledge selection model first encodes a dialogue history  $X_i$ , a QA pair  $(Q_i, A_i)$ , using the pre-trained language model BERT [135]. We process the input text as a concatenation of the question  $Q_i$  and its answer  $A_i$ , and insert a special tag [CLS] at the beginning and another special tag [SEP] to separate them. The final hidden state that corresponded to [CLS] is used as the aggregated sequence representation of the dialogue history, which is denoted as  $\mathbf{h}_i^{\text{qa}}$ . Meanwhile, we encode entities and relations with pre-trained knowledge embedding (e.g., TransE [249]), and we represent them as  $\mathbf{h}_j^e$  and  $\mathbf{h}_k^r$  for the entity  $e_j$  and the relation  $r_k$ .

We consider two versions of methods to perform the knowledge selection: attention-based and MLP-based. The entity selection works in the same way as the relation selection. Here we only introduce the entity selection for simplicity. In the attention-based version, we regard the score of selecting the entity  $e_j$  as the unnormalized dot-product attention score [250] between  $\mathbf{h}_j^e$  and  $\mathbf{h}_i^{\text{qa}}$ , and we calculated it as follows:

$$\mathbf{s}_i = \frac{\mathbf{q}_i \mathbf{K}_i^T}{\sqrt{d_k}}, \quad (5.1)$$

where  $\mathbf{q}_i$  is the query vector such that  $\mathbf{q}_i = \mathbf{W}^Q \mathbf{h}_i^{\text{qa}}$ ,  $\mathbf{K}_i$  is the key matrix which is a stack of key vectors  $\mathbf{W}^K \mathbf{h}_j^e$  for all  $e_j \in \mathcal{E}_i$ , and  $d_k$  is the dimension of queries and keys. As for the MLP-based version, we simply concatenate  $\mathbf{h}_i^{\text{qa}}$  and  $\mathbf{h}_j^e$ , then input it into an MLP to predict the score:

$$s_{ij} = \text{MLP}([\mathbf{h}_i^{\text{qa}}; \mathbf{h}_j^e]), \quad (5.2)$$

where  $[.;.]$  denotes the concatenation operation.

Finally, we treat the selection task as a binary classification for each candidate entity and set the objective as minimizing the binary cross entropy:

$$\begin{aligned} \mathcal{L}_{\text{ent}} = & -\frac{1}{N} \sum_{i=1}^N \sum_{j:e_j \in \mathcal{E}_i} [y_j^i \cdot \log \sigma(s_{ij}) \\ & + (1 - y_j^i) \cdot \log(1 - \sigma(s_{ij}))], \end{aligned} \quad (5.3)$$

where  $\sigma(\cdot)$  is the sigmoid function,  $N$  is the number of samples,  $y_j^i$  is the ground truth label for the  $j^{\text{th}}$  entity in the  $i^{\text{th}}$  sample, and it equals 1 if the corresponding entity is selected and is 0 otherwise. During the training, we take the sum of  $\mathcal{L}_{\text{ent}}$  and  $\mathcal{L}_{\text{rel}}$  as the final objective

function.

### 5.3.3 Follow-up Question Generation

We formulate the follow-up question generation as a typical language modeling problem and adopt DialoGPT [251] to solve it. DialoGPT is a neural conversational response generation model, which is inherited from GPT-2 [252] and is pre-trained on 147M conversation-like exchanges extracted from Reddit comment chains. It has achieved state-of-the-art results on various dialogue generation benchmarks[252].

In this work, we serialize the input into one sequence and then input it into DialoGPT to estimate the language model. Specifically, we first concatenate  $Q_i$ ,  $A_i$  and  $Y_i$ , and add [EOS] at the end of each utterance such as  $[Q_i, \text{EOS}, A_i, \text{EOS}, Y_i, \text{EOS}]$ . With the selected entity  $e_i^*$  and relation  $r_i^*$ , we manually design a discrete template, such as "How to ask about", to combine  $e_i^*$  and  $r_i^*$  into a prompt  $P_i$ . The prompt is treated as an utterance, and is inserted between the dialogue history and the follow-up question to guide the generation. The final input can be represented as one sequence of tokens  $[u_1, u_2, \dots, u_m] = [Q_i, \text{EOS}, A_i, \text{EOS}, P_i, \text{EOS}, Y_i, \text{EOS}]$ . At each decoding step  $t$ , we compute the negative log-likelihood between the predicted distribution and the ground truth from the given response:

$$\mathcal{L}_{\text{NLL}} = - \sum_t \log p(u_t | u_{<t}) \quad (5.4)$$

Notably, during the inference time, we only input  $[Q_i, \text{EOS}, A_i, \text{EOS}, P_i, \text{EOS}]$  into DialoGPT and let it decode the response until reaching [EOS] as the generated follow-up question.

## 5.4 GRICEAN SCORES

To systematically assess the quality of generated follow-up questions, we proposed a set of reference-free and content-based metrics, *Gricean Scores*, grounded on Gricean Maxims [180]. The Gricean Maxims are a collection of communication principles, to which both speaker and listener should adhere to engage in effective communication. Researchers have been using Gricean Maxims to evaluate both human-human conversations [253, 254] and human-agent conversations [173, 255]. In the context of conducting conversational surveys, a "cooperative" interviewer would obey all the maxims to form questions and probe quality responses. Those maxims include *Quantity*, *Quality*, *Relation*, and *Manner*. For our purpose, we relied on the Gricean Maxims to define a set of reference-free metrics that quantitatively measure the quality of a follow-up question generated by our models. We measure the quality of a

follow-up question and compute *Gricean Scores* from five aspects: *Relevance*, *Informativeness*, *Truthfulness*, *Clarity*, and *Coherence*. Our reference-free metrics allowed us to evaluate a follow-up question without relying on single ground truth, making it more robust to open-ended text generation and real-world deployment. Also, the *Relevance*, *Informativeness*, and *Truthfulness* is knowledge-driven, which makes our score more contextualized. The survey designers could select a knowledge graph, customize one, or create one that best fits the survey objective and information value.

**Notation.** We defined indexed sets  $\hat{\mathcal{Q}}, \mathcal{T}$  such that a generated follow-up question  $\hat{Q}_i \in \hat{\mathcal{Q}}$  corresponds to a target follow-up question  $T_i \in \mathcal{T}$ . We referred the dialogue history of  $\hat{Q}_i$  as its context  $C_i$  and further represent  $\hat{Q}_i$  as a sequence of tokens  $[w_1, w_2, \dots, w_N]$ . We marked the recognized entity from  $\hat{Q}_i$  as  $\hat{e}_i$ , and the set of context entities as  $\mathcal{E}_i$ .

**Relevance.** By the Gricean Maxim of relation, a high-quality follow-up question should be pertinent to the prior discussion and adhere to the overall survey objective. Irrelevant follow-up questions elicit useless information and burden the analysis process. We measured *Relevance* by checking if the recognized entity in the generated follow-up question is from the context entity set:

$$\text{REL}(\hat{Q}_i) = \mathbb{1}[\hat{e}_i \in \mathcal{E}_i] \quad (5.5)$$

**Informativeness.** From the Gricean Maxim of Quantity, effective communication should be informative. In the context of a conversational survey, an agent should always ask questions that maximize the information gain in a participant’s response. In a knowledge graph, an entity’s out-degree centrality captures the outreach of other entities which indicates the potential information gain by probing this entity. We measured the *Informativeness* of a generated follow-up question by the out-degree centrality<sup>19</sup> of the recognized entity in the question:

$$\text{INFO}(\hat{Q}_i) = \text{Centrality}(\hat{e}_i) \quad (5.6)$$

**Truthfulness.** According to the Gricean Maxim of Quality, a cooperative interlocutor should communicate truthfully. In our case, a high-quality follow-up question should always lead to truthful information. For example, a survey respondent would respond to a question asking for a book’s release year to the best of their knowledge but not a question about a book’s director since the book does not have such an attribute. We measured *Truthfulness* of a generated follow-up question from an entity-relation perspective where the question should

---

<sup>19</sup>The centrality is normalized through dividing by the maximum possible degree in the graph

ask a valid relation connected to the recognized entity based on the knowledge base.

Specifically, we first trained a BERT-based relation prediction model (similar to the knowledge selection model introduced in Sec. 5.3.2) to predict which relation is contained in  $\hat{Q}_i$ , and we denoted the predicted relation as  $\hat{r}_i$ . Then we measured *Truthfulness* of  $\hat{Q}_i$  as checking whether  $\hat{r}_i$  is connected to  $\hat{e}_i$  in the knowledge base **KG**:

$$\mathbf{TRUTH}(\hat{Q}_i) = \mathbb{1}[(\hat{e}_i, \hat{r}_i) \in \mathbf{KG}] \quad (5.7)$$

**Clarity.** The Gricean Maxim of Manner advocates that one should communicate in a clear and orderly manner, and therefore we require a follow-up question to avoid obscurity and ambiguity. We regarded *Clarity* of a generated follow-up question as how well it fits in natural language, and we used an external powerful language model as an expert to provide the measurement. Hence, the perplexity of a generated follow-up question is computed under a pre-trained language model and we pick DialoGPT because it's pre-trained on dialogue data. Specifically, we calculate it as follows:

$$\mathbf{CLA}(\hat{Q}_i) = \exp\left(-\frac{1}{N} \sum_j^N \log p_\theta(w_j | w_{<j})\right) \quad (5.8)$$

**Coherence.** A coherent conversation reflects another aspect of Manner as well. Following previous work [256], we evaluated the semantic *Coherence* of generated follow-up questions to dialogue histories by predicting the probability of each generated question conditioned on the previous QA pair using the powerful language model, BERT, whose pre-training tasks include this same objective.

$$\mathbf{COH}(\hat{Q}_i) = \mathbb{1}_{\text{BERT}}(\hat{Q}_i | C_i) \quad (5.9)$$

## 5.5 EXPERIMENTS

We first evaluated our model's accuracy in knowledge selection and then examined the quality of generated follow-up questions with selected knowledge.

### 5.5.1 Knowledge Selection

We first assessed the ability of our model to predict the salient knowledge chosen by human annotators given the dialogue history. The result can inform us of the feasibility of explicitly

Model	Entity Selection		Relation Selection		
	R@1		R@1	R@3	R@5
<b>Attention-based</b>					
KG Selector (w/ TransE)	0.642		0.271	0.593	0.761
(w/ TransD)	0.624		0.295	0.616	0.773
(w/ TransR)	0.628		0.288	0.612	0.773
<b>MLP-based</b>					
KG Selector (w/ TransE)	0.647		0.283	0.605	0.768
(w/ TransD)	0.648		0.288	0.613	0.763
(w/ TransR)	<b>0.654</b>		<b>0.302</b>	<b>0.620</b>	<b>0.776</b>

Table 5.1: Experimental results of various methods on the knowledge selection task. MLP-based models generally perform better than attention-based models for both selection tasks, and the MLP-based model with TransR reached the best performance with 0.654 recall@1 score on entity selection and 0.302 recall@1 score on relation selection.

Model	REL(%) $\uparrow$	INFO $\uparrow$	TRUTH(%) $\uparrow$	CLA $\downarrow$	COH(%) $\uparrow$
KG-FQG (w/o knowledge)	60.06	0.47	41.63	<b>2.46</b>	<b>99.07</b>
(w/ predicted knowledge)	<b>72.51</b>	0.55	<b>67.93</b>	2.92	98.55
(w/ gold-standard knowledge)	69.42	<b>0.63</b>	63.35	2.83	98.99

Table 5.2: Results of various methods on the follow-up question generation task.  $\uparrow$  indicates the higher score the better, while  $\downarrow$  means the lower score the better.

separating knowledge selection from the full generation task and choosing the best model to use in our two-phase framework. Meanwhile, separating this task could increase the interpretability of the whole framework by explicitly identifying the knowledge entity and relation that steer the follow-up question generation. We compared the two proposed baseline models, attention-based and MLP-based, and examined how well they interact with different types of knowledge embedding, including TransE [249], TransR [257] and TransD [258].

Table 5.1 shows the results of recall@k of the selection models. MLP-based models generally perform better than attention-based models for both selection tasks, and the MLP-based model with TransR reached the best performance with 0.654 recall@1 score on entity selection and 0.302 recall@1 score on relation selection. Based on this, we decided to use the best performing model, MLP-based with TransR, in the second stage.

### 5.5.2 Follow-up Question Generation With Knowledge

We evaluated our model on the follow-up question generation with selected knowledge in two settings: given the gold-standard knowledge chosen by human annotators, or the

Model	R-1	R-2	R-L
KG-FQG (w/o knowledge)	15.46	3.82	15.06
(w/ predicted knowledge)	19.66	7.12	19.32
(w/ gold-standard knowledge)	<b>34.98</b>	<b>16.58</b>	<b>34.00</b>

Table 5.3: ROUGE scores of various methods on the follow-up question generation task. Aligned with our Gricean Score, we showed that adding knowledge is an effective way to involve knowledge in follow-up question generation, as the two models with knowledge outperform their counterpart, DialoGPT without knowledge.

knowledge predicted by the best selection model. Additionally, we compared them with the baseline model DialoGPT without adding external knowledge to evaluate the effectiveness of explicitly employing knowledge.

We first used *Gricean Scores* to perform the evaluation and analyze the results using ANOVA tests with Tukey methods. As shown in Table 5.2, compared to the baseline model, our proposed knowledge-driven models significantly improved the quality of generated follow-up questions in terms of *Relevance* ( $F(2, 3009) = 19.41, p < 0.01^{**}$ ) and *Truthfulness* ( $F(2, 3009) = 85.67, p < 0.01^{**}$ ). This suggests that integrating selected knowledge is a valid approach to leverage knowledge to guide the generation. For *Clarity* we found the baseline model performed the best ( $F(2, 3009) = 15.59, p < 0.01^{**}$ ). As for *Informativeness* and *Cohherence*, three models all achieved good performance with insignificant difference, and we think it's due to the powerful backbone, DialoGPT, we chose. We found the difference between our model with predicted knowledge and gold-standard knowledge for all five dimensions is not statistically significant. This shows that even if the chosen knowledge by the knowledge selection model is different from the ones by human annotators, our generation model can generate appropriate follow-up questions, and demonstrated the applicability of our proposed method for the open-ended scenario. By examining the outputs, we noticed that including external knowledge through prompts can reduce hallucinations - the fabrication of untruthful information [259]. For example, when a dialogue history mentions "Oh, I loved Tom Arnold in *Undercover Blues*", the baseline model that does not involve knowledge generates "Who else starred in *Underpants Blues*", while our model is capable of generating the movie name correctly. This indicates that adding external knowledge through prompts can be a potential solution to entity-level hallucination in text generation.

We further adopt the common metric ROUGE [260] and report the results in Table 5.3. Our experiments again showed that adding knowledge is an effective way to involve knowledge in follow-up question generation, as the two models with knowledge outperform their counterpart, DialoGPT without knowledge. However, we can see there is still a gap between the model with predicted knowledge and gold-standard knowledge, and we think

Model	Quantity	Quality	Relation	Manner
DialoGPT	1.41	1.52	1.40	1.62
Ground Truth	1.82**	1.95**	1.82**	1.84
KG-FQG (w/ predicted knowledge)	1.74*	1.86**	1.66	1.84
(w/ gold-standard knowledge)	1.80**	1.74	1.77**	1.75

Table 5.4: Human Evaluation results of various methods on the follow-up question generation task on Likert Scales from 0-2. The \* indicates the difference against DialoGPT is statistically significant. \* $p < 0.05$ . \*\* $p < 0.01$ .

the reason comes from the different knowledge selections, which results in generating totally different follow-up questions. For example, asking about an actor and a movie can be different, but both valid follow-up questions. However, the only-one ground truth for one test instance limits the consideration of other possibilities, and this may underestimate models' performance in open-ended text generation. By comparing to the results above, this shows the superiority of our *Gricean Scores* which evaluates multiple aspects of generated text without relying on ground truth text. The reference free nature of the *Gricean Scores* opens its way to the evaluation of other text generation tasks.

### 5.5.3 Expert Evaluation

Complementing the result with our objective metrics, we randomly sampled 50 instances from the generated follow-up questions and ground truth texts to perform a expert evaluation. We first discussed a codebook based on four dimensions of Gricean Maxim on a 0 (poor) to 2 (excellent) point scale. Then two expert annotators individually rated instances blind to the generation methods. Krippendorff's alpha ranged from 0.80 to 0.92 for each set of coding. We averaged their scores (Table 5.4) and performed ANOVA tests with Tukey methods to analyze the results. Overall, the results aligned with our proposed *Gricean Scores*, where the knowledge-driven methods outperform the baseline model, DialoGPT, in all dimensions except *Manner*. The Tukey post-hoc tests showed the differences between Ground Truth and our knowledge-driven methods are not statistically significant in all four dimensions, indicating the effectiveness of our approach. In addition, our *Gricean Scores* correlates well with our expert evaluation, Quantity: 0.16, Quality = 0.37\*, Relation = 0.24\*, and Manner = 0.19<sup>20</sup>.

Dialogue History	<ul style="list-style-type: none"> <li>– Was Two by Two released in 2016?</li> <li>– It was, the same year <u>When I'm Gone</u> was released, which is another similar book.</li> </ul>
Predicted Entity: Predicted Relation: KG-FQG:	<i>When I'm Gone</i> <i>release_year</i> – When was "When I'm gone" released?
Dialogue History	<ul style="list-style-type: none"> <li>– Do you know <u>The Runaway Jury</u>?</li> <li>– The Runaway Jury is written by John Grisham, with a genre of Suspense.</li> </ul>
Predicted Entity: Predicted Relation: KG-FQG:	<i>The Runaway Jury</i> <i>subject</i> – What is the subject of the movie?

Table 5.5: Generated examples from KG-FQG with predicted knowledge for error analysis. The analysis shows two potential sources of errors, ignoring dialogue history and entity disambiguation.

## 5.6 ERROR ANALYSIS

Table 5.5 shows several examples that our model with predicted knowledge fails. In the first example, we can observe that our model picks an appropriate mentioned entity, the book *When I'm gone*, to formulate its follow-up question but asks the relation *release\_year* which has been covered in the dialogue history. This implies that our current model sometimes may not pay attention to the dialogue history when selecting a relation for asking. We believe this issue might be solved if we add additional constraints for relation prediction and only allow the knowledge selection model to select unmentioned relations.

Another kind of errors can be seen from the second example. The dialogue history mentions *The Runaway Jury* which is a novel, but our model misunderstands it as a movie even if the phrase "is written by" in the dialogue history indicates it is a book. We notice that such a problem usually happens when different entities share the same name, such as an adapted movie and its original book. Therefore, it is necessary to perform named entity disambiguation during the preprocessing, and exploring better knowledge embedding may be potentially useful in alleviating this problem as well.

---

<sup>20</sup>\* indicates the correlation is significant per Pearson's test

## 5.7 CONCLUSION

We propose a knowledge-driven framework to address three challenges in follow-up question generation in conversational surveys, diverse and complex responses interpretation, high-quality question construction, and question evaluation. Our framework first selects an entity-relation pair from dialogue histories as question topic and focus, and then uses it to guide a GPT-based model to generate high-quality follow-up questions. To verify the effectiveness of the proposed framework, we collect a new dataset and propose a new set of reference-free evaluation metrics, *Gricean Scores*. Extensive experimental results suggest that our framework outperforms competitive baseline models in both quantitative and qualitative experiments.

## CHAPTER 6: CONVERSATIONAL AI FOR ONLINE INFORMED CONSENT

### 6.1 OVERVIEW

In previous chapters, we designed and built conversational AIs for human understanding. When creating such tools to interact with people and collect information, especially sensitive information such as personal identity and health information, we need to consider the ethical considerations surrounding consent, security, and usage to ensure privacy and protect individual rights.

In this chapter, we studied ethical considerations around the online human subject study <sup>21</sup>. We examined the role of an AI-powered chatbot in improving the online informed consent process. As a core cornerstone of ethics in human subject research [10], informed consent is the process that guards prospective participants' voluntary and informed participation decisions. Through the informed consent process, the participant understands the purpose of the study, procedures to be carried out, potential risks and benefits of participation, the extent of data collection and confidentiality, and their rights. Despite its importance, studies found people often sign the form without a thorough read [38, 39, 43, 262, 263, 264].

Consent form reading research, from high-stake studies with severe ramifications to low-stake studies with minimal risk, shows that participants do not thoroughly read consent forms [38, 39, 43, 262, 263, 264]. The participants become less informed when the consent process moves online [41, 42, 43]. For example, Ghandour et al. [265] found that 65% people spent less than 30 seconds reading an online consent form with over 800 words, which should have taken roughly seven minutes. Pedersen et al. [43] showed that compared to the in-person informed consent process, participants' ability to recognize and recall information from the consent form dipped further when the same consent form was presented online. As a result, the current informed consent process may be neither informed nor consensual.

Uninformed participation decisions put both the participant and the researcher at risk. The consequence could endanger the participant's health if they failed to notice the study procedure could induce stress on their pre-existing mental or physical conditions [266]. It could also create privacy risks to participants' personal data if the participant holds incorrect assumptions of the researcher's data-sharing practice [40]. Further, a lack of a good understanding of the consent form may inhibit the participant's autonomy in making the consent decision. Cassileth et al. [267] found more than 1/4 of participants in a medical

---

<sup>21</sup>This Chapter has been published at ACM CHI Conference on Human Factors in Computing Systems (CHI 23) [261]

study thought accepting the consent form was the only option to receive treatment even though the form indicated alternatives. Meanwhile, an ineffective informed consent process could damage study validity and data quality [48, 268]. Failure to understand the study procedures and purposes may not only negatively impact performance on experimental tasks but also create confounding factors, especially for studies involving deception [48, 268]. Conversely, an effective informed consent process could improve participant engagement and promote trust and rapport between the participant and the researcher [47, 48, 49]. It could reduce the power asymmetry in participant-researcher relations by bridging the information gap of the study, informing participants of their rights, and guarding voluntary decisions [72]. Consequently, a successful informed consent process benefits both the participant and the researcher.

In recent years, especially during the COVID-19 pandemic, online informed consent has become more prevalent. More studies, both online and in-person, collected participants' consent remotely. And those studies may involve risky procedures or collect sensitive information, including people's mental health and personal genetic data. Therefore, improving online informed consent reading is becoming increasingly important.

However, improving consent-form reading is challenging, especially in an online environment [43, 44, 269]. Compared to an in-person setting where the researcher could directly interact with the participants, guide them through the consent form, and answer their questions, in an online environment, the absence of the researcher creates communication barriers and significantly demotivates and disincentivizes consent form reading [41, 48, 270]. In the past, researchers have experimented with different interventions to improve consent form reading, including simplifying the content, lowering reading grade level, and introducing interactive features [41, 46, 54]. However, two literature reviews of over 70 past studies suggest that the effect of those interventions was limited, and it is burdensome to design and develop compelling and effective interactive experiences [44, 45]. Therefore, exploring new techniques to improve online consent form reading is necessary.

We examined the role of an AI-powered chatbot <sup>22</sup> in the delivery of consent form content. We built the chatbot **Rumi** with a hybrid system that combines a rule-based system with AI components. **Rumi** can greet a participant, go through the consent form section by section, answer the participant's questions, and collect people's consent responses, similar to what an experienced researcher would do in a lab setting. Through a conversational interface, a chatbot could grab people's attention, deliver personalized experiences, and provide human-

---

<sup>22</sup>In contrast to traditional rule-based chatbots, we define an AI-powered chatbot as a chatbot that leverages artificial intelligence (AI) technologies, including machine learning, natural language processing, and advanced analytics in the building process or the deployment environment of the chatbot.

like interactions [173]. All those features could potentially benefit the online informed consent process. However, chatbots also bear several risks. First, a turn-by-turn chat requires extra time and effort to complete the informed consent process, which is a major challenge in consent form reading [44]. The risk is even higher for studies with paid participants, who would not be rewarded for taking longer to complete the study. Second, current chatbots are far from perfect. Their limited conversation capabilities may deliver incorrect answers or lead to user disappointment and frustration [14]. Therefore, it is yet unknown how a chatbot could affect the online informed consent process.

To explore the effectiveness of an AI-powered chatbot that guides a participant through an informed consent process, we asked three research questions (RQs),

**RQ6.1:** How would the participant’s consent form reading differ in a study with the AI-powered chatbot-driven consent process vs. the form-based consent process? (*Consent Form Reading*)

**RQ6.2:** How would the participant’s power relation with the researcher differ in a study with the AI-powered chatbot-driven consent process vs. the form-based consent process? (*Power Relationships*)

**RQ6.3:** How would response quality differ in a study with the AI-powered chatbot-driven consent process vs. the form-based consent process? (*Study Response Quality*)

To answer our research questions, we designed and conducted a between-subject study that compared the use of an AI-powered chatbot, **Rumi**, and a typical form-based informed consent process in an online survey study. Since no previous work has examined the use of chatbots that deliver informed consent, in this study, we focused on examining the holistic effect of a chatbot instead of the effect of individual chatbot features. With a detailed analysis of 238 study participants’ informed consent experiences and their responses to the survey study, we found 1) **Rumi** improved consent form reading, in terms of both recall and comprehension, 2) participants who completed the consent form with **Rumi** perceived a more equal power-relation with the researcher, 3) by reducing the power gap, the improved informed consent experience ultimately benefited study response quality.

This chapter provides three unique contributions.

- An understanding of the holistic effect of AI-powered chatbots conducting online informed consent. Our findings extend prior work on creating an effective informed consent process and reveal the practical value of using an AI-powered chatbot for informed consent, especially in improving consent form recall and understanding, researcher-participant relation, and study response quality. We further provided empirical evidence that could attribute observed improvement in study response quality to the reduced

power gap.

- Design implications of creating effective AI-powered chatbots for informed consent. We discussed design considerations, such as personalized reading assistance and power dynamics management, to further improve the online informed consent process.
- New opportunities for creating and operationalizing AI-enabled, consent experiences for a broader context. The demonstrated effectiveness of the chatbot-driven informed consent opens up opportunities for employing AI-powered chatbots for other types of consent in the age of AI.

## 6.2 METHOD

To answer our RQs, we designed a between-subject study that compared the outcomes of two methods to deliver the consent form online, an AI-powered chatbot (*Chatbot Condition*) and a typical form (*Form Condition*), on consent form reading, participant-researcher power relation, and study response quality.

### 6.2.1 Dummy Survey Study Design

To understand how an AI-powered chatbot facilitates online informed consent reading and how it might influence study response quality, we have to separate the study quality evaluation from the consent form reading evaluation. Therefore, we designed a dummy study dedicated to evaluating study response quality.

The dummy study is about problematic social media use. To complete the study, the participant first read a short article about problematic social media use. The goal is to familiarize participants with the issue. Then, the participant answered a survey with both close-ended questions and open-ended questions. The choice-based questions included six five-point Likert Scale questions adopted from Bergen Social Media Addiction Scale [271]. The open-ended questions are adopted from [272] with the goal of understanding people's attachment to social media and how it affects people's real life. Both question sets are widely used in social media research. We choose this topic for four reasons. First, it relates to most people online and is suitable to conduct online. Second, we could vary the level of psychological discomfort and data sensitivity to simulate a wider range of online studies. Third, the survey method is the most widely used research instrument, ensuring our finding's generalizability. Lastly, prior studies have provided us with established methods to robustly measure the response quality of a survey with both open-ended questions and choice-based

questions [173, 273]. We also considered a genetic study used by [61] that deals with genetic data. Although mishandled genetic data may cause more tangible consequences, the level of risks is difficult to vary and the generalizability is limited.

We designed three versions of our dummy study for three common risk levels in online survey studies,

- *Low* - Non-sensitive data without personal identifiers
- *Medium* - Sensitive data without personal identifiers
- *High* - Sensitive data with personal identifiers

For the *Low* risk version, the survey will ask about people's *opinions* regarding others' problematic social media use. This version is designed to evoke minimal psychological discomfort by asking for opinions about other people instead of directly recalling their own experiences. For the *Medium* risk version, the participant will answer questions regarding *their own* problematic social media use. And in the *High* risk version, we will ask participants to additionally reveal their social media handles as a personal identifier for a follow-up study. The distinction was made clear in the study description and potential risks in the consent form.

### 6.2.2 Consent Form Design: Form Condition

The consent form was based on the Social Behavioral Research Consent Form template provided by the Institutional Review Board (IRB) at the University of Illinois Urbana-Champaign. We improved the design of the form-based consent form based on recommendations from prior work [41, 53, 54, 55]. Specifically, we broke the consent form into sections to reduce information overload and used a clear and simple format to ensure clarity and readability.

### 6.2.3 Consent Form Design: Chatbot Condition

We created a chatbot, *Rumi*, with the goal of simulating an in-person informed consent process experience where a researcher goes through the consent form section by section, asks if the participant has questions, and makes clarifications. In our study, *Rumi* first greeted the participant and informed the participant that it could take questions at any time during the informed consent process. Then, *Rumi* went through the informed consent form section by section with the exact content in the Form Condition. Participants can click the "Next" button to proceed to the next section or type in the text box with their questions or other

requests. During the process, **Rumi** proactively asked if the participant had any questions twice, one after the risk section and one at the very end. Participants can skip by pressing the “No Questions” button. Then, **Rumi** confirmed the participant’s age and elicited for their consent to join the study. We included a video in the supplementary material to demonstrate how **Rumi** conducts the informed consent process.

We adopted a hybrid approach to build **Rumi** by combining a rule-based model with AI-powered modules [274]. We made this decision for the following reasons. First, rule-based models have limited capability to recognize participants’ questions and deliver diverse and engaging responses, let alone handle non-linear conversations (e.g., side-talk, request to repeat information, etc.). Second, although an end-to-end generative system could lead to a more engaging experience, it may produce incorrect or non-factual answers. Those answers may cause severe issues in high-stake scenarios [275, 276], e.g., delivering a consent form. Third, the consent form content is typically specific to the situation or context in which the form is being used. Such a few-shot or zero-shot setting poses another challenge to building an effective generative model, even fine-tuning pre-trained models (e.g., DialoGPT [277]). As a result, we decide to prioritize the answer authenticity and build **Rumi** on a hybrid system with AI-powered modules to enable **Rumi**’s ability to handle a broad set of questions and provide diverse and accurate answers.

Specifically, we built **Rumi** on the Juji platform, a hybrid chatbot building platform. Juji provides built-in AI-based functions for dialog management and effective Q&A. **Rumi** will follow a rule-based conversation agenda to go through the consent form section by section and acquire the participant’s consent. Given a question, Juji offers pre-trained Natural Language Understanding (NLU) models to identify relevant questions with known answers in a Q&A database and returns an answer or a follow-up question for clarification. When the chatbot is unsure about how to answer a question, it will recommend similar questions to give participants a chance to obtain desired answers and to learn more about the chatbot’s capabilities [278]. Juji also offers a diverse set of conversation skills such as handling side talks and conversation repair to provide an engaging conversation experience [173].

We curated the Q&A database by creating a set of seed questions ourselves and piloting **Rumi** with 54 online participants. Since the goal is to gauge potential questions participants may ask, we asked our participants to ask as many questions as possible. Researchers on the team wrote answers for each question based on the consent form and added Q&A pairs into the database. Before deploying **Rumi** for this study, the Q&A database contains over 200 Q&A pairs. To further enhance **Rumi** ability to recognize questions that may be differently phrased, we leveraged a text-generation model GPT-3 (text-davinci-002) [279] to create question paraphrases. Similarly, we used the GPT-3 to create a candidate answer set for each

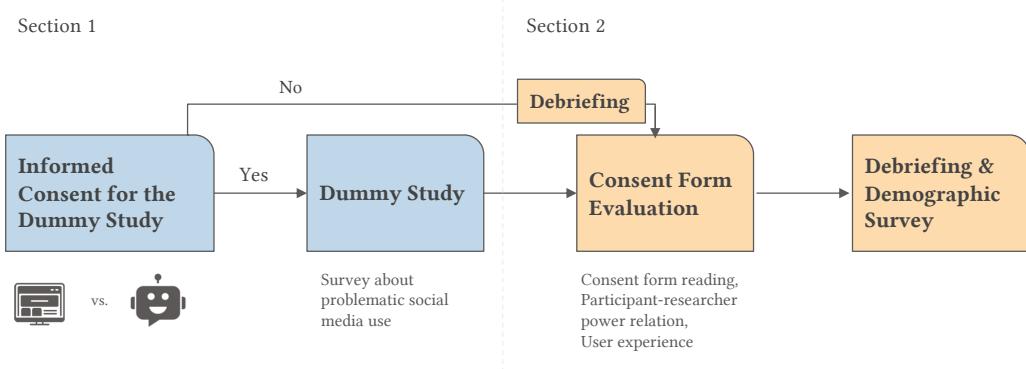


Figure 6.1: The figure shows the overall study procedure. In Section 1, based on the assigned condition, the participant interacted with Rumi or the Form for the informed consent process and then completed a dummy study about social media use. In Section 2, participants answered questions about their consent form reading experience. We debriefed our participants and collected demographic information at the end. If the participants decided not to join the study in Section 1, they would be invited to start Section 2.

question. We augmented the Q&A database with five question paraphrases and five candidate answers for each Q&A pair. We hand-checked all generated texts to ensure information authenticity.

#### 6.2.4 Study Procedure

The complete study consists of two sections (Fig 6.1) and is approved by the IRB. In the first section, participants were randomly assigned to one of the conditions (Chatbot Condition vs. Form Condition) and one of the risk levels (Low vs. Medium vs. High). The study started with the informed consent process based on the assigned condition. Upon consent, the participant completed the Dummy Survey Study.

Our consent form evaluation is in the second section. The participants answered questions about their understanding of the consent form, their perceived relationship with the researcher, the informed consent process experience, and demographic information. In the end, we debriefed the participants with the real purpose of the study and asked them to complete an additional consent form about sharing their answers to the second section of the study. Before leaving the study, we also asked participants to give open-ended comments on their informed consent.

If the participant decided not to join the study in the first section, they were debriefed on the real purpose of the study and asked if they were willing to join our second section to evaluate their consent form reading. If they agreed to join, we would direct them directly to

section two to complete the rest of the study.

### 6.2.5 Measures

**Consent Form Reading** We measure participants' consent form reading in two dimensions, *recall* and *comprehension*.

**Recall:** The ability to recall information from the consent form suggests that people pay attention to the consent form. Similar to [48], to assess participants' recall of the form, we inserted two random statements "*watch a video with orange cat*" and "*read study materials in a blue background*" into the middle of the study procedure section and the risks section of the consent form respectively. Participants were tested on their recall of the color words, "*orange*" and "*blue*". We decided to use recall of two random statements as our measure of consent form reading because it could sensitively measure thoroughness and making a correct guess is difficult [48]. Studies showed participants were more likely to read the procedure and risks sections carefully [39, 48], which provides us an opportunity to measure the lower bound of consent form reading. The participants were asked to select one or two color words from five common color words ("green", "red", "white", "orange" and "blue"). The participants would receive a score of 2 if both color words were selected correctly, a score of 1 if the participants selected only one option and the answer was correct, and a score of 0.5 if the participant selected two options but was partially correct. Otherwise, they got a 0.

**Comprehension:** Participants' comprehension of the consent form reflects the effectiveness of reading. To comprehend the consent form, the participants need to process the text, understand its meaning, and integrate it with their prior knowledge [280]. Inspired by [280], we measured comprehension with six questions that required participants to process the information presented in the consent form beyond simple recollection. These questions asked about the study procedure, potential risks, and actions to take if a certain scenario happens, e.g., how the participants could protect their privacy if a data breach happens or what a participant should do if they decide to withdraw from the study. A total of four multiple-choice questions were presented. The comprehension score measured how accurately a participant answered those questions. The final score is the percentage of correct answers, ranging from 0% to 100%.

**Participant-Researcher Power Relation** We measure the participant-researcher power relation by participants' perceived relationship with the researcher in the study and their

feeling of agency and control.

We measure two aspects of the perceived relationship, Partnership and Trust. Based on [281], we asked if the participants perceived their relationship with the researcher who ran the study as partners. Adopted from [61], we measured the trust by asking to what degree the participants trusted the researcher would handle their data properly.

According to [11, 66], the power asymmetry between research parties inhibits participants' autonomy. By power redistribution, we would expect participants to regain agency and control, two constructs of autonomy. We measured the perceived agency and control from both sense of positive agency and the sense of negative agency. We adapted scales from [282] based on the context of the informed consent process. All items adopted a 7-point Likert scale from Strongly Disagree to Strongly Agree.

**Study Response Quality** We measured the study response quality by examining participants' responses to both choice-based questions and open-ended questions in the problematic social media use survey.

**Non-differentiation:** Non-differentiation is a survey satisficing behavior where the respondents fail to differentiate between the items by giving nearly identical responses to all items using the same response scale [283]. Non-differentiation deteriorates both the reliability and validity of question responses. It further inflates intercorrelation among items within the questionnaire and suppresses differences between the items [284]. We used the mean root of pairs method [273] to measure the non-differentiation in choice-based questions. We calculated the mean of the root of the absolute differences between all pairs of items in a questionnaire. The metric ranged from 0 (The least non-differentiation) to 1 (The most non-differentiation).

**Response Quality Index (RQI):** To measure the response quality of open-ended questions, we created a *Response Quality Index* based on [173]. It measures the overall response quality of N responses given by a participant on three dimensions, relevance, specificity, and clarity, derived from Gricean Maxim [180]:

$$RQI = \sum_{n=1}^N \text{relevance}[i] \times \text{clarity}[i] \times \text{specificity}[i] \quad (6.1)$$

(N is the number of responses in a completed survey)

**Relevance.** A good response should be relevant to the context. For an open-ended question, a quality response should be relevant to the survey question. Irrelevant responses

not only provide no new information but also burden the analysis process. We manually assessed the relevance of each open-text response on three levels: 0 – Irrelevant, 1 – Somewhat Relevant, and 2 – Relevant.

**Specificity.** Quality communication is often rich in details. Specific responses provide sufficient details, which help information collectors better understand and utilize the responses and enable them to acquire more valuable, in-depth insights. We manually assessed the specificity of each open-text response on three levels: 0 – Generic description only, 1 – Specific concepts, and 2 – Specific concepts with detailed examples.

**Clarity.** Clarity is another important axis. Each text response should be easily understood by humans without ambiguity, regardless of its topical focus. We manually scored each free-text response on three levels: 0 – Illegible text, 1 – Incomplete sentences, and 2 – Clearly articulated response.

**Coding Process:** We coded a total of 1428 open-ended responses. Two researchers with a background in human-computer interaction and expertise in open-ended survey analysis conducted the coding process. They first randomly selected 10% of the responses and created a codebook on the above three dimensions with definitions and examples. Then, two researchers coded the rest of the data independently and were blind to the condition. Krippendorff’s alpha ranged from 0.83 to 0.98 for each set of coding. The final disagreement was resolved by discussion.

## Participant Experience and Demographics

**Time and Effort:** Injecting interactivity often means the participants need to spend extra time and effort to interact with the system, which is a major trade-off [44, 45]. Therefore, to measure the perceived time and effort of the informed consent process, we adapted the ASQ scale with two items [285] on how satisfied people were with the time and effort spent on the informed consent process, which we later averaged into a single score.

**Future Use:** People’s willingness to use the same system in the future is a strong indicator of good user experience and satisfaction. We used a single-item 7-point Likert scale to ask if the participant would use the chatbot or the form to complete an informed consent process in the future.

	Low	Medium	High	Total
Chatbot Condition	41	38	40	119
Form Condition	39	40	40	119
Total	80	78	80	238

Table 6.1: The table shows the participant distribution across conditions. Participants were randomly assigned to one condition based on the consent form conditions and the risk level. A total of 238 participants were included in the final analysis.

**Demographics:** Prior studies on chatbots suggests individual differences in people’s experience [173] may moderate their chatbot experience. We collected basic demographic information, including age, gender, education level, and annual household income.

#### 6.2.6 Participant Overview

We recruited fluent English speakers from the United States on Prolific <sup>23</sup>. Of the 278 participants who opened our link and started the informed consent process, 252 completed the informed form. Two participants in the Chatbot Condition explicitly declined to join the study and left the study immediately.

Out of the 250 participants who started the study, 238 (Denoted as P#) completed the study and passed our attention and duplication check. Our following analysis is based on those 238 valid responses (Table 6.1). Among those 238 participants, 97 identified as women, 136 identified as men, and 5 identified as non-binary or third gender. The median education level was a Bachelor’s degree. The median household income was between \$50,000 - \$ 100,000. And the median age of participants was between 25 - 34 years old. On average, our participants spent 1.24 mins ( $SD = 3.03$ ) completing the informed consent process in the Form Condition and 7.75 mins ( $SD = 7.06$ ) with Rumi. We compensated our participants at the rate of \$12/hr.

#### 6.2.7 Data Analysis

We used Bayesian analysis to compare the distributions of effects on consent form reading (RQ6.1), the participant’s power relation with the researcher (RQ6.2), and study response quality (RQ6.3) between two consent methods. We were motivated to use Bayesian analysis for the following reasons. First, Bayesian models allow us to foreground all aspects of the model; No modeling assumptions need checking that are not already foregrounded in the model description. Second, compared to the null hypothesis significance testing (NHST),

---

<sup>23</sup>[www.prolific.co](http://www.prolific.co)

the Bayesian analysis focuses on “how strong the effect size is” instead of “if there is an effect”. It better fits the exploratory nature of our study. Third, Bayesian models facilitate the accumulation of knowledge within the research community as study outcomes can be used as informative priors later. Kay et al. provide a detailed review of the Bayesian method’s advantages in HCI research [286].

We formulated a hierarchical Bayesian model for each outcome measure. We build two types of hierarchical Bayesian models, linear regression models for continuous measures and ordinal logistic regression for ordinal measures. For Recall, Comprehension, Agency and Control <sup>24</sup>, Non-differentiation, and RQI, we modeled the data as a Normal distribution and used linear regression models to estimate the Normal distribution means for both the Chatbot Condition and the Form Condition. By contrasting the posterior distributions of the means for the two conditions, we would know how the consent method affects outcome variables. Furthermore, we estimated the effect size of the difference of the posterior distribution with Cohen’s d for Bayesian linear regression models.

For Partnership, Trust, Time and Effort, and Future Use, we used ordinal logistic regression models to estimate the posterior distributions of the cumulative odds for a given value on the ordinal scale. Our Bayesian analysis goal was to compare whether the rating distribution was significantly different between conditions with respect to the neutral midpoint of the scale (*Neither agree nor disagree*). This would tell us whether participants were more likely to disagree with the statement in one condition over the other. We constructed the distribution of the difference between the cumulative odds of a rating of 4 or below (the midpoint of a 7-point Likert scale) between the Chatbot Condition and the Form Condition. Negative values indicate that participants in one condition had less odds of providing a neutral or negative response to the other condition. We used this for all the above measures except Recall where we estimated the posterior distributions of the cumulative odds of getting the recall question 50% correct or lower. Based on the posterior distributions, we calculated the Odds Ratio (OR) as the effect size <sup>25</sup>.

In all models, we controlled for the following covariates: study risk levels, participants’ age, gender, education level, and annual household income. We controlled for these demographics as prior studies on conversational agents suggest individual differences may affect their interaction with a conversational agent [173]. Full mathematical descriptions of each type of model are provided in the Supplementary Material. We performed the Bayesian analysis using

---

<sup>24</sup>We modeled the perceived Agency and Control as a continuous variable as it is a composite score from an 11-item scale.

<sup>25</sup>We interpreted the magnitudes of odds ratios based on Chen et al. [287] where OR = 1.68, 3.47, and 6.71 are equivalent to Cohen’s d = 0.2 (small), 0.5 (medium), and 0.8 (large), respectively

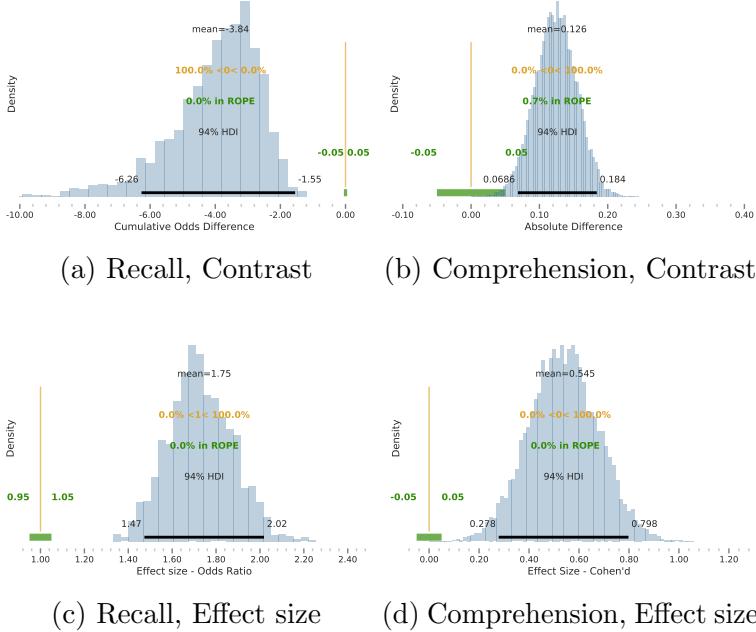


Figure 6.2: The first row represents the posterior distribution contrasts between the Chatbot Condition and the Form Condition for the cumulative odds of achieving less than 50% accuracy in the Recall task and for the means of Comprehension score. The second row shows the effect size distribution, Odds Ratio for Recall and Cohen’s d for Comprehension. Plot (a), (b), and (d) show an orange vertical line located at 0 with green bars indicating ROPE. Similarly, plot (c) has an orange vertical line located at 1 and a green ROPE interval. Effect distribution falling into these ROPE regions suggests no difference between conditions or no effect. Note that the x-axis is not the same scale for all plots. **Main finding:** Compared to the participants in the Form condition, participants in the Chatbot Condition provided more correct answers in the recall task and achieved a higher comprehension score. The differences are both statistically significant.

NumPyro<sup>26</sup>, a popular Bayesian inference framework. We used Markov Chain Monte Carlo (MCMC), a stochastic sampling technique to sample the posterior distribution  $P(\theta|D)$ , the distribution functions of the parameters in the likelihood function given the data observations  $D$ . In particular, we used the No-U Turn Sampler (NUTS) for sampling.

We supplemented our quantitative analysis with qualitative evidence by analyzing participants’ chat transcripts in the Chatbot Condition. We performed the thematic analysis [288] on the questions people asked. A member of the research team first performed open coding on the data and then refined these codes in an iterative and reflexive process. The same person then used axial coding to group these codes into larger categories to extract common themes.

---

<sup>26</sup>[num.pyro.ai/](http://num.pyro.ai/)

## 6.3 RESULTS

Overall, **Rumi** improved participants' consent form reading. Our participants who interacted with **Rumi** could recall more information from the consent form and take more correct actions based on the consent form compared to those in the traditional form-based informed consent process. We also found that in the Chatbot condition, participants perceived themselves as having a more equal power relation with the researcher and offered higher-quality responses in the Dummy Survey Study. Our exploratory path analysis revealed a potential mechanism where the chatbot-based consent method improves response quality by reducing the power gap.

### 6.3.1 People Engaged with Rumi and were Satisfied with the Experience

Although our participants spent more time chatting with **Rumi** to complete the informed consent process, they were, in general, satisfied with the Time and Effort (Chatbot:  $M = 4.72$ ,  $SD = 1.54$ ; Form:  $M = 4.48$ ,  $SD = 1.57$ ) used in completing the informed consent process. They also indicated that they are willing to use such a chatbot for future informed consent experiences (Chatbot:  $M = 4.21$ ,  $SD = 1.89$ ; Form:  $M = 4.03$ ,  $SD = 1.84$ ). We modeled both measures as ordinal variables and contrasted the posterior distributions of the cumulative likelihoods of a rating of 4 for both conditions. We found that participants' perceived time and effort and their indicated future use did not differ significantly between the two conditions. The High-Posterior Density Interval (HPDI)<sup>27</sup> for the cumulative odds difference overlapped with the ROPE (Region of Practical Equivalence)<sup>28</sup> of  $0 \pm 0.05$  in all cases (Time and Effort:  $M = 0.05$ , 94% HPDI: [0.01, 0.08]; Future Use:  $M = -0.22$ , 94% HPDI: [-0.49, 0.01]), indicating that participants were not significantly more likely to disagree or agree in one condition over the other for both Time and Effort and Future Use.

We dug into the participants' transcripts to further understand participants' interaction patterns. We found our participants engaged actively with **Rumi**. Our participants raised

---

<sup>27</sup>The HPDI is the location of 94% of the posterior density. It is similar to, but different from, the idea of the confidence interval used in non-Bayesian Statistics. In non-Bayesian Statistics, a 94% confidence interval is informally interpreted as “with 94% probability the parameter of interest lies in a specific interval; the tails are of equal width (i.e., 3%)”; the HPDI is the *densest* interval covering 94% of the posterior. The HPDI is guaranteed to include the most likely value, but this is not always true for confidence intervals; see McElreath [289]. For a more careful definition of the confidence interval, see Hoekstra et al. [290].

<sup>28</sup>Unlike non-Bayesian Statistics, where one can ask, if the two means for two conditions are different  $P(\mu_1 \neq \mu_2)$ , in Bayesian statistics, one asks if the HPDI of the distribution  $P(\mu_1 - \mu_2)$ , the distribution of the difference of the means of the two conditions, excludes an interval where we can consider the two treatments equivalent. This equivalence interval is domain-dependent. A posterior distribution HPDI that lies outside the ROPE is considered a significant result in Bayesian data analysis.

a total of 449 questions ( $M = 3.77$ ,  $SD = 2.56$ ), and **Rumi** answered 389 (85.97%) of them. We identified four major categories of questions, **Rumi**'s capability(12.69%), research team information (11.58%), study information (56.15%), and side-talking (19.59%). Our participants ask about what **Rumi** can do and what questions could **Rumi** answer (e.g., “*What do you know?*”[P107]). Another type of question is about the research team (e.g., “*Who is [Researcher Name]?*”[P57]). Through those questions, the participants could learn more about the research team to start rapport building. Unsurprisingly, people asked the most questions regarding the study itself. Specifically, our participants asked questions about the study procedure (45.23%; e.g., “*What do I do after the survey?*”[P41]), risks (28.17%; e.g., “*Will my information be safe?*”[P83]), compensation details (12.30%; e.g., “*What will I get after this?*”[P94]), and general information about the study (14.30%; e.g., study purpose, survey topic, etc.). Interestingly, some participants started side-talking with **Rumi**, such as “*How's your day?*”[P43] or “*I didn't sleep well yesterday. Do you sleep?*”[P67], which suggests an even higher engagement. Those questions indicate our participants were willing to spend the effort interacting with **Rumi**. And **Rumi** helped our participant to clarify important information regarding the consent form.

Prior work suggests that introducing interactivity often creates user burdens which may deter user experience, especially for consent form reading where the required time and effort is one major roadblock [44, 45]. Our results indicate that people are willing to actively engage with **Rumi** to complete the informed consent process, even though it took a longer time than normal online informed consent. We believe scale and speed should not be the only value in the informed consent process. Sacrificing speed and scale for a more engaging experience and grabbing people's attention, especially in this high-stake scenario, is important to consider.

### 6.3.2 Rumi Led to Better Consent Form Reading

Overall, similar to prior work [43, 48, 54], the consent form reading is poor. In terms of recall, only 26 participants in both conditions in total recalled both “orange” and “blue” correctly from the color phrase list. On average, people scored 0.63 out of 2 ( $SD = 0.64$ ) on the recall task and scored 53.22% ( $SD = 24\%$ ) on the comprehension task.

Compared to the static form, going through the consent form with **Rumi** leads to significant improvement in consent form reading. Participants who interacted with **Rumi** recalled more correct color words from the consent form (Chatbot:  $M = 0.76$ ,  $SD = 0.71$ ; Form:  $M = 0.51$ ,  $SD = 0.54$ ). We modeled Recall as an ordinal variable and estimated the posterior distributions of the cumulative likelihoods of getting 50% of the total score (1 out of 2). The HPDI for the cumulative odds difference excluded the ROPE of  $0 \pm 0.05$  ( $M = -3.84$ ,

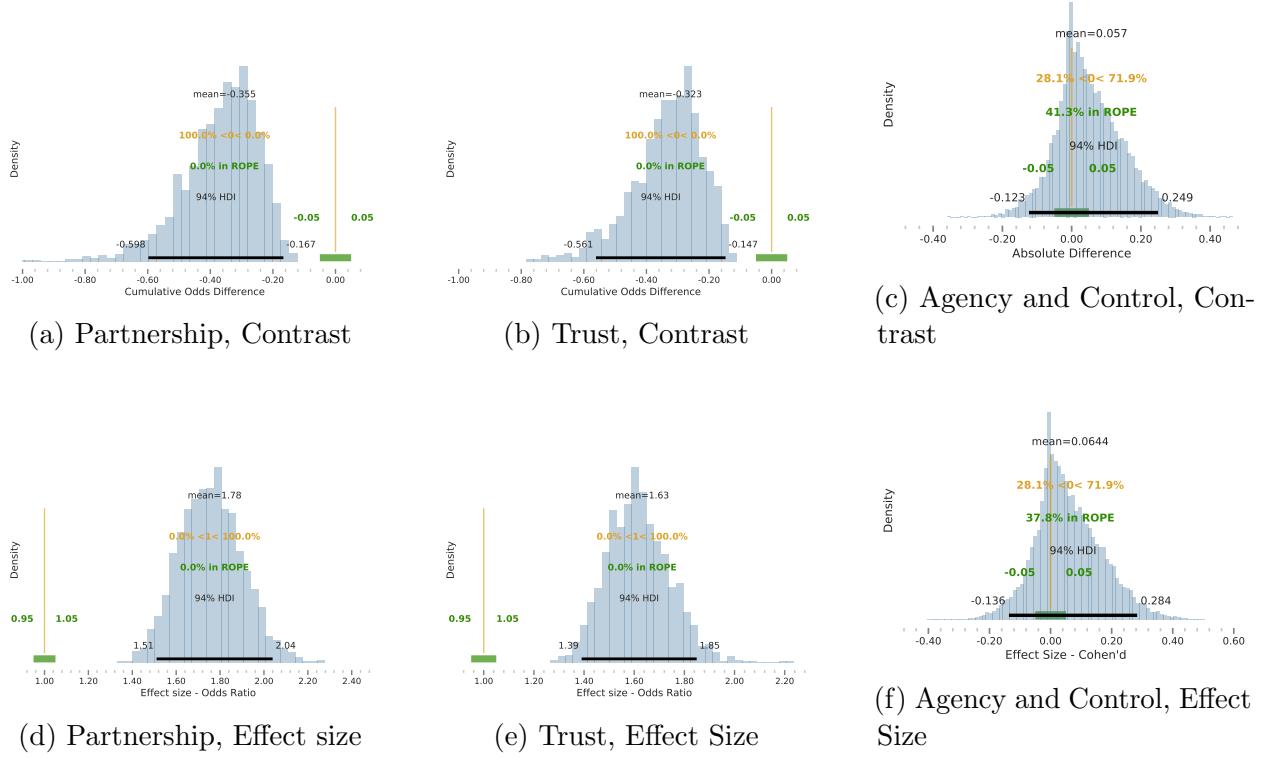


Figure 6.3: The first row represents the posterior distribution contrasts between the Chatbot Condition and the Form Condition for the cumulative odds of observing a neutral or lower rating in Partnership and Trust and for the means of Agency and Control. The second row shows the effect size distribution, Odds Ratio for Partnership and Trust and Cohen's d for Agency and Control. Plots (a), (b), (c), (f) show an orange vertical line located at 0 with green bars indicating ROPE. Similarly, plots (d) and (e) have an orange vertical line located at 1 and a green ROPE interval. Effect distribution falling into these ROPE regions suggests no difference between conditions or no effect. Note that the x-axis is not the same scale for all plots. **Main finding:** Compared to the participants in the Form Condition, participants who interacted with Rumi perceived their relationship with the researcher more like a partnership and trusted the researcher more. The difference is significant. Although participants in the Chatbot Condition reported higher perceived agency and control, the difference is not statistically significant.

94% HPDI: [-6.26, -1.55]). The estimated effect size (Odds Ratio) is small ( $M = 1.75$ , 94% HPDI: [1.47, 2.02], excluding the ROPE of  $0 \pm 0.05$ ; See Fig 6.2). The result indicates our participants could recall better with Rumi.

Participants in the Chatbot Condition also better comprehended the consent form (Chatbot:  $M = 61\%$ ,  $SD = 22\%$ ; Form:  $M = 46\%$ ,  $SD = 23\%$ ). We estimated the posterior distributions of the mean for the Chatbot Condition and the Form Condition. We found the difference between the Chatbot Condition ( $M = 51\%$ , 94% HPDI: [39%, 62%]) and the Form Condition ( $M = 51\%$ , 94% HPDI: [27%, 48%]) is statistically significant (excluding ROPE  $0 \pm 0.05$ ), with a medium to large effect size (Cohen's d) ( $M = 0.55$ , 94% HPDI: [0.28, 0.80], excluding

a ROPE of  $0 \pm 0.05$ ; See Fig 6.2). This suggests that the participants understood the content and could take better actions according to the consent form to protect their rights.

The results answer RQ6.1 clearly. The chatbot-driven informed consent process improves consent form reading in terms of both recalling information from the consent form and comprehending its content. Two factors may play a role in the observed improvement. First, the participant may be better engaged. The interactive features of **Rumi** were designed to simulate an in-person experience in which the research assistant actively engaged with the participant. Many of our study participants enjoyed this human-like experience and commented “*the bot is very friendly.*”[P61] and “*i liked how the bot talks to me.*”[P58]. And participants appreciated that **Rumi** went through the consent form with them, “*It was easier and nicer to read the consent form with the bot using texts other than a wall of text. Thank you :D*”[P15]. As we showed in Sec. 6.3.1, in the Chatbot Condition, the participant spent significantly more time during the informed consent process. Although time spent may not always lead to better reading [45], it could suggest higher engagement which plays a key role in reading comprehension [280]. Second, **Rumi** is designed to answer people’s questions. A total of 449 questions were raised and **Rumi** answered 85.97% of them. Our participants appreciated **Rumi**’s ability to answer their questions in real-time, “*It’s pretty cool that the chatbot can answer my questions right away*”[P32] but some participants mentioned that **Rumi** cannot fully understand their questions. Although the chatbot has limited capability, the ability to answer people’s questions on the fly might contribute to the improved consent form reading.

### 6.3.3 Rumi Aided Participant-researcher Relationship

The participants who interacted with **Rumi** perceived themselves as having a more equal power relation with the researcher in charge of the study. Our results indicated that people in the Chatbot Condition trust the researcher more (Chatbot:  $M = 5.60$ ,  $SD = 1.40$ ; Form:  $M = 4.92$ ,  $SD = 1.93$ ) and believed their relationship with the researcher is more like a partnership (Partnership) compared to the Form Condition, which indicates a smaller power gap (Chatbot:  $M = 4.26$ ,  $SD = 1.70$ ; Form:  $M = 3.63$ ,  $SD = 1.74$ ). Similar to Sec 6.3.1, we treated both variables as ordinal and estimated the posterior distributions of the cumulative likelihoods. Again, we constructed the distribution of the difference between the cumulative odds of a rating of 4 (the midpoint of the 7-point Likert scale) in the Chatbot condition and the cumulative odds of the rating in the Form condition. Based on the estimated posterior distributions, we found the differences of the cumulative likelihoods in both measures are statistically significant (Partnership:  $M = -0.36$ , 94% HPDI: [-0.60, -0.17], excluding the

ROPE of  $0 \pm 0.05$ ; Trust:  $M = -0.67$ , 94% HPDI: [-1.14, -0.28], excluding the ROPE of  $0 \pm 0.05$ ). The odd ratio suggests a small effect size for both measures (Partnership:  $M = 1.78$ , 94% HPDI: [1.51, 2.04], excluding the ROPE of  $0 \pm 0.05$ ; Trust:  $M = 1.63$ , 94% HPDI: [1.39, 1.85], excluding the ROPE of  $0 \pm 0.05$ ).

However, we did not observe a significant difference in participants' feelings of agency and control after the informed consent process. The participant who interacted with **Rumi** ( $M = 5.15$ ;  $SD = 0.78$ ) reports a higher feeling of agency and control, measured by a composite score of positive and negative agency, compared to the Form Condition ( $M = 5.04$ ;  $SD = 0.91$ ). Since we measured participants' feelings of agency and control with a composite score of an 11-item scale, we treated the score as a continuous variable and contrasted the mean between the Chatbot Condition and the Form Condition. We estimated the posterior distribution of the mean difference and found the difference is not statistically significant ( $M = 0.06$ , 94% HPDI: [-0.12, 0.25], overlapping the ROPE of  $0 \pm 0.05$ ).

Going through the online informed consent process with **Rumi** increases the trust between the participant and researcher and closes the power gap. Two potential factors may explain the observed effect. First, as mentioned in [72], a more effective consent form reading could reduce the power gap by bridging the information gap and assuring a voluntary decision. We did observe a significant correlation between participants' consent form reading and their power relation with the researcher (Recall:  $r(236) = 0.14$ ,  $p = 0.03$ ; Comprehension:  $r(236) = 0.21$ ,  $p < 0.01$ ). The observed difference in the power relation could potentially be attributed to more effective communication. Secondly, the humanness of **Rumi**'s design may help with rapport building between the researcher and the participants, which potentially reduces the power gap [66]. In our study, **Rumi** is framed as a virtual research assistant and represents the research team. As the first interaction between the researcher and the participant, the informed consent process may also serve the role of rapport building beyond informing the study participant.

#### 6.3.4 Rumi Led to Higher Survey Response Quality

Participants who interacted with the chatbot provided higher-quality responses to the dummy survey. For choice-based questions, participants in the Chatbot Condition exhibit less survey satisficing behavior (Chatbot:  $M = 0.42$ ,  $SD = 0.22$ ; Form:  $M = 0.47$ ,  $SD = 0.25$ ). The posterior distributions show the observed difference is not statistically significant ( $M = -0.03$ , 94% HPDI: [-0.08, 0.03], overlapping a ROPE of  $0 \pm 0.05$ ).

For open-ended questions, participants in the Chatbot Condition provided higher-quality responses (Chatbot:  $M = 5.38$ ,  $SD = 2.29$ ; Form:  $M = 4.17$ ,  $SD = 2.43$ ). The posterior

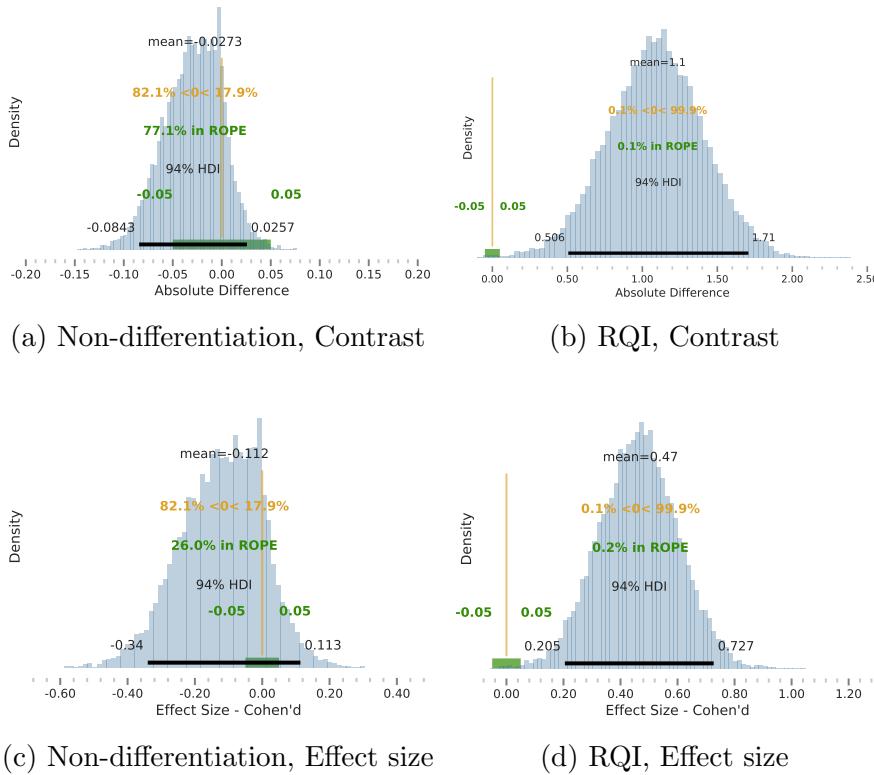


Figure 6.4: The first row represents the posterior distribution contrast of two means between the Chatbot Condition and the Form Condition for the Non-differentiation and Response Quality Index (RQI). The second row shows the effect size distribution, Cohen’s  $d$ , for both measures based on the posterior distribution. Each plot shows an orange vertical line located at 0 with green bars indicating ROPE. This represents that there was no difference between conditions or no effect. Note that the x-axis is not the same scale for all plots. **Main finding:** Participants who interacted with Rumi displayed less non-differentiation in close-ended questions but the difference is not statistically significant. We also observed participants in the Chatbot Condition contribute higher quality answers to the open-ended questions. The posterior distributions indicate the difference is statistically significant with a medium effect size.

distribution on the difference of RQI between two conditions shows a mean contrast of 1.1 with HPDI of [0.50, 1.71]. Since the 94% HPDI lies outside a significant ROPE of  $0 \pm 0.05$ , the result implies a significant effect with a medium effect (Cohen’s  $d$ :  $M = 0.47$ , 94% HPDI: [0.21, 0.73], excluding the ROPE of  $0 \pm 0.05$ ).

The results answered our RQ6.3. The participants who interacted with the chatbot during the informed consent process contributed significantly higher-quality responses to the open-ended questions. Participants who are in the High-risk condition were also willing to elaborate on their answers. Based on existing work, the observed effect could be due to a more effective consent form reading [48, 266] and a reduced power gap in the researcher-participant relationship [11, 66]. We further examined two potential pathways in Sec. 6.3.5.

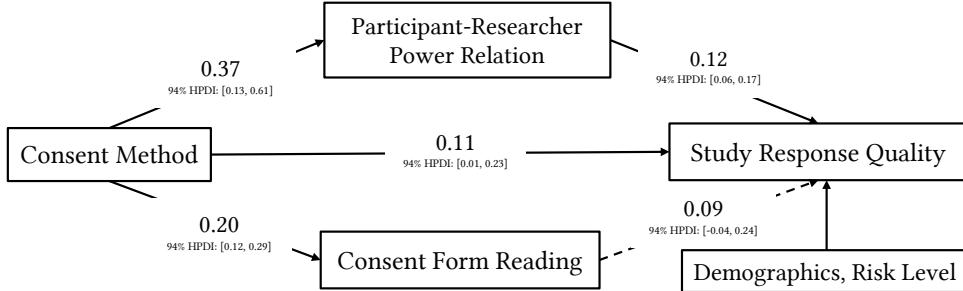


Figure 6.5: The figure shows a Bayesian Structural Equation Model (SEM) model to understand why chatbot-driven informed consent leads to improved study quality. As suggested by previous literature, we examined two potential pathways, 1) reduced power gap in the researcher-participant relationship, and 2) improved consent form reading. The results indicate a significant partial mediation effect from the reduced power gap.

### 6.3.5 Reduced Power Gap may Explain the Study Response Quality Improvement

Existing studies suggest that the power gap between the researcher and the participant plays a key role in the participant’s engagement and level of self-disclosure [11, 66]. A more equal power dynamic promotes trust and enhances the participant’s feeling of agency and control, potentially benefiting the study response quality. On the other hand, better consent form reading may also improve study response quality as the participant understands the study procedure better [48, 266]. Therefore, we explore two potential pathways that may mediate the effect of chatbot-driven informed consent on study response quality: through reduced power gap and through improved consent form reading.

We built a Structural Equation Model (SEM) model to answer our question. As shown in Fig 6.5, the Consent Method (Chatbot vs. Form) was set to predict the Researcher-Participant Power Relation and Consent Form Reading. The Researcher-Participant Power Relation is a composite score of the Researcher-participant relationship and the participant’s perceived agency and control. The Consent Form Reading combines recall and comprehension into one single measure with normalization. The Study Response Quality combines Non-differentiation for choice-based questions and Study Response Quality Index for open-ended questions. Both Researcher-Participant Power Relation and Consent Form Reading were set to predict the Study Response Quality, which is also predicted by the Consent Method. We believe that the Researcher-Participant Power Relation predicts Study Response Quality because the power dynamics influence trust and autonomy, two key factors in self-disclosure [11, 66]. We believe that the Consent Form Reading predicts Study Response Quality by how much the participant understands and follows the study instruction. Whether the Consent Method predicts Study Response Quality is our main question; the other two pathways are paths that mediate the

effect of the Consent Method on Response Quality. Like all other models, we controlled participants' demographics and risk levels. All variables were treated as manifest variables and modeled as Normal Distributions. We used Bayesian inference to fit the proposed SEM model with the No-U-Turn Sampler<sup>29</sup>. The model fits perfectly, with posterior predictive p-value = 0.48; BRMSEA = 0.01, 94% HPDI: [0.00, 0.06]; BMc = 0.99, 94% HPDI: [0.98, 1.00]; adjB $\hat{\Gamma}$  = 0.99, 94% HPDI: [0.96, 1.00].

Results indicated a significant total effect of the Consent Method on the Study Response Quality ( $\beta = 0.18$ , 94% HPDI: [0.08, 0.29], excluding the ROPE of  $0 \pm 0.05$ ). Notably, as shown in Fig 6.5, there was an indirect path from the Consent Method to the Study Response Quality via Participant-Researcher Power Relation ( $\beta = 0.12$ , 94% HPDI: [0.06, 0.17], excluding the ROPE of  $0 \pm 0.05$ ). This path reduces the total effect by 39%. However, the direct effect remained significant, reflecting only partial mediation. Meanwhile, we did not observe a significant path from the Consent Method to the Study Response Quality via the Consent Form Reading ( $\beta = 0.09$ , 94% HPDI: [-0.04, 0.24], overlapping the ROPE of  $0 \pm 0.05$ ).

The Bayesian SEM model shows that the potential reason that chatbot-driven informed consent improved later study response quality may be the reduced power asymmetry between the researcher and the participant. This finding aligns with prior studies that as one of the early interactions between the participant and the researcher, the informed consent process could bridge the information gap and shape the researcher-participant power relation, which may ultimately benefit study response quality and data richness [11, 66, 72]. However, the model does not indicate a potential pathway from the improved consent form reading. We believe there are two potential explanations. First, our study procedure is straightforward. An improved consent form reading may not have a strong effect in preparing the participants for the later study. Second, our study is low-stake compared to medical trials. Existing studies' findings on clinical trials may not generalize [266].

## 6.4 DISCUSSION

In this study, we found that a chatbot-driven informed consent process could effectively improve consent form reading, reduce the power gap, and ultimately benefit study response quality. In this section, we discuss design implications for a more effective informed consent experience with conversational AI.

---

<sup>29</sup>[cran.r-project.org/web/packages/blavaan/](http://cran.r-project.org/web/packages/blavaan/)

#### 6.4.1 Personalized Chatbot-driven Informed Consent Experience

In a turn-by-turn conversation, a chatbot can ask questions about participants' experiences and preferences to initiate personalization [19]. In our study, we found our participants enjoyed the feeling of a personalized experience. For example, one participant commented, “*Never seen a bot like this. I like the feeling Rumi is talking to ME.*” [P11]. Given this opportunity, we should consider incorporating personalization into the chatbot-driven informed consent experience. However, we need to be extremely cautious about unwanted persuasive effects [291] that violate the principle of voluntary participation when introducing personalization into online informed consent. The participation decision should be fully voluntary. The goal of personalization should be to facilitate the understanding of the consent form, not to *nudge* the participant to participate.

An AI-powered chatbot could highlight important content based on the participant's previous experiences with consent forms. On the one hand, existing studies suggest that more experienced study participants spent less time reading the consent form and often missed important information [292] because they tended to assume all consent forms were similar. Such incorrect assumptions exposed participants to unwanted risks. In this case, their pre-existing experience hindered consent form reading. On the other hand, regulations, such as General Data Protection Regulation (GDPR), often require a consent form to contain information that a participant may already be familiar with. In this case, a chatbot could summarize those materials to help participants better allocate their attention to new content. Therefore, it will be useful for the chatbot to learn about participants' prior informed consent experiences, analyze the new consent form, summarize contents a participant may be familiar with, and highlight new content to ensure a thorough read. An interesting idea to explore is a centralized chatbot that helps participants to manage all consent forms while preserving anonymity from prior study experiences. Although the IRB requires the researcher to draft the consent form in layman's language, some study procedures are indeed complicated. And sometimes certain terminology is necessary for clear communication, especially for high-stake clinical trials. Participants in our study appreciated *Rumi*'s ability to offer clarifications and answer their questions. Therefore, the chatbot should include interactive features to help participants understand the consent form. We could also borrow a thread of chatbot research focused on education where the chatbot helps people study new content and review materials [293, 294]. For example, a chatbot could first assess the participant's existing knowledge about the study topic to determine the necessary explanations and ask questions at the end to ensure learning outcomes.

However, as time and effort are among the biggest hurdles in consent form reading, the

trade-off between the benefits and risks of the interactive features needs proper calibration and consideration. Specifically, we need to consider the time cost a participant will spend on consent form reading. A participant’s compensation is often associated with time spent in a study procedure. If the total compensation is fixed, the interactive informed consent procedure will reduce the pay rate. In our study, we compensated participants in the Chatbot condition with an extra bonus at the end to ensure the promised pay rate. Although the research team may need to budget more for each participant, we argue, as the study shows, the improved response quality will ultimately benefit the research results. It saves time for researchers to clean up low-quality data from participants who did not read the study procedure carefully, especially in cases where unattended participants create confounding factors that endanger the study quality [48].

#### 6.4.2 Managing Power Dynamics

Our results echo prior studies on the role of an effective informed consent process for closing the power gap in the researcher-participant power relation and its benefit in study response quality. We believe that the chatbot could influence researcher-participant relations by adjusting its own power relation with the participant. We could further extend the utility by designing an informed consent chatbot that actively manages the power dynamic.

From a power relation point of view, one’s identity has a strong influence on their relative power over their counterpart [65]. Cassell mentioned that a researcher’s chosen identity, e.g., Interviewer, Facilitator, Initiator, Researcher, could change the power dynamic with the study participant [11]. We could carefully design an informed consent chatbot’s identity to suit various contexts. For example, to reduce the power gap, we could design the chatbot as a research partner rather than as a researcher. However, designing a virtual agent’s identity is complicated. Many design dimensions, including appearance, language style, etc., need to be carefully considered. Any incoherence may mar the entire experience.

The informed consent process could be considered as a negotiation about information disclosure between the researcher and the participant [66]. The researcher holds the information about the study, and the participants gain the knowledge and experience needed to perform the study. Karnieli-Miller et al. pointed out that such negotiation has the potential to change power relations by giving participants more information [66]. Thus, we should prepare an informed consent chatbot with conversation skills for such negotiation, so that the chatbot could understand participants’ requests, clarify their information needs, and actively manage information disclosure about the study.

We could further empower the participant by considering the ownership of the informed

consent chatbot. In this study, the chatbot acted on the researcher’s behalf and as a part of the research team. Although in most chatbot use cases the chatbot is owned by the creator instead of the user, a participant-owned informed consent chatbot may provide several benefits. First, the participant will have total authority over the conversation history. In this case, the participant could have a safe space to ask questions without feeling judged. Second, such an informed consent chatbot could become the central hub for all informed consent needs. It could act on behalf of the participant, analyze consent forms based on the participant’s preferences, and proactively ask the researcher questions to satisfy the participant’s information needs.

#### 6.4.3 Combining Human Expertise with LLMs

Many of our participants liked *Rumi*’s ability to respond to their questions in real-time with answers grounded in the consent form content. However, creating a chatbot that can accurately answer people’s questions, especially in high-stake contexts, is challenging. Due to limited natural language understanding ability, the current Q&A functionality for most commercial chatbot building platforms relies on a database of handcrafted Q&A pairs. It is especially time-consuming in the informed consent context as participants’ questions are specific to the consent form, further limiting the reusability of a Q&A database. Although some questions could be reused, for example, an institution may share the same template, and some study procedures could be similar, future studies are needed to design tools to support such a sharing practice.

Large language models (LLMs), like GPT-3, show promise in a new way to build conversational agents to answer people’s natural language questions. Although one could use off-the-shelf LLMs with in-context learning to build a chatbot to answer a wide range of domain-specific questions, LLMs sometimes generate non-factual information and have limited capability to memorize a long document [21]. Both shortcomings should be avoided in high-stake contexts, e.g., delivering consent forms, as non-factual information could mislead a participant to make an uninformed decision. It is not only an ethical concern but also could lead to severe consequences. For example, a participant who agrees to join a study without full knowledge of the specific study procedures may experience unexpected extreme physical or mental stress.

Therefore, we should consider leveraging LLMs carefully with the above shortcomings in mind. One framework to consider is combining LLMs with human expertise. In our study, we used GPT-3 to augment Q&A pair generation to empower *Rumi*. To ensure correctness, we acted as a validation layer to check if the GPT-3 generated paraphrased questions and

answers were correct and appropriate. The augmented Q&A database enables Rumi to capture more participant questions and delivered more diverse answers. We believe LLMs could facilitate more chatbot development tasks by teaming with human experts. For example, one could use an LLM as a testing tool by generating question sets to identify issues and develop fix. Such method could enable a faster iteration that traditionally relies on bootstrapping conversations on the fly [274]. Besides using human expertise as a gatekeeper, we should study better LLMs control mechanisms for factual Q&A. For example, we could leverage a knowledge-driven approach [295] by parsing a plain text consent form into a structured knowledge graph and using the graph to steer LLMs to generate factual answers that are grounded in the consent form content. Again, given the shortcomings of generative models, we believe a human-in-the-loop framework is preferred to safely take the advantage of generative models for more capable informed consent chatbots.

In summary, future work should study effective human-in-the-loop frameworks that can support research teams, especially teams without AI expertise and resources, to build and test an informed consent chatbot that consistently delivers factual answers.

## 6.5 LIMITATIONS

We recognize several limitations in our approach. First, as the first study of this kind, our main goal was to explore the potential benefits and limitations of the informed consent process driven by an AI-powered chatbot. Through an SEM model, we further explored the potential path that may explain our observed effect; namely, the chatbot-driven informed consent process improved the study response quality by altering the power relationship. However, due to the exploratory nature of this study, our study could not infer strong causal relationships. Future confirmatory studies are needed to confirm the observed effects and explain the mechanism.

Second, the scope of our study design was limited to online studies with surveys. Although we designed three risk levels to simulate studies that collect different types of data, compared to high-stake clinical trials that may involve severe ramifications, the risk of an uninformed decision is lower in our case. Two factors may limit the generalizability of such a design. First, people tend to pay less attention to consent forms for lower stake studies [43]. We may observe a smaller difference if people were more attentive to the consent form in both conditions. Second, the study procedure was simple and straightforward, e.g., complete a survey. Although it represents the majority of online studies, some studies may include more complicated study procedures, for example, playing a game, where a good understanding of the procedural details may play a stronger effect in the later study. Thus, we need to study

chatbot-driven informed consent under various contexts and studies with different levels of complexity.

Third, participants who declined to join the study are missing from our analysis. Even though, in our study design, participants who decided not to join the study were redirected to Section 2 and offered the opportunity to complete the consent form evaluation. However, some participants may have closed the consent form without answering it. Although other factors may play a role, e.g., usability issues, those participants could read the consent form carefully and make an informed not-to-participate decision. The current study design did not include those participants in the analysis. In our study, 26 out of 278 participants opened our consent form (Chatbot Condition: N = 18; Form Condition: N = 8) without completing it. We believe the effect of this potential confounding factor on our results is limited, but a future study is necessary.

Fourth, our study was designed to investigate the *holistic* effect of using an AI-powered chatbot to lead the informed consent process. However, the design of a chatbot (e.g., language style, name, and appearance) and its capability (e.g., natural language interpretation, question answering, dialogue management) are important design dimensions that may have an effect on the final outcome. As the first step, we aimed to build *Rumi* to deliver the best possible experience. The data collected in this study were inadequate to tease apart and quantify the contribution of each individual design factor. Since each of the interaction features have both benefits and risks [173], it is valuable to rigorously quantify the contribution of different features. This, however, requires additional, fully controlled experiments that are beyond the scope of the current study.

Lastly, although chatbots are increasingly adopted in our daily lives, from customer service to conversational surveys [14], it is still uncommon to use chatbots to conduct an informed consent process. Used in the first study of its kind, *Rumi* was novel to most participants. Since we could not control for the novelty effect in our current study design, we did not know the impact of novelty factors. While we are planning longitudinal studies to examine the influence of the novelty effect, the novelty effect may wear off, like any novel technology, as chatbots become a norm. In our case, as chatbot-driven informed consent becomes more common, the effect is most likely to wear off, similar to how more studies are using e-consent forms today [45].

## 6.6 CONCLUSION

In this chapter, we examine the role of an AI-powered chatbot in improving informed consent online. We built, *Rumi*, an AI-powered chatbot that can greet a participant, go through

the consent form section by section, answer the participant's questions, and collect people's consent responses, to simulate an in-person informed consent experience. We designed and conducted a between-subject study that compared *Rumi* with a typical form-based informed consent process in the context of an online survey study about people's social media use to examine the holistic effect of a chatbot in leading an online informed consent process. We found *Rumi* improved consent form reading, promoted a more equal power relationship between the participant and the researcher, and improved the study response quality. Our exploratory path model indicated the improved study response quality may be attributed to the reduced power gap by the chatbot-driven informed consent process. Given our study results and the simplicity of creating such a chatbot, our work suggests a new and promising method for conducting effective online informed consent. As chatbots become more popular, our results also present important design implications for creating more effective informed consent chatbots.

## CHAPTER 7: USING A CONVERSATIONAL AI TO UNDERSTAND INDIVIDUALS AND HELP TEAMING

### 7.1 OVERVIEW

In this chapter, we aimed to study how people respond to a conversational AI’s information request in a *real world* setting and how information collectors could gain additional insight from the collected data. To do this, we deployed a conversational AI in an education context to study teaming<sup>30</sup>.

In many real-world tasks, from workplace collaborations to student course projects, team effort not only is required, but it also plays a pivotal role in their success. Many organizations thus consider teamwork—one’s ability to collaborate with others and effectively work in a team—as one of the most important and desirable skills of their employees. Numerous research efforts show that team performance can be influenced by a large number of factors and their combinations, such as individual team members’ characteristics and team personality compositions [296, 297, 298, 299]. It is thus a daunting task to understand the interactions of these factors and their effects on team performance, let alone using these factors to guide team formation and behavior. Although there are existing tools that aid in teaming process (e.g., team formation<sup>31</sup>), there is a lack of tools to help understand individual team members, team compositions, and their relationships with team performance [300].

Given the importance of teamwork in organizations, there has been an increasing focus on engaging students in team projects. On the one hand, students can practice and thereby improve their teamwork skills. On the other hand, research on cooperative learning has shown that learning can be enhanced in group environments in which students can learn more actively (e.g., [301]). Educational environments therefore serve as a good testbed for studying teaming, since there are often many teams with varying characteristics, and team performance can be easily measured and compared by the same or similar set of assignments and projects.

To study and potentially guide student teaming efforts in an educational setting, we have developed a conversational AI called INDIGO (Individual Differences for Group Optimization) to interact with a user through one-on-one text-based conversations. INDIGO can also automatically infer a user’s personality traits based on his/her chat behavior without directly asking any personality test questions, which is known to be problematic [302]. We intend to

---

<sup>30</sup>This Chapter has been published at the 24th International Conference on Intelligent User Interfaces (IUI 19) [228]

<sup>31</sup>[catme.org](http://catme.org)

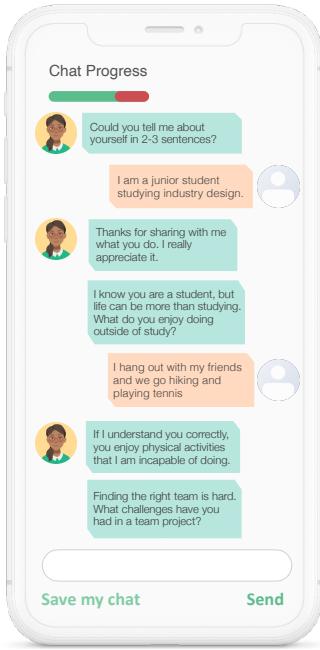


Figure 7.1: A screenshot of an example conversation that INDIGO talks to a student. In the conversation, the agent asks about the student’s past team experience and their team preferences.

use INDIGO to achieve three goals. First, we hope to use INDIGO as a more engaging tool than a traditional online survey to gather initial information from students, such as their team preferences and expectations. Interacting with a chatbot is a new experience, which may help combat survey fatigue and collect higher-quality information.

Second, we hope to use INDIGO to gauge students’ personality traits objectively, preventing potential faking in traditional personality tests [302]. Such personality measures can then be used to study team personality composition and its effect on team performance [297, 298, 303]. In an educational setting, students might provide fake answers to self-assessment surveys to make themselves more desirable to potential teammates. To prevent fake answers, we want to automatically infer a student’s personality traits without asking any traditional personality test questions (e.g., self-reported rating on “I get angry easily”). While personality inference tools exist (e.g., IBM Watson Personality Insights<sup>32</sup>), they often require a certain amount of personal data (e.g., a minimal 500 words from one’s social media account) and many students may feel reluctant to provide such data. Naturally, a chatbot, as described in [101], would help us address this challenge. In short, we create INDIGO to kill two birds with one stone: eliciting user input regarding teaming (e.g., a student’s team preferences and team

---

<sup>32</sup>[www.ibm.com/watson/services/personality-insights/](http://www.ibm.com/watson/services/personality-insights/)

experience) and using the same input to automatically infer user’s personality traits without asking any additional personality test questions.

Third, we hope to test the feasibility of using INDIGO as a long-term team companion that can follow a team and interact with team members continuously during their team efforts. It can be used to detect any change in team dynamics and potentially guide team behavior based on the changes. If it is feasible, INDIGO may therefore serve as a useful tool to help conduct longitudinal team studies.

To achieve our goals, we deployed INDIGO in a real-world setting, where it was used to interact with 191 students who enrolled in a large engineering class at a university, formed 40 teams, and engaged in semester-long team projects. INDIGO interacted with each student at the beginning and end of a class to elicit their views and opinions about teaming, including team preferences and reflections. It also automatically inferred the students’ personality traits based on their chat behavior [101]. We recorded students’ interaction with INDIGO, including their impression of INDIGO. We also tracked each team’s performance throughout the semester. For comparison purpose, we also used traditional surveys to gather students’ input at the beginning and end of the semester to understand their team preferences and perceptions.

Based on the collected data, we performed a series of analyses to answer three key research questions.

- **RQ7.1:** How well do students interact with INDIGO?
- **RQ7.2:** How effectively can INDIGO gather information from students?
- **RQ7.3:** How the personality traits computed by INDIGO can provide insights into the relations between team composition and team performance?

To preview our results, we found that students interacted with INDIGO for an extensive period of time (e.g., 60-minute chat in their pre-course interview) and offered open and honest views. Moreover, INDIGO was able to elicit much richer information from the students compared to the traditional surveys. Third, specific team personality compositions based on the inferred personality traits *significantly* predict team perception and team performance.

As a result, this chapter offers three unique contributions. First, it suggests a new approach to team formation. Through text-based conversations, INDIGO can interview students to gather their team preferences and expectations, and thereby measure their personality traits. It can then use such results to recommend team compositions for optimal teaming. Second, our use of INDIGO through a real-world teaming task demonstrates its practical value beyond

team formation. Given that the students were willing to spend an extensive period of time with INDIGO, INDIGO could be potentially used as a team companion to accompany teams longitudinally, track team changes, and guide team behavior in real-time. Third, we also present a novel, effective method for researchers to investigate teaming in the real world and better understand how team composition impacts team performance.

## 7.2 INDIGO FIELD DEPLOYMENT

To evaluate INDIGO, we conducted a field study in a Computer Science course offered at a large public university. This course required students to team up and collaborate on a semester-long project to create a web/mobile application, which determined 40 % of their final course grade.

### 7.2.1 Participants

A total of 191 students enrolled in the class and 184 of them consented to participate in the study ( $N=184$ ). While the majority of the students majored in Computer Science, there were students from non-engineering majors, such as Industrial Design, Information Science, and Psychology.

### 7.2.2 INDIGO Interviews

We deployed INDIGO at the beginning and the end of the semester to interview each student enrolled in the class.

Each pre-course interview included five sections. The first section was a warm-up conversation, during which INDIGO and a student were introduced to each other and chatted about their hobbies and favorite movie. The second section included a 20-item Impression Management (IM) Scale questionnaire to measure how students consciously favor themselves to impress others [304]. Since IM is found highly related to self-control and social adaptation in workplace [305], we hypothesized that it might affect team performance. In the third section, INDIGO asked the student a set of open-ended questions regarding his/her team experience and team preferences. For example, it asked a student “*what is your preferred role in a team*” and “*what kind of individuals do you want to have on your project team?*”. In the fourth part of the interview, INDIGO discussed with the student about his/her own characteristics, such as his/her strengths and weaknesses. The purpose of this discussion was to gauge how much a student is willing to trust INDIGO and share personal information

including sensitive information such as weaknesses. The last section was to solicit the student's feedback about INDOGO. The student was asked to express his/her impression of INDIGO and rate INDIGO on a number of scales, such as its helpfulness and likeability.

At the end of the semester, INDOGO interviewed each student again to elicit their views and opinions on their overall team experience. The post-course interview included questions, such as "*What is your overall team working experience*" and "*What kind of suggestions do you have to improve the teaming experience*".

### 7.2.3 Team Formation

To investigate whether INDIGO could offer useful insights for team formation, we wanted to learn the effect of existing team formation tools. Thus, the course instructor used CATME, an existing team formation tool<sup>33</sup>, to help team assignments. In this tool, the instructor first selected a set of criteria, such as team skills, working style, and demographics. By these criteria, the students were asked to take a series of surveys to obtain their self-reported measures. Based on the completed surveys, the instructor then set the weight for each criterion and ran the algorithm to obtain team assignments.

In our study, the instructor weighted the following criteria the most: languages skills, skill set (e.g., programming, UI design, and teamwork), programming capability, leadership preference (e.g., single leader vs. shared leadership), leadership role (follower or leader), and thinking style (e.g., big picture vs. detail-oriented). The algorithm was also configured to make the teams as diverse as possible by the weighted criteria.

Although all students in the course were told the teams were selected by an algorithm, only half of the teams ( $N=19$ ) were assigned by the algorithm and another half ( $N=21$ ) were assigned randomly. A total of 40 teams were formed with 4-5 students in each team.

### 7.2.4 Team Personality Composition

Existing studies showed that team personality composition influences team dynamics and team performance [297, 298, 299, 303, 306]. To collect a team's personality composition, we used the students' Big 5 personality traits inferred by INDIGO during their interview.

---

<sup>33</sup>[catme.org](http://catme.org)

### 7.2.5 Team Perception Survey

In addition to personality composition, team dynamics is known to affect team performance [307, 308, 309, 310]. At the end of the course, we used a survey provided by the team formation tool CATME to gather students' perception of their own team. As we will describe below, this survey allows us to characterize a team using a scale with seven dimensions, with each dimension consisted of a set of statements on a 1-5 Likert scale (1 totally disagree to 5 totally agree):

- *Psychological Safety.* This is a 7-item survey to measure how safe a person feels to take interpersonal risks in a team. It is known to affect team performance [311].
- *Interpersonal Cohesiveness.* This 3-item survey measures the interpersonal relationship between team members. In particular, it measures interpersonal attraction perceived by team members toward each other [312].
- *Task Commitment.* This dimension uses 3 items to measure the level of team member's commitment toward the group goal and how much effort was made by each teammate.
- *Task Attraction.* This dimension also includes 3 items to gauge the overall working atmosphere in a group. For example, it assesses how much the members enjoyed the group activities or the group's work as a whole.
- *Relationship Conflict.* This 3-item survey measures the level of tension perceived in the workgroup and the frequency of negative emotion generated during the working process.
- *Group Task Conflict.* This dimension uses 3 items to assess the conflicts among team members in terms of the group task, specifically, the frequency of conflicting opinions in the working process.
- *Process Conflict.* This 2-item dimension measures the conflicts during work distribution and resource allocation, e.g., how much conflict was there on task responsibilities.

## 7.3 RESULTS

To evaluate INDIGO and answer our research questions, we have examined multiple sources of data from INDIGO's field deployment, including the chat transcripts between each student and INDIGO and INDIGO-inferred personality traits of the students. Here we report the

findings to answer our three research questions, respectively: (a) users' interaction with INDIGO, (b) the effectiveness of INDIGO in eliciting information from the students, and (c) the effect of INDIGO-derived personality traits on team performance and team perception.

### 7.3.1 RQ7.1: Students' Interaction with INDIGO

We first examined students' engagement with INDIGO. Our results showed that on average each student spent about 60 minutes ( $SD = 26$  minutes) with INDIGO in their pre-course interview and 26 minutes ( $SD = 7$  minutes) in the post-course interview. In addition to *engagement duration*, we also computed the *response length*, defined by the number of words in each student's responses, to gauge the amount of information that the students were willing to provide during their interviews. On average, each student provided 620 words ( $SD = 291$  words) in their pre-course interview and 289 words ( $SD = 184$  words) in their post-course interview. These measurements demonstrated that the students were willing to spend a considerable amount of time with INDIGO and offer information during their interaction with INDIGO.

**Perceived Characteristics of INDIGO** We examined students' impressions of INDIGO by examining their descriptions and ratings of INDIGO.

The students were asked to describe their impression of INDIGO in three keywords. The top-5 most mentioned keywords were *friendly*, *robotic*, *kind*, *nice* and *polite*. From these words, it seemed that the students perceived their interaction with INDIGO positively. In fact, 80% of students provided *all* positive expressions when describing their impression of INDIGO, such as "*agreeable, friendly, perceptive*"; "*nice, charming, funny*"; and "*sweet, smart, and well built*"

The students were also asked to rate INDIGO on three dimensions, *likeable*, *helpful*, *enjoyable*, on a scale of 1 to 5, 1 being not at all and 5 being very much. The average rating for each dimension was: *likeable* 3.14 ( $SD = 1.29$ ), *helpfulness* 3.12 ( $SD = 1.26$ ), *enjoyable* 2.53 ( $SD = 1.27$ ). Overall, the students seemed ambivalent about INDIGO and did not quite enjoy their chat with INDIGO. In the hope of finding explanations, we examined the students' chat transcripts with INDIGO. The data helped explain the ratings from a couple of angles. First, the chat transcripts revealed that INDIGO was limited at understanding a student's complex input, which certainly did not make the chat enjoyable. For example, one student commented on:

*"[INDIGO] doesn't understand context when giving responses"*

Another student also stated:

*“[INDIGO is] not a huge conversationalist (not too much enthusiasm or talking outside of the script)”*

This was also consistent with the students’ description of INDIGO since quite a few used the word “robotic” to describe their impression of INDIGO. Although we leveraged the best conversational agent that is available to us, INDIGO still has much to improve especially its ability to interpret a user’s complex and diverse input.

Additionally, the interviews with INDIGO especially the pre-course interview were long (e.g., 60 minutes), which might have made the experience less enjoyable. However, considering people’s tolerance with traditional surveys [313], INDIGO’s engagement duration with the students is quite remarkable.

Despite their ambivalence about INDIGO, it is encouraging to observe that the students were still willing to interact with INDIGO and offer rich information.

**Perceived Role of INDIGO** One of our goals was to investigate whether a conversational agent like INDIGO could serve as a team companion. To find out in which role INDIGO could serve a team, we asked students’ perceived role of INDIGO by rating INDIGO on two roles: *like a friend* and *like a counselor* on a scale 1-5. Students perceived INDIGO more like a counselor ( $M = 3.24$ ,  $SD = 3.26$ ) than a friend ( $M = 2.52$ ,  $SD = 1.28$ ). This might be another reason why students felt the conversation was not enjoyable as it was not like chatting with a friend.

Moreover, when the students were asked to rate how much they trusted INDIGO on a 5-point Likert scale, they indicated that they somewhat trusted INDIGO:  $M=3.49$ ,  $SD=0.99$ . To understand the students’ trust in INDIGO, we further examined the chat transcripts and found that the students were indeed quite open and honest in offering their opinions to INDIGO. For example, when asked about their weaknesses, one student stated:

*“I suck at comprehending things... I feel like I’m pretty slow. It takes me a while to grasps concepts and that along with my slight laziness doesn’t make for the best combo.”*

Similarly, another answered:

*“I need someone to guide me, in other words, it’s hard for me to start one thing without any guidance.”*

	Similar Personality
Enjoyable	0.56, p<0.05*
Likeable	0.49, p<0.05*
Helpful	0.50, p<0.05*
Trust	0.36, p<0.05*

Table 7.1: The correlation between students perceived similarity about INDIGO’s personality traits and their impression of INDIGO

Considering that these conversations occurred *before* they had found their teammates for their class project, many students seemed to have provided honest answers and did not try to hide their weaknesses.

Similarly, when asked what kind of role they want to play in a team, the students were honest to state the role they wanted to play. For example, one student stated:

*“I would prefer to not be a leader , esp . not this semester because I have a lot of other things going on.”*

Likewise, another student mentioned:

*“I don’t want to be a leader because I have too many assignments to work on.”*

Similar to the findings in [101], students’ perception and their behavior with INDIGO showed that they somewhat trusted an agent like INDIGO and were willing to disclose personal information during the interaction. Although in our study, we did not intentionally frame INDIGO as a counselor, many students considered it as one, which might have also encouraged them to open up and offer truthful information. Understanding the students’ perception of the role of INDIGO is important especially if we wish to use INDIGO as a team companion, which must be effective at eliciting authentic team information to understand the true team dynamics.

**Perceived Relation with INDIGO** Previous research shows that users enjoyed their interaction more with an agent if they perceive the agent has a personality similar to theirs [314]. The students were asked to rate how similar they were to the personality of INDIGO (*SimilarPersonality*). We then examined the correlation between the *SimilarPersonality* rating and all other user ratings, such as *likeable* and *enjoyable*. The analysis revealed a moderate correlation with all these ratings (Table 7.1), which suggests the potential to adapt INDIGO’s personality to that of a user to improve user experience with INDIGO.

### 7.3.2 RQ7.2: Effectiveness of INDIGO in Information Gathering

One of the main purposes of using INDIGO is to provide a more engaging way to gather information from the students, compared to the traditional, static online surveys. We thus evaluated INDIGO on its effectiveness of gathering information from the perspective of the instructor. To form teams and understand team dynamics, instructors often use traditional online surveys to learn about the students and their team experience. Compared to these surveys, the conversational interview conducted by INDIGO used more open-ended questions. While open-ended questions help elicit richer information and provide rationales behind quantitative ratings, research shows that collecting responses to open-ended questions is often difficult [33]. We thus conducted a series of analyses to examine whether the open-ended questions posed by INDIGO helped elicit useful information that can benefit the course instructor.

We first compared student's responses to the question, "*What is your preferred role in a team*" which was asked in both the CATME survey and INDIGO's interview before the teams were formed. The CATME survey used a choice-based question with five options from *strongly prefer to be a follower* to *strongly prefer to be a leader*. In contrast, INDIGO posed it as an open-ended question. To capture the gist of student responses to this question, we used an enhanced Latent Dirichlet Allocation (LDA) model [230] to analyze the 184 responses and automatically derive a set of semantic themes covered by the responses. We also used LexRank [234] to find representative sentences within each theme.

The themes produced by the LDA model not only covered *all* options presented in the choice-based question but also gave additional information, such as how they wish to play a role and the rationale why they wanted to serve a particular role, for example, one student mentioned

*"I prefer to take turns leading and following."*

It would be difficult to put this student's answer into a category yet the information is valuable. Similarly, another student stated his preferred role on a team:

*"I prefer to be the coordinator in the team. I would like to collect different ideas from team members, do some conclusion, share ideas among different groups and ask for advices from professor and TAs."*

Not only was additional information collected, but the instructor could also better understand the "*why*" behind the students' ratings. For example, at the end of the course, the students were asked to rate their overall team experience on a scale of 1 to 5, where 1 being

poor and 5 being excellent. INDIGO also conducted a post-course interview that asked the students how they felt about their team experience. Coupling the students' responses to INDIGO with their ratings, the instructor got a more comprehensive picture on how the teams worked together. For example, a student who gave a 5 rating wrote to INDIGO,

*“... all of us were very supportive of each other and we split up the work evenly”*

In contrast, a student who gave a 1 rating mentioned,

*“It was a little tiring when others wanted to leave off the work until the last second”*

From the above examples, INDIGO was able to elicit useful information from the students. One might argue that the open-ended questions could be inserted into a regular survey to collect the needed information. However, extensive survey statistics show that people are willing to spend only a few seconds per question on a survey that lasts more than 5 minutes [313]. Our use of INDIGO indicated that the students were willing to spend time interacting with it and offering useful information, which suggests an alternative to static surveys.

### 7.3.3 RQ7.3: Effect of INDIGO's Personality Insights on Team Results

To examine whether and how INDIGO could help team formation, we looked into its inferred student personality traits and investigated the effect of team personality composition on team results. Specifically, we wanted to answer two questions:

- **RQ7.3a:** How does team personality composition impact student team performance?
- **RQ7.3b:** How does team personality composition impact team members' perceptions of their own team?

We used the data from three sources: (a) 184 students' 35 Big 5 personality traits inferred by INDIGO, (b) 184 students' self-reported team perception by seven dimensions, and (c) 40 teams' project performance. Because of the number of data dimensions involved, we first performed factor analyses to examine the relationships among the relevant data dimensions.

**Factor Analysis of Inferred Personality Traits** We first examined the factorability of the inferred 35 personality traits. The results showed that 33 of 35 measures correlated with at least one other item ( $R^2 \geq 0.3$ ). Moreover, the Bartlett's test of sphericity was significant

Factors	Personality Traits
Emotional	<b>Neuroticism</b> , Depression, Impulsiveness, Vulnerability
Collaborative	<b>Agreeableness</b> , Cooperation, Sympathy
Social	<b>Extroversion</b> , Friendliness, Gregariousness
Open-minded	<b>Openness</b> , Imagination, Intellectual Curiosity
Responsible	<b>Conscientiousness</b> , Dutifulness, Cautiousness
Sensitive	Feelings
Self-Disciplined	Self-Discipline, Anxiety, Vulnerability

Table 7.2: Personality traits loaded onto 7 separate factors. The bold trait indicates the trait is one of the Big five personality traits

$(\chi^2(595) = 35091.39, p < .05)$ . Thus factor analysis was suitable for all 35 items. A Principal Components Analysis (PCA) indicated a seven-factor solution, which explained 53% of the variance. A screen plot also showed the sharp leveling off of Eigenvalues after the seven factors. Table 7.2 lists the seven factors. For each of the seven factors, we then measured a composite trait score, a regression-weighted mean of items with primary loadings greater than 0.5 in the factor. Internal consistency for each score was also examined using Cronbach's alpha. The alphas were: 0.78 for *Emotional* (4 items), 0.63 for *Collaborative* (3 items), 0.53 for *Social* (3 items), 0.55 for *Open-minded* (3 items), 0.43 for *Responsible* (3 items), 0.73 for *Sensitive* (3 items), and 0.39 for *Self-Disciplined* (2 items). Overall, our analyses indicated that seven distinct factors were underlying the students' personality measures with reasonable internal consistency. Only one item had a cross-loading above 0.5 (*Vulnerability*), however, this item had a strong primary loading of 0.64.

**Factor Analysis on Team Perception Measures** We also performed factor analysis on the seven dimensions that measured students' preception of their own team. The Bartlett's test of sphericity was significant  $(\chi^2(21) = 119.74, p < 0.05)$ . A two-factor solution was derived from PCA, which explained 59% of the variances. Accordingly, we created two composite scores for both factors, respectively. Table 7.3 showed these two factors. Each composite score was regression-weighted on the relevant items with their primary loadings greater than 0.5. Cronbach's alphas were also computed: 0.84 for *Team Cohesion* (4 items) and 0.69 for *Team Conflict* (3 items). And no cross loading was found.

**Analysis Variables** To answer our questions above, we computed a set of measures as independent and dependent variables, respectively.

	Team Relationship	Team Conflict
Psychology Safety	0.55	
Interpersonal Cohesiveness	0.91	
Task Commitment	0.58	
Task Attraction	0.85	
Relationship Conflict		0.73
Task Conflict		0.56
Process Conflict		0.76

Table 7.3: Factor loadings for two factors from the Student’s Team reflection. The two-factor solution explained 59% of the variances.

## Independent Variables

- *Team Personality Composition.* Numerous studies show that both mean and variance of team personality composition scores influence team performance [298, 299, 303]. For each team, we measured its personality composition by the mean and variance of the individual team members’ personality scores. Each individual’s personality scores were computed by the seven extracted personality factors (Table 7.2).
- *Team Formation Method.* A binary variable indicates whether a team was determined by an algorithm (value=1) or randomly assigned (value=0).
- *Impression Management Score (IM Score).* IM score reflects how students consciously favor themselves to impress others. For each team, the IM Score is a mean of its team member’s individual IM scores.

## Dependent Variables

- *Team Performance.* For each team, its performance was based on the team’s final project score, which included the scores of all project milestones throughout the course.
- *Team Perception.* For each team, we computed the mean of each of the two factors, Team Relationship and Team Conflict, extracted based on the ratings reported by each team member in their survey.

**Analysis Methods** Before examining the effect of the above independent variables on each of the dependent variables, we first analyzed the relationship between the dependent variables. A correlation test showed no significant relationships between *Team Relationship* and *Team Performance*:  $r(38) = -0.05$ ,  $p = 0.76$  and neither between *Team Conflict* and *Team Performance*:  $r(38) = 0.29$ ,  $p = 0.07$ .

Predictor	<i>b</i>	<i>beta</i>	<i>sr</i> <sup>2</sup>	<i>r</i>
(Intercept)	91.05**			
Emotional	.07*	.43	.12	.34*
Collaborative	-.03	-.12	-.01	-.05
Social	-.04	-.16	.02	.01
Open-minded	.01	.04	.00	.03
Responsible	-.05	-.22	.05	-.13
Sensitive	-.01	-.02	.00	.20
Self-Disciplined	-.06	-.23	.05	-.10
Impression Management	.56*	.37	.11	.36*
Team-formation Method	-.05	-.01	.00	.10

Table 7.4: Regression results using Team Performance as the outcome variable. The analysis results showed that the *Emotional* variance in team personality composition and the level of Impression Management significantly predicted team performance.

**RQ7.3a: Effect of Team Personality Composition on Team Performance** Using *Team Personality Composition* as an independent variable and *Team Performance* as a dependent variable, we built a regression model controlling for Impression Management (IM) scores and *Team Formation Method*.

The analysis results showed that the *Emotional* variance in team personality composition ( $\beta = .43$ ,  $t(30) = 2.30$ ,  $p < 0.05$ ) and the level of Impression Management ( $\beta = 0.37$ ,  $t(30) = 2.22$ ,  $p < 0.05$ ) significantly predicted team performance (Table 7.4). The *Emotional* variance in team personality composition alone explained a significant proportion of variance in team performance,  $R^2 = 0.34$ ,  $p < 0.05$ . The level of IM alone explained a significant of the variance as well,  $R^2 = 0.36$ ,  $p < 0.05$ . No other effect was significant.

In other words, the more diverse a team was in its Emotional makeup, the better its project performance was. Although no prior study on teaming reports such findings, this result is consistent with prior findings in personality research. In particular, the *Emotional* factor consisted of 4 items on the Neuroticism dimension (Table 7.2). According to Oertig et al. [315], people high in Neuroticism tend to perform better in short-term goals with deadlines, which is very similar to our class project setting. On the other hand, people low in Neuroticism possess higher emotional intelligence, which in turn helps overall team effort [316]. In summary, a team with members at varied levels of Neuroticism could benefit from both sides to perform better. Moreover, not only is our unique finding derived from a field study, but it also helps verify several previous teaming theories. For example, Lykourentzou shows a balanced personality composition benefit team performance [299] and how emotional traits affect teamwork [299, 316].

Team Member	Team A ( $Var_{Emotional} = 713$ )				Team B ( $Var_{Emotional} = 219.7$ )				
	P1	P2	P3	P4	P1	P2	P3	P4	P5
Emotional	232	222	214	274	258	225	236	247	223
GPA	3.92	3.4	3.5	2.30	3.9	3.96	3.93	3.0	3.72
Skill Set	T,W,P	T,P	P	T,U,P	T,W,U,P	P	T,W	T,P	T,W,P
Leadership Preference	O	O	O	S	O	O	O	S	S
Leadership Role	N	F	F	N	F	F	F	N	N
Thinking Style	I	B	B	I	B	B	B	I	I

Note. T indicates Team-work. P indicates Programming. W indicates Writing Skill. D indicates Design Skill. O indicates One Leader with Input. S indicates Shared Leadership. F indicates Follower. B indicates No Preference between Follower and Leader. I indicates Idea Oriented. B indicates Balanced between Idea Oriented and Detail Oriented.

Table 7.5: Team characteristics comparison between Team A and Team B. Team A had a diverse Emotional makeup and achieved a high team performance and Team B had a more homogeneous Emotional makeup and achieved a low team performance.

Predictor	<i>b</i>	<i>beta</i>	<i>sr^2</i>	<i>r</i>
(Intercept)	3.30			
Emotional	.01	.23	.03	-.09
Collaborative	.01	.25	.08	-.36*
Social	-.00	-.07	.01	-.12
Open-minded	.01	.28	.03	.26
Responsible	.00	-.02	.00	-.01
Sensitive	-.02*	-.42	.10	.23
Self-Disciplined	-.01	-.14	.02	-.13
Impression	.02	.08	.01	-.01
Management				
Team-formation				
Method	-.27*	-.33	.10	.35*

Table 7.6: Regression results using Team Conflict as the outcome variable. The results showed that the mean of a team's *Sensitive* score significantly predicted *Team Conflict*.

To better illustrate our finding, we chose and compared the characteristics of one team that had a diverse Emotional makeup and achieved a high team performance (Team A) with another team that had a more homogeneous Emotional makeup and achieved a low team performance (Team B) (see Table 7.5). The compositions of these two teams were very similar in terms of the skills, leadership preference, leadership role preference, and thinking style. Those criteria were the most weighted criteria set by the instructor during the team formation, which are also commonly used by other project-based classes [317]. While these characteristics are very similar, the variances of team *Emotional* composition were very different ( $Var_{TeamA} = 713$ ,  $Var_{TeamB} = 219.7$ ).

To understand how the team's *Emotional* makeup impacted team performance, we in-

terviewed the teaching assistant (TA) who mentored both teams. The TA mentioned that “Team A was often concerned with their grade and kept sending me emails to ask what’s the requirement and when the next assignment will be due.” It is also interesting to observe that although instructors often intuitively believe that general academic performance (e.g., GPA) of students is most predictive of team performance, we can see from Table 7.5 that members of Team A ( $M = 3.28$ ) actually had a *lower* mean GPA than that of those in Team B ( $M = 3.70$ ). Also, one student on team A who had a relatively high emotional score also had a low GPA. Based on the TA’s comments, one explanation why Team A performed better was that the members with a high *Emotional* score tended to remind the team about upcoming deadlines and nudge the team to finish the work on time. On the other hand, emotionally more calm members helped hold the team together without being overwhelmed by the deadlines. It is possible that this was the reason that Team A achieved a much higher overall project score than Team B (98.7% vs 90.4%). As an example, the analyses afforded by INDIGO led us to focus on important aspects of the teams, which provided important insights on how team composition could play a pivotal role in influencing team performance.

Our finding above suggests that emotional makeup of a team impact team performance. To create teams with high performance especially for accomplishing short-term goals, INDIGO could be used to first understand individuals’ personality traits and then suggest teams that are made up of members with varied emotional characteristics.

As we hypothesized, our analysis indicated that the Impression Management Scale (IM) also influenced team performance. In particular, the higher the average IM score was in a team, the better the team performed. Our finding seems consistent with previous research on relating IM scale with self-control and social adaptation [305]. In other words, a team would perform better, if all team members have a high level of self-control and can adapt well socially in a team setting. In addition, we further examined whether the inferred personality traits could predict the IM scores. We found that among the seven factors, *Sensitive* (beta = 0.25,  $p = 0.01^{**}$ ), *Emotional* (marginal, beta = -0.19,  $p = 0.05$ ), and *Social* (marginal, beta = -0.15,  $p = 0.07$ ). This suggests that personality traits may be used to infer IM scores *automatically*, which can then be used to suggest team formations (e.g., trying to form teams with a higher average IM scores). The significant relationship between team personality composition, impression management score, and team performance indicates the potential of an agent like INDIGO: it could be used to understand individuals by inferring their personality traits and then use the inferred traits to recommend high-performance teams.

**RQ7.3b: Effect of Team Personality Composition on Team Perception** To answer our second question above, we built a regression model that used team personality composition as independent variables and team perception as dependent variables. Regression results showed that the mean of a team's *Sensitive* score significantly predicted *Team Conflict*,  $\beta = -0.42$ ,  $t(30) = -2.10$ ,  $p < 0.05$  (See Table 7.6). In particular, the less sensitive a team was, the fewer conflicts a team experienced. Moreover, *Team Formation Method* played a role. Teams assigned by the tool reported fewer team conflicts,  $\beta = -0.33$ ,  $t(30) = 2.24$ ,  $p < 0.05$ . No other effect was significant, see Table 7.6.

Our results suggest that teams that were lower on the *Sensitive* measure experienced fewer conflicts during the teaming process. The *Sensitive* measure is highly loaded on *Feelings*, *Anxiety* and *Vulnerability*, which suggest that people high on these dimensions are more sensitive and vulnerable to conflicts and negative feelings [318]. When facing upcoming deadlines, those team members might have expressed more negative emotions. Furthermore, their similarly vulnerable teammates could not cope with the negative feelings and might cast their own negative emotions. The intensified negative feelings would then create more team conflicts and affect their team relationship.

In addition, we found teams that were assigned by the team formation tool were more satisfied with their team. This finding was aligned with previous research that team formation algorithm could help improve team dynamics [317]. We further tested whether the team formation method interacted with the team personality composition. The result showed no significant effect.

Again, the relationship between students' inferred personality and team perception indicates INDIGO's potential to recommend teams based on team personality compositions that will optimize team experience.

## 7.4 DISCUSSION

The field deployment of INDIGO and its demonstrated value offered encouraging results. Students actively engaged with INDIGO and offered rich information for forming effective teams. The instructor was able to gain deeper insights into students as unique individuals along with their teaming preferences and experience. We believe that the novel use of a conversational agent like INDIGO contributes to the rich information harvested. Before INDIGO, traditional teaming tools require students to take a long survey before assigning teams. Survey fatigue may prevent students from giving truthful and in-depth information. On the other hand, the interactivity of INDIGO may reduce survey fatigue as it had an hourly long chat with each student.

Moreover, INDIGO inferred team personality composition predicted team outcomes. Our result shows that teams with a higher variance in their *Emotional* makeup performed better and teams with lower on the *Sensitive* measure experienced fewer conflicts. Without INDIGO, such relationships were difficult for traditional tools to discover. Additionally, INDIGO saved the extra effort required for the students to assess their personality, not mentioning the objectivity in these results due to social desirability bias. The relationships discovered by INDIGO also provide guidance for instructors to choose the appropriate criteria to form effective teams.

## 7.5 DESIGN IMPLICATIONS

We believe the design of the intelligent teaming tool can benefit from our findings. First, from the interaction between INDIGO and students in the class, we learned that students trusted INDIGO and perceived INDIGO as a counselor. In the future design of an intelligent agent for teaming, we could leverage such perception by framing the agent as a team coach. Similar to a counselor in real life, who would follow up with their clients, a team coach can follow up with individual team members throughout their teaming efforts. Such a team coach can collect students' team perceptions in real-time and track their changes. Instructors and teaching assistants can then use the gathered information to intervene or guide group activities, such as helping reduce interpersonal tension or resolve interpersonal conflicts.

Second, our results show that students prefer INDIGO more if they perceive INDIGO having a personality similar to theirs. As one of the design considerations, we could adapt an intelligent agent to that of a student to better engage with the student.

Third, since our findings reveal that team personality composition predicts team outcomes, we can use such findings to augment INDIGO. In particular, we can extend INDIGO to automatically recommend team formation based on its inferred personality traits. For example, it can select each team by maximizing the variance of emotional characteristics of the team members.

## 7.6 LIMITATIONS

Our current work has several limitations. First, the measure of team performance in our study was limited to one project score. The goal of education should go beyond a simple score. For example, the project score does not fully reflect how many teamwork skills students have learned in the process. Moreover, research suggests that a lower score sometimes may even

imply better learning [319, 320, 321]. Second, our field deployment of INDIGO was situated in a Computer Science class in a large U.S. public university, which may not be representative of student teaming situations in other cultures since culture often influences teaming and team success [322, 323]. Third, in our evaluation, we collected the self-reported team perception at the end of the semester. Such measures may not fully reflect a team’s status, let alone capture the changes in such status. Ideally, we want to collect team perceptions throughout the semester and detect their changes over time to better assess team dynamics. Although the student teaming effort lasted for a full semester, it is still considered short-term teaming. It is unclear whether our findings would hold for longer teaming efforts, for example, sports teams or workplace teams that may last for many years.

## 7.7 CONCLUSIONS

In this paper, we presented the novel use of a conversational agent called INDIGO (Individual Differences for Group Optimization) to understand student teaming efforts. We evaluated INDIGO through a field deployment involving 191 university students in 40 teams working on semester-long team projects. INDIGO interviewed each student twice to learn the student’s team preference and team experience. From the perspective of students and instructors, INDIGO demonstrates its ability to collect valuable teaming information through students that could help instructors to form effective teams. The relationship discovered between INDIGO inferred student’s personality traits and team outcomes demonstrates the potential to recommend team compositions for optimal teaming experience.

Our findings bear design implications on developing intelligent agents for aiding teaming efforts at different stages, such as recommending team formation, tracking team dynamics, and guiding team behavior.

## CHAPTER 8: DESIGNING VOICE INTERFACE FOR COLLECTING EXPLICIT USER FEEDBACK

### 8.1 OVERVIEW

In recent years, voice assistants, such as Google Assistant, Amazon Alexa, or Apple Siri, have become increasingly popular. Due to the rapid development of machine learning models, voice assistants nowadays deliver more diverse and more personalized content, including daily news briefings, outfit suggestions, and music recommendations. In this chapter, instead of focusing on text-based conversational AI, we moved our focus to a different modality of interaction, voice, and studied how people react and perceive voice assistants' information requests for explicit user feedback <sup>34</sup>.

Researchers and engineers rely on user signals to build quality machine learning models that enable voice assistants to provide personalized experiences [325]. These user signals also help recommender systems to evaluate the recommended content and inform future recommendations. One such signal is explicit user feedback, where users tell the system explicitly about their experience, opinions, or preferences [95]. Compared to other types of user signals, such as interaction logs, explicit user feedback contains less noise [77, 96, 326].

Recent studies found that users are willing to engage with conversational agents and provide meaningful information [173]. This indicates a large, currently under-utilized opportunity to create tighter feedback loops via voice assistants to connect system builders and end users and improve the underlying recommender systems to provide better personalization. However, feedback mechanisms also need to be carefully designed to avoid degrading the user experience [88, 327]. Therefore, without established and well-understood techniques for collecting explicit user feedback, this type of data will not be collected, depriving researchers and engineers of a potential high-quality source of user signal. This missed opportunity might become even more significant as recommender systems — traditionally powered by collaborative filtering — continue their shift towards more responsive algorithms like reinforcement learning, where user feedback plays a central role [328].

Explicit user feedback is especially useful when it is collected *in situ* [88, 329], i.e., in the place and in the moment in which the interaction of interest occurs. It avoids introducing a memory bias when asking users to recall their impression of an interaction that happened in the past [330]. Also, having access to rich contextual information can help system builders improve recommender systems by tailoring content to specific contexts [331]. For example,

---

<sup>34</sup>This Chapter has been published at Proceedings of the ACM on Human-Computer Interaction (CSCW 21) [324]

users who often listen to slow songs may want something more energetic when going for a run. But collecting explicit feedback in situ is challenging: unlike graphical interfaces where system builders can use pop-up questions or upvote/downvote buttons to prompt for feedback, the lack of visual affordances on voice interfaces limits the system’s ability to prompt for user feedback. This “invisible” nature of voice commands [332] makes it difficult for users to even know about the options to provide feedback. Moreover, the voice assistant must interrupt an ongoing interaction to ask for feedback, which can hurt the user experience. For example, if the voice assistant wants to know whether the user likes a song that is playing, the assistant needs to stop the music to ask the user.

Picking the right situations and frequency to ask for feedback is another challenge: frequent feedback requests may cause fatigue that hurts user experience and feedback quality [25]. It is crucial for the voice assistant to ask strategically to reduce experience friction. It should focus on the feedback request that maximizes system gain with minimal user experience friction. To identify what this feedback might be, it is key to involve not only the end users but also experienced machine learning practitioners. These experts know what user signals would be most beneficial to the development of the underlying systems, and they are well-attuned to potential user concerns.

Motivated by the benefits of explicit in situ user feedback and the potential challenges of implementing proactive voice interactions, our research aims to provide an overview and a better understanding of the potential opportunities for voice assistants to engage with users for feedback elicitation. With our work, we are addressing two research questions:

- **RQ8.1:** What types of explicit user feedback are valuable to machine learning practitioners?

We conducted interviews with machine learning (ML) practitioners ( $N = 12$ ) to identify categories of explicit user feedback that are particularly valuable for ML model inspection and improvement. Our analysis revealed four categories, 1) Clarifying user input, 2) Clarifying behavioral signals, 3) Collecting feature feedback, and 4) Understanding user context. By collecting explicit user feedback in situ, a voice assistant could help ML experts evaluate underlying algorithms and enable user models to deliver more accurate recommendations and more engaging experiences. With the understanding of what types of feedback are valuable, we then asked our second research question:

- **RQ8.2:** How do different voice assistant design dimensions affect user willingness to respond to the feedback elicitation request and the perceived friction to the experience?

We conducted an online study ( $N = 294$ ) with realistic scenarios to evaluate user perceptions of different voice assistant designs when collecting explicit user feedback. We look at three design dimensions of a voice assistant: 1) Framing: How the voice assistant is being presented to the user — either as an Assistant, a Collaborator, or a Learner; 2) Elicitation Strategy: How the voice assistant elicits the feedback — either as a direct question that the user is expected to respond to, or as an instruction on how the user can initiate providing feedback, 3) Level of Proactivity: When the feedback elicitation occurs in relation to the user’s interaction with the assistant — as an extension of expected response to a voice command, as a response to a user interaction that typically does not involve a response from the voice assistant, or independently from any user interaction.

Our results show that the above design dimensions of the voice assistant are indeed important aspects to consider when aiming to collect explicit user feedback *in situ*. Users are more willing to respond to voice assistants that are framed as a Collaborator or Learner than to those framed as an Assistant. Voice assistants that provide instructions on how to give feedback are perceived as less disruptive than those asking direct questions. Finally, the choice of the elicitation strategy becomes more consequential as the voice assistant becomes more proactive (i.e. when it attempts to elicit feedback with no direct user interaction preceding it).

Our work contributes in three ways: first, we derived four usage categories through interviews with ML experts in recommender systems and a review of related work. These categories provide potential directions for further investigation to collect explicit user feedback through voice assistants. Second, our online study shows design decisions that can affect a user’s perception and experience when a voice assistant is asking for explicit feedback. Lastly, we discuss design considerations and implications for creating voice assistants that can effectively elicit quality feedback in different contexts and mediate collaborations between developers and users in recommender systems.

## 8.2 METHOD OVERVIEW

Our work aims to address two research questions. In RQ8.1, we identify different types of explicit user feedback that are valuable to ML practitioners. We then use these insights to investigate RQ8.2, learning how voice assistant design decisions can affect users’ willingness to respond to the feedback elicitation and the perceived friction to the experience.

To address RQ8.1, and to lay the foundation for a study that would allow us to explore RQ8.2, our first step was to conduct expert interviews with ML practitioners. Our second step, to address RQ8.2, was to run an online study in which we applied different voice

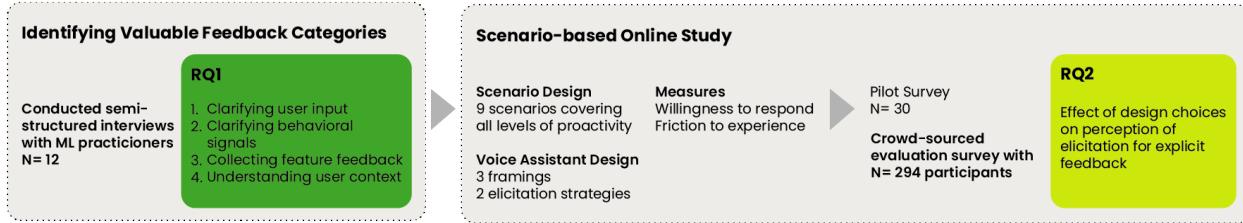


Figure 8.1: Overview of the two steps of our research: from identifying valuable feedback categories through expert interviews to generating scenarios evaluating the effect of different design choices on the user perception.

assistant design decisions to realistic scenarios that were generated based on insights from the interviews and related work.

### 8.3 IDENTIFYING VALUABLE FEEDBACK CATEGORIES THROUGH EXPERT INTERVIEWS

To identify categories of explicit user feedback that are valuable to collect via voice assistants, we conducted semi-structured in-depth interviews with machine learning experts who design, develop, or evaluate machine learning models that directly shape the interactions of end users. We recruited 12 machine learning practitioners (six women and six men) from a content-streaming company for which recommender systems play a central role in most of its products. Six participants have directly worked or are working on recommender systems or user models that interact with users through voice assistants. We conducted semi-structured interviews of approximately 45 minutes with each participant via a video conferencing software<sup>35</sup>. At the beginning of each interview, we asked the participants about their roles, past experience, and challenges in building and evaluating machine learning models. Then we focused on their opinions and thoughts on explicit user feedback *in situ* in the context of their work. For example, we asked what kinds of user feedback are particularly useful to them and, if applicable to their experience, we asked specifically about collecting feedback through conversational voice interactions. All interviews were transcribed and then analyzed by two authors through inductive coding and clustering to identify common categories of scenarios in which such feedback would be useful to machine learning practitioners.

<sup>35</sup>Please refer to the supplementary material for the interview protocol.

### 8.3.1 Usage Categories

Through our interviews, we identified four usage categories that represent a broad range of scenarios in which explicit in situ user feedback through conversational voice interactions could benefit the underlying machine learning models or recommender systems: (1) Clarifying user input, (2) Clarifying behavioral signals, (3) Collecting feature feedback, and (4) Understanding user context. In the following, we will briefly describe these categories.

**Clarifying user input** This category describes scenarios where the user issues a query to the system but the query is too ambiguous for the system to make a decision with high confidence. The voice assistant can then prompt the user for clarification. One machine learning expert (P2) gave the following example,

*“If somebody says … Excuse my vulgarity, but if somebody says, ‘F\*\*\*k you.’ to a voice assistant, are they upset with the voice assistant, or are they looking for a song by CeeLo Green? It’s really tough to just look at just those two words, and be able to say, ‘I am 100% [sure, I] know what that is.’ ”*

In this example, a voice assistant could disambiguate the user input by proactively prompting the user to clarify their request, saying, “Do you want the song by CeeLo Green?” The interviewed machine learning practitioners described this usage category as particularly valuable, as “[user] utterances are short and ambiguous” (P2). The relevance of this category is further backed by prior studies, which found that clarifying ambiguous user input can help deliver content that best matches the user’s needs [78, 114, 333]. Another benefit mentioned by the machine learning practitioners is that such explicit user input in response to the voice assistant’s clarifying question can help create higher quality datasets with less noise. For example, participant (P4) said,

*“Most of the ambiguity could be cleared up I think, if we could just ask users questions. That would really help us understand which rows have labels that are pure noise and which rows have labels that are reliable.”*

**Clarifying behavioral signals.** The second common theme mentioned is about clarifying behavioral signals. This refers to occasions when a user’s behavioral signals — such as duration of user engagement, skipping content, or completing content to the end—might contradict each other. One participant (P5) mentioned,

*“Follow[ing]<sup>36</sup> and play[ing] an artist or podcast] could be contradictory. It seems like people tend to follow something, which one might think means they are really interested in it, but [then] they don’t actually engage with.”*

Nowadays, machine learning systems, especially reinforcement learning systems, heavily rely on users' behavioral signals to reward or punish the system actions and eventually learn user preferences. Here, the voice assistant could proactively ask for explicit feedback when detecting such ambiguity in the behavioral signals. Clarifying contradictory signals could not only help the models to infer user preference with confidence, but it could also help machine learning experts to discover new user patterns [334].

**Collecting feature feedback.** When introducing new features, it is often initially unclear what implicit signals are most meaningful or reliable. For example, one machine learning expert (P2) spoke about having to come up with evaluation metrics before knowing how users will respond to it,

*“... Because you can’t really collect any data from users if there’s no feature to actually collect that data from. So that was the initial challenge that we faced.”*

In cases like these, user researchers are often brought on to conduct user studies. However, our participants and previous studies mentioned that this kind of user research often takes a long time, and the data they provide is retrospective and disconnected from the data used to build the model. Voice assistants might open new opportunities to prompt the targeted collection of feedback on early feature prototypes at the moment of interaction and with a faster turn-around time. Collected feedback could be immediately available to the machine learning experts and directly linked to the relevant interaction logs or model output facilitating the creation of reliable evaluation metrics.

**Understanding user context.** Lastly, our participants reported that by taking the user's context into account, many user models and recommender systems could be improved. For example, in the music context, the activity a user is currently engaged in, or whether they are with other people at the moment, can strongly affect what they might want to listen to [335]. One expert (P12) states,

*“I think [collecting user context] is pretty challenging because where our user is right now, whether people are with them or not, is really difficult to figure out, but at the same time, very important.”*

---

<sup>36</sup>Following in this context means subscribing to notifications about an artist's or a podcast show's new releases.

Our participants also raised privacy concerns on collecting user context, as expressed by P6:

*“I think a lot of the hurdles that we face on that front are more around privacy than anything else, which I greatly respect. ”*

Thus, the system should acquire a user’s consent prior to collecting such data. We also learned from our interviews that some contextual data, such as mood or the social setting of a user, can be technologically difficult to capture and thus relies on users’ self-reporting. Such challenges were also mentioned in previous literature [88, 331]. Eliciting explicit feedback from the user could help alleviate both the privacy as well as technological challenges. We believe that if voice assistants directly ask the user for such information, it might improve the transparency about what data is gathered, implicitly get the user’s consent, and facilitate context detection that might be technologically challenging or inherently subjective.

## 8.4 SCENARIO-BASED ONLINE STUDY

Our online study directly addresses RQ8.2, making use of scenarios we generated for the identified usage categories to understand how certain design decisions affect users’ willingness to respond and the perceived friction to the experience. We chose an online scenario-based survey for several reasons: first, an online survey allows us to reach a large and diverse participant pool. Second, with a large group of participants, we can compare many design outcomes in multiple scenarios with sufficient statistical power. Third, presenting the scenarios to participants as text-only protects us from unwanted confounding individual effects of the assistant voice itself, which is known to be a major factor [336]. Lastly, online scenario-based studies have been successfully used to investigate user perceptions in other contexts [337, 338].

### 8.4.1 Scenario Design

For each of the scenario categories, we generated scenarios based on examples mentioned by our participants in the interview study (see Tab 8.1). Overall, we included 9 scenarios: 3 for Clarifying user input, 2 for Clarifying behavioral signals, 2 for Collecting feature feedback, and 2 for Understanding user context. Each of the scenario descriptions consists of three parts: setting the interaction context, the interaction itself, and a prompt by the voice assistant.

While our scenario generation process made sure that the scenarios were realistic from a machine learning engineer perspective, we wanted to ensure the same from the user perspective.

Category	Example Scenario	Voice Assistant Prompt Variations
Clarifying user input	You say to your assistant: “Hey VA, play Thriller.”	[Question] Did you want the song Thriller, or the album Thriller? [Instruction] Okay, here’s the song Thriller by Michael Jackson. If you wanted the whole album, say, “play the album Thriller.”
Clarifying behavioral signals	<p>You say to your voice assistant: “Hey VA, play me some Beatles music.”</p> <p>The voice assistant plays “Hey Jude” by the Beatles, which you’ve listened to many times in the past.</p> <p>You really like the song, but this time you want to try something else. So you say: “Hey VA, skip this song.”</p>	<p>[Question] Oh okay, did I not play what you wanted?</p> <p>[Instruction] Sure, here’s another Beatles song. If you want to tell me why you skipped ‘Hey Jude’, say something like, “Hey VA, I skipped because I wanted to hear something new.”</p>
Collecting feature feedback	<p>You feel like hearing some music you’ve never heard before.</p> <p>You say to your voice assistant: “Hey VA, recommend some music.”</p> <p>Your voice assistant says: “Okay, check out this new release by Arianna Grande, an artist who you have listened to a lot in the past.” Then it plays that song.</p>	<p>[Question] After the song ends, the voice assistant says, “What did you think? Was hearing my explanation before the song useful?”</p> <p>[Instruction] After the song ends, the voice assistant says, “If you want to tell me if my explanation before the song was useful, say ‘Hey VA, I have some feedback.’ ”</p>
Understanding user context	<p>You’re feeling like you want some music but not sure exactly what to play.</p> <p>You say to your voice assistant: “Hey VA, play some music I’d like.”</p>	<p>[Question] Sure, I can give you better suggestions if you let me know your mood. How are you feeling at the moment?</p> <p>[Instruction] Sure, I can find something based on your mood if you want. You can say things like ‘I’m happy’ or ‘I’m feeling kind of down.’</p>

Table 8.1: Overview of the four usage categories with an example scenario for each and two voice assistant prompt variations.

Thus, inspired by Gabriele et al. [337], we introduced a validity check in our study, asking participants to rate how plausible and how likely to happen the described scenario is, each on a 3-point Likert Scale (1-3). Our participants considered all scenarios to be plausible ( $M = 2.47$ ,  $SD = 0.73$ ) and likely to happen in their daily interactions with a voice assistant ( $M = 2.46$ ,  $SD = 0.73$ ), without any significant difference among the four categories. This suggests that users perceived our scenarios to be realistic.

#### 8.4.2 Voice Assistant Design

Each participant was assigned to one voice assistant design condition. In this section, we describe the design decisions we wanted to investigate in our study and how we modified our scenarios to do so.

**Framing** Previous research has shown that people attribute human-like interpersonal behaviors to AI agents [155]. There is evidence that in human-AI interaction, the personality or the metaphor of an intelligent agent affects how people perceive and interact with the agent [158, 159]. Thus, we wanted to investigate the role of a voice assistant’s framing on users’ perception of user feedback elicitation. For that, we introduced the voice assistant in three different framings: (1) as an Assistant, (2) as a Learner, and (3) as a Collaborator.

The Assistant condition represents how voice assistants are typically presented – framed as virtual helpers. The Learner condition was designed to signal the limited initial capabilities of a voice assistant, as prior studies showed that people are more friendly to a virtual agent when they know the virtual agent is imperfect [158]. Lastly, in the Collaborator condition, our framing echos prior studies where framing human-AI interactions as collaborative tasks or creating shared goals would induce positive attitudes toward the virtual agent and lead to better outcomes [339]. Before reading the scenarios, each participant was primed with an introductory sentence that clearly highlighted one framing as shown in Table 8.2.

**Elicitation Strategy** Elicitation Strategy is a key component when collecting high-quality explicit user feedback *in situ*. Different feedback prompts may induce different levels of friction, which ultimately affects the quality of user feedback. Multiple studies have indicated that virtual assistants should be adaptive in terms of how they pose questions or requests [115, 228]. In our study, we considered two elicitation strategies: (1) directly asking the question, and (2) providing instructions for the user to provide feedback to the voice assistant. We created two variations for each interaction scenario. For example, when getting ambiguous user requests like “Play Thriller”, the voice assistant could ask directly:

Framing	Voice Assistant Introduction
Assistant	Imagine you have a voice assistant that helps you with your daily tasks.
Learner	Imagine you have a voice assistant that's always learning how to better help you with your daily tasks.
Collaborator	Imagine you have a voice assistant that collaboratively works with you on your daily tasks.

Table 8.2: Introductions for the three different voice assistant framings in our study. We examined three framings, Assistant, Learner, and Collaborator, which indicate different levels of capabilities.

*“Did you want the song Thriller or the album Thriller?”*

or it can provide instructions

*“Okay, here’s the song Thriller by Michael Jackson. If you wanted the whole album, say ‘play the album Thriller.’ ”*

**Level of Proactivity** Each scenario in our study is further characterized by its Level of Proactivity. While in all our scenarios the voice assistant initiates the prompt to elicit the feedback, in some this prompt is directly connected to a user input that preceded it, while others are more removed from a specific user interaction. In many scenarios, this is less of a design choice, but a result of the nature of the interaction.

To understand the effects of these different circumstances, we defined three levels of proactivity: (1) A low level of proactivity describes scenarios where the voice assistant’s question directly follows a user’s command and asks for information that is closely related to the current session. For example, the voice assistant is asking for clarification when the user’s initial utterance is ambiguous, as in the “Thriller” example described above. (2) A middle level of proactivity describes scenarios where the voice assistant’s prompt is not a direct response to a user’s request to the assistant, for example when the user skips multiple songs or “likes” a song. Finally, (3) a high level of proactivity describes scenarios where the voice assistant initiates a new session with a prompt that is not triggered by the user’s interaction, for example asking the user if they liked a song once it has finished playing.

Based on the voice assistant’s Framing and Elicitation Strategy, we yielded a  $3 \times 2$  experiment design with six design variations.

### 8.4.3 Measures

To understand the participants' perception of the voice assistant's prompt for explicit feedback, we measured two major aspects: willingness to respond and friction to the user experience.

**Willingness** We assessed participants' willingness to respond to the voice assistant's prompt for explicit feedback. Studies have shown that the strength of a person's willingness to act indicates the likelihood of actual behavior, such as adopting new technologies [340] and providing feedback [341]. In our case, we want to know how people may respond to the feedback request from a voice assistant with different designs. Based on previous studies that measured people's behavioral intention [342], we also used a 4-point Likert scale and asked a similar question, "How would you react to what the voice assistant said?" The 4-point scale ranged from "It's extremely likely that I would say nothing."(1) to "It's extremely likely that I would say something back to the assistant."(4) The question directly measures participants' willingness to respond to the voice assistant's prompt. As recommended by [343], the item is clearly worded, free of jargon and easy to understand. To ensure that the study participant properly read and understand the scenario and the voice assistant's prompt, we asked a follow-up question, "If you were to say something back to the voice assistant, what would you say?"

**Friction** We then evaluated the friction that such voice assistant prompts would introduce to a user's experience. The user perceived experience friction indicates the cost of a voice assistant's feedback request. Assistant design that can induce lower friction may collect better feedback quality and achieve better user retention [173]. Also, previous work suggests a potential link between a user's perceived friction and the assistant's level of proactivity [344], which we also considered in our voice assistant's design. We adapted three 5-point scale questions used by prior studies to the context of our scenarios. Those questions have been shown to be effective indicators of friction [345]. First, we asked directly about the perceived experience friction: "How disruptive to your overall experience is the voice assistant speaking at that moment?" Second, we assessed the friction indirectly by asking for the desired frequency of such interactions: "In a scenario like this one, how often do you think a voice assistant should respond the way it did?" And third, we assessed the perceived cognitive load: "How hard was it for you to figure out what to say in your response to the voice assistant?" All three questions are 5-point Likert Scale questions, ranging from 1 (Not disruptive at all/ Never/ Extremely easy) to 5 (Extremely disruptive/ Always/ Extremely

difficult).

**Demographic Measures and Individual Characteristics** As prior literature suggests that certain demographic information and pre-existing attitude correlate with people’s perception of technology [191, 192] and their experience with voice assistants specifically [159], we also collected basic demographic information, participants’ personalities, and their attitude to and experience with voice assistants.

The basic demographic information included age, gender, education level, and annual household income. Participants’ personalities were measured by TIPI, a well-examined 10-item personality scale [346] where participants rate ten sentences about their personality on 7-point Likert scales.

To assess the participants’ prior voice assistant experience, we asked them what voice assistants and devices they use or have used in the past. They further self-reported their interaction frequency with their voice assistant(s) on a scale from “Multiple times per day” to “Never”. Based on the Technology Acceptance Model [347], we collected participants’ existing attitude towards voice assistants from the perspective of Usefulness and Satisfaction on 5-point Likert scales (from -2 to 2). To also assess their trust towards voice assistants, we added a single-item scale, similar to [159].

#### 8.4.4 Study Procedure and Participant Recruiting

Our study had four sections: In the first section, we asked our participants about their prior voice assistant experience and their attitude towards voice assistants (see Section 8.4.3). For the second section, each participant was randomly assigned to one of the six potential combinations of three different framings (Assistant, Learner, or Collaborator) and two elicitation strategies (asking questions or providing instructions for giving feedback). Participants were asked to read the Framing, a short paragraph of text as described earlier in Section 8.2. To ensure that participants read the passage, we asked them to rephrase how they would describe a voice assistant and used it as an attention checker. In the third section, we presented each participant with four randomly selected scenarios from the generated scenario pool (see Table 8.1), one from each of the four usage categories. For each participant, the voice assistant followed the same elicitation strategy in all scenarios. After each scenario, we asked the participants to answer a short questionnaire about their reaction and perception of the voice assistant’s prompt. In the last section, the participants were asked to answer a series of demographic questions.

In the survey, participants read a total of four scenarios, one after another and for each scenario, they answered a short series of eight questions.

Before deploying our survey to a broader audience, we ran a pilot study with 30 participants recruited from Appen with the same criteria as the actual study (described below). In the pilot study, we explicitly asked our participants about the clarity and validity of the generated scenarios, and about confusion around the experiment procedure and question wording. We then revised our study based on participants' feedback.

For the actual study, we recruited our participants from Appen<sup>37</sup>. This platform has been widely used to gather human intelligence in AI research and social science experiments [348, 349]. We sent out our survey in five batches over the course of a week: three during weekdays and two on the weekend to recruit a larger variety of participants. Participants were paid \$12.5/hr regardless of their performance and spent approximately 20 minutes on the survey ( $M = 19.03$  mins,  $SD = 18.31$  mins). Since our study materials were written in English, our study targeted participants only from English-speaking countries<sup>38</sup>. To ensure quality, we made our study limited to people who are qualified as high-quality workers based on their history with Appen<sup>39</sup>.

#### 8.4.5 Limitations

The goal of this study is to explore the design space for eliciting explicit user feedback using voice assistants. Given the inevitable cost to the user experience when prompting for feedback, we wanted to start with scenarios that would bring the most benefits to machine learning practitioners. Using a scenario-based study design allowed us to evaluate multiple design dimensions effectively and gather input on which ones to pursue further, e.g., as prototypes. However, the hypothetical nature of the scenario-based design also introduces some limitations that should be addressed in future work. First, although our validity check showed that all scenarios in the study are perceived to be plausible and likely to happen for the participants, it still requires them to imagine being in that situation. Second, we measured users' self-reported willingness to react and the perceived disruptiveness to the experience instead of their actual behavior. Thus, as a next step, we would develop prototypes to evaluate the scenarios as tangible experiences in the field and measure users' more natural behavior directly.

The generalizability of our study is also limited by the scenario design focusing on one

---

<sup>37</sup>[appen.com/](http://appen.com/)

<sup>38</sup>[en.wikipedia.org/wiki/List\\_of\\_territorial\\_entities\\_where\\_English\\_is\\_an\\_official\\_language](https://en.wikipedia.org/wiki/List_of_territorial_entities_where_English_is_an_official_language)

<sup>39</sup>[success.append.com/hc/en-us/articles/203219195-Guide-To-Contributors-Channels-Page](https://success.append.com/hc/en-us/articles/203219195-Guide-To-Contributors-Channels-Page)

	Question	Instruction	Total
Assistant Condition	48	48	96
Learner Condition	57	50	107
Collaborator Condition	45	46	91
Total	150	144	294

Table 8.3: Breakdown of the conditions our participants were assigned to based on a combination of Framing and Elicitation Strategy.

particular application domain, i.e., music. Voice assistant users hold different privacy norms across usage domains [350]. This means that in other application domains of recommender systems, such as health [351] or finance [352], people might consider a voice assistant’s feedback request more as a threat to their privacy. Although one of our study scenarios asked for sensitive data such as the user’s context and their emotional state [353], it will require further investigation to understand how people perceive and react when the voice assistant is asking for more sensitive information.

Lastly, we want to acknowledge the absence of text-to-speech voice in our study design. The different characteristics of text-to-speech voices can have a strong impact on people’s preferences and behavior [354], which may influence how a user reacts to a request from a voice assistant. However, in a large-scale online study, it is challenging to make our voice assistant’s voice adaptive to every participant’s preference. Any mismatch in our voice choice with the participant’s preference may create unwanted effects that undermine our results. Therefore, in the future, it is important to understand people’s reactions to voice assistants’ feedback requests in a real-world setting where the participant is using their own voice assistant. Additionally, it is important to understand how different text-to-speech voice choices may affect an assistant’s feedback elicitation and the interplay with the design dimensions examined in our study.

## 8.5 RESULTS

Our results suggest that the choice of Framing and Elicitation Strategy impacts the user’s willingness to respond and the perceived disruptiveness, while the Level of Proactivity moderates these effects. In this section, we will first provide an overview of the data we collected, and then present our findings in detail.

### 8.5.1 Analysis

We analyzed our data using linear mixed effects method for its robustness in modeling non-independent data. Since each participant in our study responded to four scenarios with the same set of questions, we treated each participant in the linear mixed effects model as a random effect [355], and thus each participant’s response to each scenario is a single data point in our analysis. This results in 1,176 (294 participants  $\times$  4 scenarios per participant) data points in our dataset.

We implemented Satterthwaite’s method in our linear mixed effects model to estimate the effective degrees of freedom [355]. We tested assumptions including linearity, homogeneity of variance, and the normality of residuals. All p-values reported in the results are adjusted via the Bonferroni correction.

We treated each measure as a dependent variable. For each scenario, the independent variables were the Framing of the voice assistant, its Elicitation Strategy, and the scenario’s Level of Proactivity. We also included participant’s demographic, personality, and prior attitude towards voice assistants as covariates to control for potential confounding effects suggested by prior literature [159, 191, 356].

### 8.5.2 Participant Overview

Out of the 356 participants we recruited, 294 completed the study and passed our attention check question. Our analysis is based on those 294 valid responses. Participants were randomly assigned to one of the six conditions (see Table 8.3). Among those 294 participants, 126 identified as women, 145 identified as men, and 23 preferred not to disclose. The median household income was \$25,000 - \$50,000. The median education level was a Bachelor’s degree. The participants’ ages ranged from 18-60 years old ( $M = 28.01$ ,  $SD = 6.61$ ).

All 294 participants currently owned or used to own at least one voice assistant (e.g., Google Home/Google Assistant, Amazon’s Echo/Alexa, Apple’s HomePod/Siri, Microsoft’s Cortana, Samsung’s Bixby). 62.59 % of the participants indicated that they interact with voice assistants at least once per day. 5.44 % of the participants said they had no recent interaction with a voice assistant. The results from the adapted Technology Acceptance Model questions showed that our participants consider voice assistants generally to be Useful ( $M=1.10$ ,  $SD=0.74$ ), Satisfying ( $M=0.60$ ,  $SD=0.47$ ), and Trustworthy ( $M=1.00$ ,  $SD=1.01$ ).

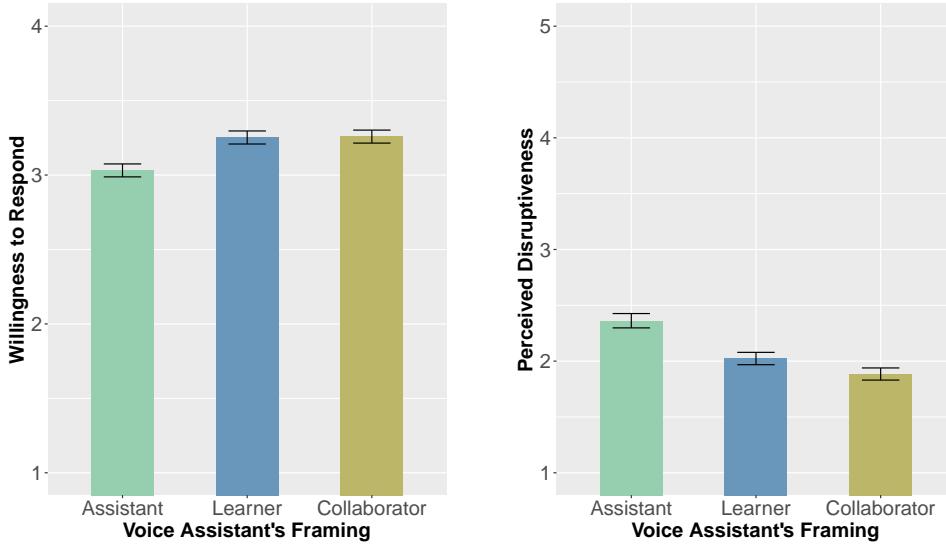


Figure 8.2: Differences in participants' willingness to respond (4-point Likert scale) and perceived disruptiveness (5-point Likert scale) for different voice assistant Framings. Overall, people were more willing to respond to a voice assistant's feedback request when it is framed as a learner or a collaborator. And people perceived a voice assistant's feedback request as less disruptive when it is framed as a learner or a collaborator.

### 8.5.3 Higher Willingness to Respond to a Learning or Collaborating Voice Assistant

We found a significant main effect of the voice assistant's Framing: participants were more willing to respond when the voice assistant was introduced as a Learner ( $M = 3.25$ ,  $SD = 0.91$ ) compared to when the voice assistant was introduced as an Assistant ( $M = 3.03$ ,  $SD = 0.85$ ;  $\beta = 0.23$ ,  $SE = 0.09$ ,  $t = 2.50$ ,  $p < 0.05^*$ ). Meanwhile, participants were also more willing to respond when the voice assistant was introduced as a Collaborator ( $M = 3.26$ ,  $SD = 0.83$ ), compared to when the voice assistant was introduced as an Assistant ( $M = 3.03$ ,  $SD = 0.85$ ;  $\beta = 0.28$ ,  $SE = 0.09$ ,  $t = 2.90$ ,  $p < 0.01^{**}$ ). Our model showed no significant difference between the Collaborator framing and the Learner framing. We found no significant effect of the Elicitation Strategy and the Level of Proactivity on the willingness to respond.

We also looked at the effect size of the Framing on participant's willingness to respond. We found a small effect with a Cohen's  $d$  of 0.38 between the Learner and Assistant framing, and a Cohen's  $d$  of 0.39 between Collaborator and Assistant. Figure 8.2 shows the differences in participants' willingness to respond between the different conditions.

#### 8.5.4 Less Perceived Disruption by a Collaborating or Learning Assistant

We found a significant main effect of the voice assistant's Framing on the perceived disruptiveness. Participants perceived the voice assistant to be less disruptive when it was framed as a Learner ( $M = 2.02$ ,  $SD = 1.14$ ), compared to when the voice assistant was framed as an Assistant ( $M = 2.36$ ,  $SD = 1.26$ ;  $\beta = -0.39$ ,  $SE = 0.11$ ,  $t = -3.56$ ,  $p < 0.01^{***}$ ). Furthermore, participants perceived the voice assistant to be even less disruptive when it was framed as a Collaborator ( $M = 1.89$ ,  $SD = 1.04$ ;  $\beta = -0.44$ ,  $SE = 0.12$ ,  $t = -3.76$ ,  $p < 0.01^{***}$ ), compared to the Assistant condition ( $M = 2.36$ ,  $SD = 1.26$ ). However, we found no significant difference between the Learner condition and the Collaborator condition.

We found a small effect with a Cohen's  $d$  of 0.39 between Learner and Assistant framing, and a medium effect with a Cohen's  $d$  of 0.57 between Collaborator and Assistant. Figure 8.2 shows the differences in perceived disruptiveness between the different conditions. There were no significant effects of the voice assistant's Framing on the other two measures of friction, acceptable frequency and perceived cognitive load to answer.

The results are consistent with prior research on the CASA paradigm [155], where people responded to computer systems as though the computers were social entities. When the voice assistant is framed as a Learner, it signals initially limited capabilities and the agency for improvement. Similar to what has been shown in [158], signaling the incompetence of artificial intelligence makes people more tolerant to its shortcomings. In our case, the voice assistant signaling incompetence by being introduced as a Learner, might make users feel like teachers, and as such feel more inclined to provide feedback to their "student"—the voice assistant. Moreover, signaling the agency to improve may make the user feel that their feedback will ultimately enhance their experience, which increases the voice assistant's chance to collect feedback without disrupting much.

The Collaborator framing emphasizes a shared goal of creating a better user experience for the user. Such a shared goal can make users more cooperative with an agent's requests [339]. This, in turn, might have led to the observed higher willingness to respond.

#### 8.5.5 Instructions were Perceived as Less Disruptive than Questions, without a Negative Impact on the Willingness to Respond

We found a significant main effect of the voice assistant's Elicitation Strategy: people perceived the voice assistant to be less disruptive when the voice assistant gives instructions on how to provide feedback ( $M = 1.93$ ,  $SD = 1.25$ ), compared to when the voice assistant directly asks a question ( $M = 2.24$ ,  $SD = 1.06$ ;  $\beta = -0.40$ ,  $SE = 0.09$ ,  $t = -4.34$ ,  $p < 0.01^{**}$ ). However, we observed no significant effect of the Elicitation Strategy on the willingness to

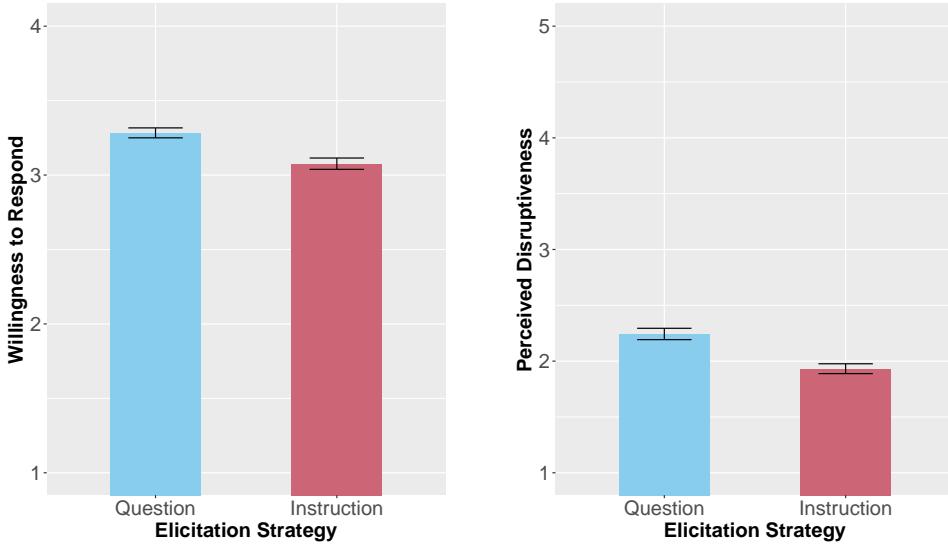


Figure 8.3: Differences in participants' willingness to respond (4-point Likert Scale) and the perceived disruptiveness (5-point Likert Scale) for different Elicitation Strategies in our scenarios. People perceived the voice assistant to be less disruptive when the voice assistant gives instructions instead of direct question. However, we found no significant effect of the Elicitation Strategy on the willingness to respond.

respond, the acceptable frequency, or the perceived cognitive load to answer.

We found a small effect with Cohen's  $d = 0.37$  between the question and instruction Elicitation Strategy on the perceived disruptiveness. Figure 8.3 shows the differences in participants' willingness to respond and the perceived disruptiveness for the different strategies.

In our study, instructions by the voice assistant tend to be longer ( $M_{Instruction} = 20.56$  words  $SD_{Instruction} = 6.65$  words;  $M_{Question} = 14.78$  words;  $SD_{Question} = 7.10$  words), although the difference is not significant ( $t = 1.83$ ,  $p = 0.09$ ). Intuitively, longer requests plus the additional effort of the user having to provide the feedback, seem to create more experience friction. However, our results suggest the opposite. Despite these results, further studies are needed to learn how the length of a voice assistant's request for feedback affects people's perceptions irrespective of the elicitation strategy.

A potential explanation for the lower disruptiveness of instructions is that it might feel less forced and more polite to be given the choice to provide feedback. However, with a choice, there is always the risk that the user will not respond. However, in our results, we did not observe a significant difference in users' willingness to respond. Therefore, providing instructions may elicit a similar amount of user feedback while avoiding too much perceived disruptiveness.

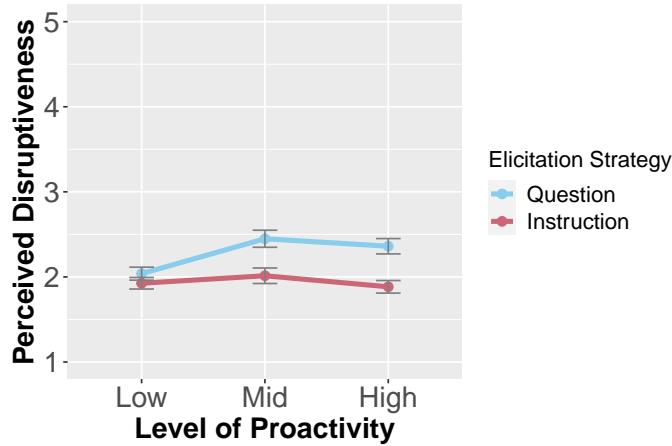


Figure 8.4: Impact of Level of Proactivity on perceived disruptiveness for different Elicitation Strategies in our scenarios. Instruction shows a stronger effect on perceived disruptiveness with higher levels of proactivity.

### 8.5.6 Elicitation Strategies Became more Important with Higher Levels of Proactivity

As defined in the Study Design Section 8.4.2, each scenario we tested represented one of three levels of a voice assistant's proactivity. We found that although higher Levels of Proactivity reduced participants' willingness to respond (Low:  $M = 3.30$ ,  $SD = 0.81$ ; Mid:  $M = 3.10$ ,  $SD = 0.88$ ; High:  $M = 3.10$ ,  $SD = 0.92$ ) and increased the perceived disruptiveness (Low:  $M = 1.99$ ,  $SD = 1.11$ ; Mid:  $M = 2.24$ ,  $SD = 1.20$ ; High:  $M = 2.10$ ,  $SD = 1.17$ ), those differences are not significant (Willingness:  $\beta = -0.31$ ,  $SE = 0.31$ ,  $t = 0.59$ ,  $p = 0.43$ ; Disruptiveness:  $\beta = 0.07$ ,  $SE = 0.08$ ,  $t = 0.94$ ,  $p = 0.35$ ).

We further looked at the interaction effects between Level of Proactivity and Framing, and between Level of Proactivity and Elicitation Strategy. We found a significant interaction effect of the latter on the perceived disruptiveness ( $t = -3.247$ ,  $p < 0.01^{**}$ ): when the Level of Proactivity increases, so does the difference in perceived disruptiveness for the two Elicitation Strategies. Figure 8.4 shows this effect on the difference in the perceived disruptiveness for the two elicitation strategies.

Thus, the Elicitation Strategy plays a more important role when the feedback prompt is separate from an expected response to a user's request to the assistant. Scenarios where the voice assistant is asking for user feedback on a specific feature, or when trying to clarify behavioral signals, are common of the highest level of proactivity. For example, if a user leaves the music on for a long time and then the voice assistant initiates a new conversation to ask if the user is still listening, that would be considered a high level of proactivity. For those cases, our results suggest that instead of asking directly, the voice assistant should strongly consider providing instruction to avoid being too disruptive, e.g.,

*“Hey just checking in. If you’re still listening, say, ‘Hey VA, I’m still listening.’”*

### 8.5.7 A Positive Prior Attitude towards Voice Assistants Decreases the Perceived Disruptiveness

While not part of our main research question, we also found that participants who had a more positive attitude towards the voice assistant prior to the study perceived voice assistant prompts as less disruptive ( $\beta = -0.38$ , SE = 0.09,  $t = -4.32$ ,  $p < 0.01^{**}$ ), and more acceptable to be asked frequently ( $\beta = -0.26$ , SE = 0.11,  $t = 2.29$ ,  $p < 0.05^*$ ). Participants who perceived voice assistants as more useful ( $\beta = -0.30$ , SE = 0.14,  $t = -2.25$ ,  $p < 0.05^*$ ) and trustworthy ( $\beta = -0.12$ , SE = 0.06,  $t = -2.01$ ,  $p < 0.05^*$ ), felt that the voice assistant’s prompts were easier to answer.

## 8.6 DISCUSSION

Our results show that we can influence users’ willingness to respond to voice assistant requests for explicit user feedback and mitigate potential experience friction through intentional design choices. In this section, we discuss design implications and potential future directions for considering voice assistants as mediators between end users and the people building them.

### 8.6.1 Optimizing Elicitation Strategies for Context and Incentive

Prior to deciding on an elicitation strategy for the voice assistant’s prompt, the user’s general context should be considered. Voice assistants are ubiquitous now: people can interact with them on their phones, their headphones, their smart home devices, in their cars, and other places. This ubiquity provides rich opportunities for a voice assistant to collect user feedback in various situations, but also brings a new challenge.

As mentioned in [156], the interaction design of a voice assistant should account for various social settings in which the user’s request and the voice assistant’s response should be appropriate. For example, when the user is in a social setting with others nearby, probing for sensitive data, such as mood, may be inappropriate. People may consider such an interaction as a privacy threat, since other people on the scene may overhear the conversation [357]. To mitigate a user’s privacy concern, one thread of research suggests giving the user control [358, 359]. The system should let the user decide what information to give and when to give it. In our case, this implies that for feedback requests that may contain sensitive data, the

safer design decision would be the elicitation strategy of providing instructions on how to give feedback whenever the user deems the moment more appropriate.

Our results suggest that people perceived the voice assistant’s instructions to provide feedback as less disruptive compared to directly asking questions, without a significant difference in terms of participants’ willingness to respond. However, the expected value of the elicited feedback should be considered for both the machine learning practitioner and for the user. If it is deemed crucial to improve or fix an underlying model, the strategy of asking a targeted question could help to more efficiently collect the most relevant user feedback even if it risks more disruption.

Of course, people are more encouraged to provide feedback if there is an incentive [360]. In some scenarios, the incentive can be implicit; for example, when the voice assistant is asking for input clarification for an ambiguous request, the user will immediately be rewarded with a better result. But when the voice assistant is probing for context, it might not be clear whether and how the collected information will benefit the user. Thus, including information on how it will benefit the user, or even offering other incentives, such as a monetary reward or free services, could likely increase a user’s willingness to respond further.

### 8.6.2 Mediating Collaborations in Recommender systems

Building a recommender system can be described as a collaborative process between the stakeholders who build these systems (e.g., machine learning engineers, designers, and researchers) and millions of end users. Each stakeholder plays a different role in this collaboration. End users interact with recommender systems and leave traces of behavior that can be used by machine learning practitioners to build models. Researchers and designers run user studies to collect feedback about these systems to improve the user experience of these systems. In our study, we found people are willing to respond to a voice assistant’s feedback requests, and perceive the interaction as less disruptive, when the voice assistant is framed as a collaborator. We argue we can further reinforce a voice assistant’s collaborator role by letting a voice assistant connect the users of these systems with those who design and build them.

For example, user researchers and designers can work with machine learning engineers to design the types of voice assistant interactions that can help elicit feedback in a more enjoyable way. Once the feedback is collected, the voice assistant can then communicate the impact of the feedback to the user. This virtuous cycle can highlight users’ contributions in the collaboration, which ultimately encourages them to provide more high-quality explicit feedback.

We can further extend the collaborative role of a voice assistant as a mediator, supporting the mutual interest between the system builders and users in improving the user experience. The voice assistant not only can collect feedback that a system builder needs, but also open the loop for users to advocate for their needs and ideas directly to the system builder. By opening the loop, the voice of a user could be directly heard by the system builder, which encourages broader participation in product design and development.

### 8.6.3 Taking Advantage of the Long-term and Evolving Relationship with an Assistant

An important consideration for the design of voice assistants is that a user’s experience is not shaped by a single interaction [361], and small language cues used by a voice assistant may play key roles in the user’s perception [159]. Based on our results, one specific design implication would be to consistently reinforce the collaborative framing of the assistant. This would mean not only introducing it as a collaborator, but also having the assistant act accordingly in other interactions, using inclusive terminology like “we” and “let’s” wherever appropriate to reinforce a collaborative relationship. Such framing would also be reinforced by other design dimensions of a voice assistant such as voice and physical appearance. For example, the device that hosts the voice assistant could be designed like a student with a child’s voice to reinforce the learner framing.

Although framing a voice assistant as a collaborator or a learner may be effective in eliciting user feedback, we should be careful to manage user expectations between the voice assistant’s actual capability and the chosen framing. The discrepancy between the user’s high expectations and the voice assistant’s limited capabilities may backfire, resulting in a lower quality user experience [160, 161]. In this case, we could take advantage of the evolving relationship and consider a framing with low competence (e.g., learner) and gradually transition to high competence framing, such as a collaborator when the voice assistant becomes more capable.

Another opportunity of users’ repeated engagement with a voice assistant is that it allows for a longitudinal approach to feedback elicitation. Feedback can be collected repeatedly over a longer period of time and lead to insights about changes in a user. This idea came up several times in our interviews, with the machine learning practitioners stressing that explicit *in situ* user feedback is particularly valuable when it captures change over time. Prior work on recommendation models also highlights the importance of incorporating the temporal change of a user’s preference [362]. In such scenarios, the elicitation strategy and its design could adapt over time: as the user becomes more familiar with the ways the voice assistant asks for feedback, its prompts could be adjusted. For example, instructions could become

shorter, or they could let the user know that feedback on this feature would help it improve its capabilities.

Overall, we found that a user’s positive pre-existing attitude towards the voice assistant positively correlates with their willingness to answer a voice assistant’s prompt and negatively correlates with the perceived disruptiveness. This suggests it is most effective to focus on creating the best possible experience in the first interactions with the assistant, and start asking for feedback only later in the relationship, when one can benefit from the positive impact of the earlier experiences. We could also consider a foot-in-the-door approach, in which a voice assistant starts with small, infrequent feedback requests for insensitive information (e.g., clarifying user input) and asks for more (e.g., understanding user context) later.

It also suggests that as voice assistants become more reliable and intuitive, user attitudes towards their capabilities may improve. Thus, we believe in the potential opportunities for explicit feedback elicitation as a core part of the voice user experience.

## 8.7 CONCLUSION

From reviewing related work and interviewing machine learning experts, we derived a set of four different usage categories where eliciting explicit feedback *in situ* on voice assistants would be valuable to improve the underlying user models and recommender systems powering today’s voice assistants. We tested different design approaches, namely the Framing of the voice assistant, its Elicitation Strategy, and the Level of Proactivity in different scenarios, to better understand the influence that such decisions have on users’ willingness to respond and the perceived disruptiveness in their experience. Our findings indicate that framing a voice assistant as a learning or collaborating entity instead of just an assistant can positively affect the users’ perception, and that providing users with instructions on how to give feedback is perceived as less disruptive than asking direct questions. We also discuss design implications of these learnings and lay out future voice assistant design directions. The goal would be a future in which voice assistants could enable a seamless, collaborative process between end users and machine learning practitioners to further improve and personalize the assistants’ services.

## CHAPTER 9: FUTURE DIRECTIONS

### 9.1 INTERACTIVE EVALUATION OF CONVERSATIONAL AI

Today's evaluation processes of NLP models produce gaps in the sociotechnical system, especially evaluating conversational AI where the conversation often has multiple turns and no preset paths [124, 363]. The evaluation metrics often fail to capture the semantic meanings, and the gap between the evaluation goal and task objectives is huge [364]. Therefore, it is important to inject human intelligence into the evaluation metrics and make the evaluation metrics could be tailored toward different real-world task objectives. In this thesis, we have proposed new metrics to evaluate a conversational agent's ability to conduct surveys (Chapter 3) and knowledge-augmented metrics for evaluating follow-up questions in an AI-driven conventional survey (Chapter 6). In the future, we believe it is important to borrow the rich literature in behavioral science and linguistic research to design and validate new evaluation methods that could incorporate human intelligence and tailor it to real-world scenarios.

For example, similar to developing a reliable program, a good AI agent also needs extensive testing and auditing. As more generative models have been incorporated into the development process, on the one hand, the increased language understanding and generative allows an individual to chat almost freely with natural language. On the other hand, both non-transparency in large language models and the complex and immense interaction space create additional layers of challenge for evaluation. How could we systematically explore whether a conversational AI can fulfill people's needs in downstream contexts while delivering positive experiences? In our future study, one idea is to study the simulated worlds to evaluate an AI agent based on real-world scenarios. Such a sandbox could create a safe environment for testing to make the evaluation process more proactive than reactive, e.g., responding to inappropriate actions in a real-world deployment. And most importantly, it could provide an interactive interface for non-AI experts to qualitatively examine the effectiveness of their agents. Prior works have tried to create simulated environments for human-human interactions [365, 366]. However, those models only account for limited social signals, especially through natural languages. Also, when building simulated environments for evaluation, we need to pay attention to the biases in the simulated world, because the simulated environments are often built upon existing datasets or past data that may contain biased or harmful views. Therefore, it is another challenge yet an important aspect. We must ensure the evaluation process reflex people's voices and values fairly. Otherwise, the biases in the evaluation process will cascade to real-world users.

## 9.2 POWER-AWARE INTERACTION

We are living in a world of power. Our power status, e.g., social structural position, access to material resources, expertise, etc., shapes our goals and behaviors in organizational settings [65]. For example, high-power individuals could strongly influence others' actions, force others to compromise, and achieve their own goals. The power status also affects one's psychological power [367]. Feeling agency and control benefits an empowered individual's efficacy and mental health status. Therefore, strategically signaling and managing the power dynamics between individuals could benefit the outcome of collaborations, negotiations, and competitions.

Power dynamics also play a role in human-AI interactions. In Chapter 5, we learned a conversational agent could moderate the power dynamic between the researcher and the participants to improve study quality. And recent studies showed that the power difference in human-AI teams could change how the individual perceives the agent's ability, responds to the agent's request, and reacts to the agent's mistake [368, 369]. We believe a power-aware agent could signal its power status to steer the interaction for good or empower the person to take responsibility and lead the interaction. For example, when an agent is uncertain about a novel task, they could use low-power language to put the person at a higher power status to take more control for team success. However, little work has explored how power dynamics manifest in human-AI interaction and how an agent could signal its own power status, let alone enable an agent to actively manage the power dynamic during the interaction to achieve optimal outcomes.

As today's AI becomes more capable, more human-AI teams are formed to help people achieve their goals and make high-stake decisions. Given the potential benefit of a power-aware agent in human-AI interaction, in the future, our goals are analyzing human-AI interaction through the lens of power, modeling power dynamics to inform agent's conversation strategies, and building evaluation metrics for power-dynamic in the context of human-AI interaction.

## 9.3 UNDERSTANDING HUMAN BEHAVIOR THROUGH EVERYDAY AI

A longitudinal study reveals the rapid fluctuations in behaviors, thoughts, and emotions from moment to moment or day to day [370, 371]. However, it is expensive to keep participants engaged over a long period. The ubiquity of conversational agents, such as conversational AIs in personal devices, smart speakers, and websites, opens opportunities to intervene and collect data through daily interaction. In this thesis, we showed the promise of using conversational AIs to understand human behavior. In Chapter 8, we showed people

are open to voice assistants' information requests with proper design. In the future, we plan to further leverage everyday AIs for human understanding.

However, today's conversational agents are mostly designed for single-shot interactions. The conversational agent will face more challenges in conducting effective longitudinal studies: How can an agent adapt and react to unique individual experiences? How can an agent build and manage a long-term relationship with people? In the future, we will explore affective computing and knowledge-driven models to create agents capable of delivering engaging adaptive longitudinal studies. Such an AI agent will help researchers study individual behaviors across the lifespan, at scale, and at low cost.

The ubiquity of every AI offers another unique opportunity, information collection *in situ*. Asking the right question at the right time ensures data diagnosticity, enhances external validity, and reduces redundant data [86, 87, 91]. In the future, we want to explore sampling techniques beyond location, time, and activity in the near future, i.e., Could we sample social context and interactions? Based on this thesis, we built a novel behavioral research platform, *Petri*, that combines mobile sensing and conversational agents to sense social contexts. We plan to develop sampling methods with few-shot learning to query the participants while immersed in real-life situational contexts accurately. The ability to probe into diverse and complex social interactions will empower behavioral scientists to dissect human behavior at a new level of granularity.

Ubiquitous technology grants us access to a large variety of data to learn an individual's emotional state, health, daily activities and so on [88]. However, it poses new threats to a participant's privacy [372]. How could we protect participants from potential privacy risks? It is crucial to consider people's agency and control when designing such novel data collection methods. To probe into this future topic, we plan to look at two processes, *informed consent*, how AI could help participants make informed decisions before joining the study, and *control over data*, how we should design a system that grants participant's control over their own data with the consideration of academic rigor and unwanted data biases. In Chapter 5, we took the preliminary step and studied how AI could be designed to facilitate informed consent. We will expand our work further for border contexts. For privacy, we need to carefully understand the trade-off between in-situ data collection and protecting people's privacy. To do this, we must understand people's needs, invite them into the design process, and co-design the potential solution.

## 9.4 AI FOR SOCIAL SCIENCE

In this thesis, we conducted a series of studies focused on the uses of conversational AI for information collection and informed consent, core research procedures in social science. Expanding on prior related work of using conversational agents for collecting information [102, 106], building psychological profiles [19], delivering interventions [373], and instructing study procedures [374], we did see a future opportunity to study the board use of AI for social science discovery. For example, a virtual research assistant could help researchers reach a worldwide population (if necessary for the study), engage with the participant and build rapport, deliver the intervention, collect high-quality data, and debrief the participants. Such a virtual research assistant could be especially helpful for longitudinal field studies where keeping participants engaged and collecting high-quality data over a long period of time is particularly difficult and expensive.

However, creating such a virtual research assistant is challenging. First, repeated interactions with a virtual research assistant at different stages of a study pose new challenges to interaction design. The agent needs to adapt and react to unique individual experiences over time. Second, personalization is a double-edged sword in the context of human subject research. Although it could increase engagement, a highly personalized chatbot could induce unwanted confounding factors due to the inconsistency across participants. Third, mediating the communication between the participants and the researchers requires a new interaction paradigm. The agent needs to mediate the communication and actively engage both parties for study success. Though challenging, such a virtual research assistant could help researchers conduct scalable, robust, and engaging human-subject studies.

Many other domains are deemed valuable to explore the role of AI, including simulated experiments, psychometric validation, hypothesis generation, meta-analysis, social cognition mapping, and cognitive modeling. Building upon this thesis, we hope to work with researchers who are interested in human behavior to further design and build AI to understand ourselves.

## CHAPTER 10: CONCLUSION

Understanding human behavior at a large scale is crucial for addressing some of society's most consequential challenges. By gaining insights into why people act as they do, we can make informed decisions that have a positive societal impact, whether it's improving public health, developing a sustainable economy, or fair education. However, the complexity and variability of human behavior require sophisticated tools and methods to capture cultural, social, environmental, and individual characteristics that influence the observed behavior patterns. Moreover, interdisciplinary expertise and a deep understanding of ethical concerns surrounding informed consent, privacy, and data collection are essential to study human behavior responsibly. In this dissertation, we take up this challenge: we explore the use of Artificial Intelligence (AI) in the context of behavioral science, design and build effective conversational AIs for information collection and informed consent.

We start by focusing on one of the most widely-used research methods in behavioral research, survey. We aimed to study conversational AIs to address three challenges in today's survey research, survey fatigue, inflexible question structure, and lack of personalization. In an AI-driven conversational survey, an agent asks questions, interprets a participant's responses, and probes answers whenever needed. We first studied the response quality and participant engagement of an AI-driven conversational survey by comparing it with form-based ones. After establishing the promise, we improved a conversational survey by equipping the chatbot with active listening skills. We built natural language processing models to enable flexible and personalized survey structures by automatically generating informative follow-up questions on the fly.

We then looked at the ethical practices in behavioral science research, the informed consent procedure in online studies. Current online participants often fail to make an informed decision without the researcher's presence and guidance. We examined the role of an AI-powered chatbot in improving informed consent online by re-introducing interactivity to the process. By comparing the chatbot with form-based interaction, we found the chatbot improved consent form reading, promoted participants' feelings of agency, and closed the power gap between the participant and the researcher.

We end with a series of empirical explorations about how people interact with such conversational AIs. We deployed a conversational AI in the real world to understand students' team preferences to improve the student teaming experience. Then, drawing from the rich use of voice assistants, we consider voice as another modality and study how voice assistants could be used to collect explicit user feedback. This thesis further offered multiple design

implications for future virtual research assistants that can engage with people responsibly and, most importantly, assist researchers in operationalizing their intuitions and curiosities about how humans think and behave.

## REFERENCES

- [1] G. S. Becker, *The economic approach to human behavior*. University of Chicago press, 1976, vol. 803.
- [2] G. T. Gardner and P. C. Stern, *Environmental problems and human behavior*. Allyn and Bacon Boston, 1996.
- [3] S. Mullainathan and R. H. Thaler, “Behavioral economics,” 2000.
- [4] R. B. Cialdini, C. A. Kallgren, and R. R. Reno, “A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior,” in *Advances in experimental social psychology*. Elsevier, 1991, vol. 24, pp. 201–234.
- [5] S. V. Paunonen and M. C. Ashton, “Big five factors and facets and the prediction of behavior.” *Journal of personality and social psychology*, vol. 81, no. 3, p. 524, 2001.
- [6] D. Cohen, R. E. Nisbett, B. F. Bowdle, and N. Schwarz, “Insult, aggression, and the southern culture of honor: An” experimental ethnography.”,” *Journal of personality and social psychology*, vol. 70, no. 5, p. 945, 1996.
- [7] P. Kline, *The new psychometrics: Science, psychology, and measurement*. Psychology Press, 1998.
- [8] G. W. Harrison and J. A. List, “Field experiments,” *Journal of Economic literature*, vol. 42, no. 4, pp. 1009–1055, 2004.
- [9] J. Henrich, S. J. Heine, and A. Norenzayan, “Most people are not weird,” *Nature*, vol. 466, no. 7302, pp. 29–29, 2010.
- [10] R. R. Faden and T. L. Beauchamp, *A history and theory of informed consent*. Oxford University Press, 1986.
- [11] J. Cassell, “Ethical principles for conducting fieldwork,” *American anthropologist*, vol. 82, no. 1, pp. 28–41, 1980.
- [12] S. E. Brennan, “Conversation as direct manipulation: An iconoclastic view,” *The art of human-computer interface design*, pp. 393–404, 1990.
- [13] G. López, L. Quesada, and L. A. Guerrero, “Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces,” in *International Conference on Applied Human Factors and Ergonomics*. Springer, 2017, pp. 241–250.
- [14] J. Grudin and R. Jacques, “Chatbots, humbots, and the quest for artificial general intelligence,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, p. 209.

- [15] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, “A new chatbot for customer service on social media,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 3506–3510.
- [16] J. Schroeder, C. Wilkes, K. Rowan, A. Toledo, A. Paradiso, M. Czerwinski, G. Mark, and M. M. Linehan, “Pocket skills: A conversational mobile web app to support dialectical behavioral therapy,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 398.
- [17] T. Ammari, J. Kaye, J. Y. Tsai, and F. Bentley, “Music, search, and iot: How people (really) use voice assistants,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 26, no. 3, pp. 1–28, 2019.
- [18] J. Hirschberg and C. D. Manning, “Advances in natural language processing,” *Science*, vol. 349, no. 6245, pp. 261–266, 2015.
- [19] M. X. Zhou, G. Mark, J. Li, and H. Yang, “Trusting virtual agents: The effect of personality,” *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 9, no. 2-3, pp. 1–36, 2019.
- [20] D. Traum, “Computational approaches to dialogue,” *The Routledge Handbook of Language and Dialogue*. Taylor & Francis, pp. 143–161, 2017.
- [21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [22] H. Muller, A. Sedley, and E. Ferrall-Nunge, “Survey research in hci,” in *Ways of Knowing in HCI*. Springer, 2014, pp. 229–266.
- [23] J. R. Evans and A. Mathur, “The value of online surveys,” *Internet research*, vol. 15, no. 2, pp. 195–219, 2005.
- [24] R. D. Fricker and M. Schonlau, “Advantages and disadvantages of internet research surveys: Evidence from the literature,” *Field methods*, vol. 14, no. 4, pp. 347–367, 2002.
- [25] S. R. Porter, “Overcoming survey research problems,” *New Directions for Institutional Research*, vol. 2004, no. 121, pp. 1–97, 2004.
- [26] P. Ben-Nun, “Respondent fatigue,” *Encyclopedia of survey research methods*, vol. 2, pp. 742–743, 2008.
- [27] S. R. Porter, M. E. Whitcomb, and W. H. Weitzer, “Multiple surveys of students and survey fatigue,” *New directions for institutional research*, vol. 2004, no. 121, pp. 63–73, 2004.
- [28] Y. Chen, *The Effects of Question Customization on the Quality of an Open-Ended Question*. Nebraska Department of Education, Data, Research, and Evaluation, 2017.

- [29] P. J. Lavrakas, *Encyclopedia of survey research methods*. Sage Publications, 2008.
- [30] E. Riiskjær, J. Ammentorp, and P.-E. Kofoed, “The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective,” *International Journal for Quality in Health Care*, vol. 24, no. 5, pp. 509–516, 2012.
- [31] U. Reja, K. L. Manfreda, V. Hlebec, and V. Vehovar, “Open-ended vs. close-ended questions in web questionnaires,” *Developments in applied statistics*, vol. 19, no. 1, pp. 159–177, 2003.
- [32] F. G. Conrad, M. P. Couper, R. Tourangeau, M. Galesic, and T. Yan, “Interactive feedback can improve the quality of responses in web surveys,” in *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass*, 2005, pp. 3835–3840.
- [33] M. Oudejans and L. M. Christian, “Using interactive features to motivate and probe responses to open-ended questions,” *Social and behavioral research and the internet: Advances in applied methods and research strategies*, pp. 304–332, 2010.
- [34] D. Heerwegh and G. Loosveldt, “Personalizing e-mail contacts: Its influence on web survey response rate and social desirability response bias,” *International Journal of Public Opinion Research*, vol. 19, no. 2, pp. 258–268, 2006.
- [35] D. Behr, L. Kaczmirek, W. Bandilla, and M. Braun, “Asking probing questions in web surveys,” *Journal of Social Science Computer Review*, vol. 30, no. 4, pp. 487–498, 2012.
- [36] E. de Leeuw, J. Hox, and A. Boev, “Handling do-not-know answers: Exploring new approaches in online and mixed-mode surveys,” *Journal of Social Science Computer Review*, vol. 34, no. 1, pp. 116–132, 2015.
- [37] J. D. Smyth, D. A. Dillman, L. M. Christian, and M. McBride, “Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality?” *Public Opinion Quarterly*, vol. 73, no. 2, pp. 325–337, 2009.
- [38] C. Lavelle-Jones, D. J. Byrne, P. Rice, and A. Cuschieri, “Factors affecting quality of informed consent.” *British Medical Journal*, vol. 306, no. 6882, pp. 885–890, 1993.
- [39] C. K. Varnhagen, M. Gushta, J. Daniels, T. C. Peters, N. Parmar, D. Law, R. Hirsch, B. Sadler Takach, and T. Johnson, “How informed is online informed consent?” *Ethics & Behavior*, vol. 15, no. 1, pp. 37–48, 2005.
- [40] J. A. Cummings, J. M. Zagrodney, and T. E. Day, “Impact of open data policies on consent to participate in human subjects research: Discrepancies between participant action and reported concerns,” *PLoS One*, vol. 10, no. 5, p. e0125208, 2015.

- [41] E. K. Perrault and S. A. Nazione, “Informed consent—uninformed participants: shortcomings of online social science consent forms and recommendations for improvement,” *Journal of Empirical Research on Human Research Ethics*, vol. 11, no. 3, pp. 274–280, 2016.
- [42] C. Geier, R. B. Adams, K. M. Mitchell, and B. E. Holtz, “Informed consent for online research—is anybody reading?: Assessing comprehension and individual differences in readings of digital consent forms,” *Journal of Empirical Research on Human Research Ethics*, vol. 16, no. 3, pp. 154–164, 2021.
- [43] E. R. Pedersen, C. Neighbors, J. Tidwell, and T. W. Lostutter, “Do undergraduate student research participants read psychological research consent forms? examining memory effects, condition effects, and individual differences,” *Ethics & Behavior*, vol. 21, no. 4, pp. 332–350, 2011.
- [44] J. Flory and E. Emanuel, “Interventions to improve research participants’ understanding in informed consent for research: a systematic review,” *Jama*, vol. 292, no. 13, pp. 1593–1601, 2004.
- [45] E. De Sutter, D. Zaçe, S. Boccia, M. L. Di Pietro, D. Geerts, P. Borry, I. Huys et al., “Implementation of electronic informed consent in biomedical research and stakeholders’ perspectives: systematic review,” *Journal of medical Internet research*, vol. 22, no. 10, p. e19129, 2020.
- [46] E. E. Anderson, S. B. Newman, and A. K. Matthews, “Improving informed consent: Stakeholder views,” *AJOB empirical bioethics*, vol. 8, no. 3, pp. 178–188, 2017.
- [47] M. C. Rowbotham, J. Astin, K. Greene, and S. R. Cummings, “Interactive informed consent: randomized comparison with paper consents,” *PloS one*, vol. 8, no. 3, p. e58603, 2013.
- [48] B. D. Douglas, E. L. McGorray, and P. J. Ewell, “Some researchers wear yellow pants, but even fewer participants read consent forms: Exploring and improving consent form reading in human subjects research.” *Psychological methods*, vol. 26, no. 1, p. 61, 2021.
- [49] S. D. Halpern, J. H. Karlawish, and J. A. Berlin, “The continuing unethical conduct of underpowered clinical trials,” *Jama*, vol. 288, no. 3, pp. 358–362, 2002.
- [50] G. R. Morrow, “How readable are subject consent forms?” *Jama*, vol. 244, no. 1, pp. 56–58, 1980.
- [51] B. H. Gray, R. A. Cooke, and A. S. Tannenbaum, “Research involving human subjects: The performance of institutional review boards is assessed in this empirical study.” *Science*, vol. 201, no. 4361, pp. 1094–1101, 1978.
- [52] E. Bjørn, P. Rossel, and S. Holm, “Can the written information to research subjects be improved?—an empirical study.” *Journal of Medical Ethics*, vol. 25, no. 3, pp. 263–267, 1999.

- [53] G. M. Dresden and M. A. Levitt, "Modifying a standard industry clinical trial consent form improves patient information retention as part of the informed consent process," *Academic Emergency Medicine*, vol. 8, no. 3, pp. 246–252, 2001.
- [54] E. K. Perrault and S. P. McCulloch, "Concise consent forms appreciated—still not comprehended: Applying revised common rule guidelines in online studies," *Journal of Empirical Research on Human Research Ethics*, vol. 14, no. 4, pp. 299–306, 2019.
- [55] C. A. Coyne, R. Xu, P. Raich, K. Plomer, M. Dignan, L. B. Wenzel, D. Fairclough, T. Habermann, L. Schnell, S. Quella et al., "Randomized, controlled trial of an easy-to-read informed consent statement for clinical trial participation: a study of the eastern cooperative oncology group," *Journal of Clinical Oncology*, vol. 21, no. 5, pp. 836–842, 2003.
- [56] P. Agre, F. A. Campbell, B. D. Goldman, M. L. Boccia, N. Kass, L. B. McCullough, J. F. Merz, S. M. Miller, J. Mintz, B. Rapkin et al., "Improving informed consent: the medium is not the message," *IRB: Ethics & Human Research*, vol. 25, no. 5, pp. S11–S19, 2003.
- [57] J. A. Friedlander, G. S. Loeben, P. K. Finnegan, A. E. Puma, X. Zhang, E. F. De Zoeten, D. A. Piccoli, and P. Mamula, "A novel method to enhance informed consent: a prospective and randomised trial of form-based versus electronic assisted informed consent in paediatric endoscopy," *Journal of medical ethics*, vol. 37, no. 4, pp. 194–200, 2011.
- [58] A. R. Tait, T. Voepel-Lewis, S. J. Chetcuti, C. Brennan-Martinez, and R. Levine, "Enhancing patient understanding of medical procedures: evaluation of an interactive multimedia program with in-line exercises," *International Journal of Medical Informatics*, vol. 83, no. 5, pp. 376–384, 2014.
- [59] B. W. Palmer, N. M. Lanouette, and D. V. Jeste, "Effectiveness of multimedia aids to enhance comprehension during research consent: A systematic review," *Irb*, vol. 34, no. 6, p. 1, 2012.
- [60] T. Bickmore, D. Utami, S. Zhou, C. Sidner, L. Quintiliani, and M. K. Paasche-Orlow, "Automated explanation of research informed consent by virtual agents," in *International Conference on Intelligent Virtual Agents*. Springer, 2015, pp. 260–269.
- [61] M. Balestra, O. Shaer, J. Okerlund, M. Ball, and O. Nov, "The effect of exposure to social annotation on online informed consent beliefs and behavior," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016, pp. 900–912.
- [62] M. M. Knepp, "Using questions to improve informed consent form reading behavior in students," *Ethics & Behavior*, vol. 28, no. 7, pp. 560–577, 2018.

- [63] C. Chen, P.-I. Lee, K. J. Pain, D. Delgado, C. L. Cole, and T. R. Campion Jr, “Replacing paper informed consent with electronic informed consent for research in academic medical centers: a scoping review,” *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 80, 2020.
- [64] S. Zhou, T. Bickmore, M. Paasche-Orlow, and B. Jack, “Agent-user concordance and satisfaction with a virtual hospital discharge nurse,” in *International conference on intelligent virtual agents*. Springer, 2014, pp. 528–541.
- [65] M. Foucault, “The subject and power,” *Critical inquiry*, vol. 8, no. 4, pp. 777–795, 1982.
- [66] O. Karnieli-Miller, R. Strier, and L. Pessach, “Power relations in qualitative research,” *Qualitative health research*, vol. 19, no. 2, pp. 279–289, 2009.
- [67] C. A. Ebbs, “Qualitative research inquiry: Issues of power and ethics,” *Education*, vol. 117, no. 2, pp. 217–223, 1996.
- [68] D. L. Wolf, “Situating feminist dilemmas in fieldwork,” in *Feminist dilemmas in fieldwork*. Routledge, 2018, pp. 1–55.
- [69] S.-H. Chen, “Power relations between the researcher and the researched: An analysis of native and nonnative ethnographic interviews,” *Field methods*, vol. 23, no. 2, pp. 119–135, May 2011.
- [70] Y. S. Lincoln and E. G. Guba, *Naturalistic inquiry*. sage, 1985.
- [71] V. Torres and M. B. Baxter Magolda, “The evolving role of the researcher in constructivist longitudinal studies.” *Journal of College Student Development*, 2002.
- [72] P. H. Schuck, “Rethinking informed consent,” *Yale Law Journal*, pp. 899–959, 1994.
- [73] M. Kustatscher, “Informed consent in school-based ethnography: Using visual magnets to explore participation, power and research relationships,” *International Journal of Child, Youth and Family Studies*, vol. 5, no. 4.1, pp. 686–701, 2014.
- [74] Š. Pero and T. Horváth, “Opinion-driven matrix factorization for rating prediction,” in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2013, pp. 1–13.
- [75] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens: an open architecture for collaborative filtering of netnews,” in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 1994, pp. 175–186.
- [76] N. Tintarev and J. Masthoff, “Designing and evaluating explanations for recommender systems,” in *Recommender systems handbook*. Springer, 2011, pp. 479–510.

- [77] G. Jawaheer, M. Szomszor, and P. Kostkova, “Comparison of implicit and explicit feedback from an online music recommendation service,” in *proceedings of the 1st international workshop on information heterogeneity and fusion in recommender systems*, 2010, pp. 47–51.
- [78] W. Maalej, H.-J. Happel, and A. Rashid, “When users become collaborators: towards continuous and context-aware user input,” in *Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications*, 2009, pp. 981–990.
- [79] L. Zheng, V. Noroozi, and P. S. Yu, “Joint deep modeling of users and items using reviews for recommendation,” in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, pp. 425–434.
- [80] C.-H. Tsai and P. Brusilovsky, “Explaining recommendations in an interactive hybrid social recommender,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 391–396.
- [81] P. Pu, L. Chen, and R. Hu, “A user-centric evaluation framework for recommender systems,” in *Proceedings of the fifth ACM conference on Recommender systems*, 2011, pp. 157–164.
- [82] B. M. Marlin, “Modeling user rating profiles for collaborative filtering,” in *Advances in neural information processing systems*, 2004, pp. 627–634.
- [83] D. O’Sullivan, B. Smyth, and D. Wilson, “Explicit vs implicit profiling-a case-study in electronic programme guides,” in *IJCAI*, 2003, p. 1351.
- [84] D. Nichols, “Implicit rating and filtering.” ERCIM, 1998.
- [85] A. Pommeranz, J. Broekens, P. Wiggers, W.-P. Brinkman, and C. M. Jonker, “Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process,” *User Modeling and User-Adapted Interaction*, vol. 22, no. 4-5, pp. 357–397, 2012.
- [86] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, “Emotionsense: a mobile phones based adaptive platform for experimental social psychology research,” in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, 2010, pp. 281–290.
- [87] M. N. Alexiades, “Ethnobotany in the third millennium: expectations and unresolved issues,” *Delpinoa*, vol. 45, no. 1, pp. 15–28, 2003.
- [88] J. Froehlich, M. Y. Chen, S. Consolvo, B. Harrison, and J. A. Landay, “Myexperience: a system for in situ tracing and capturing of user feedback on mobile phones,” in *Proceedings of the 5th international conference on Mobile systems, applications and services*, 2007, pp. 57–70.

- [89] K. Zheng, D. A. Hanauer, N. Weibel, and Z. Agha, “Computational ethnography: automated and unobtrusive means for collecting data in situ for human–computer interaction evaluation studies,” in *Cognitive informatics for biomedicine*. Springer, 2015, pp. 111–140.
- [90] C. N. Scollon, C.-K. Prieto, and E. Diener, “Experience sampling: promises and pitfalls, strength and weaknesses,” in *Assessing well-being*. Springer, 2009, pp. 157–180.
- [91] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, “Towards a better understanding of context and context-awareness,” in *International symposium on handheld and ubiquitous computing*. Springer, 1999, pp. 304–307.
- [92] V. Pejovic, N. Lathia, C. Mascolo, and M. Musolesi, “Mobile-based experience sampling for behaviour research,” in *Emotions and personality in personalized services*. Springer, 2016, pp. 141–161.
- [93] A. Bachmann, R. Zetsche, A. Schankin, T. Riedel, M. Beigl, M. Reichert, P. Santangelo, and U. Ebner-Priemer, “Esmac: a web-based configurator for context-aware experience sampling apps in ambulatory assessment,” in *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, 2015, pp. 15–18.
- [94] L. Baltrunas, M. Kaminskas, B. Ludwig, O. Moling, F. Ricci, A. Aydin, K.-H. Lüke, and R. Schwaiger, “Incarmusic: Context-aware music recommendations in a car,” in *International conference on electronic commerce and web technologies*. Springer, 2011, pp. 89–100.
- [95] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, “Recommender system application developments: a survey,” *Decision Support Systems*, vol. 74, pp. 12–32, 2015.
- [96] X. Amatriain, J. M. Pujol, and N. Oliver, “I like it... i like it not: Evaluating user ratings noise in recommender systems,” in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2009, pp. 247–258.
- [97] A. Mariakakis, M. Goel, M. T. I. Aumi, S. N. Patel, and J. O. Wobbrock, “Switchback: Using focus and saccade tracking to guide users’ attention for mobile task resumption,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI ’15. New York, NY, USA: Association for Computing Machinery, 2015. [Online]. Available: <https://doi.org/10.1145/2702123.2702539> p. 2953–2962.
- [98] V. Q. Liao, M. M. Hussain, P. Chandar, M. Davis, M. Crasso, D. Wang, M. Muller, S. N. Shami, and W. Geyer, “All work and no play? conversations with a question-and-answer chatbot in the wild,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI’18)*. ACM, New York, NY, USA, vol. 13, 2018.
- [99] A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney, “Autotutor: An intelligent tutoring system with mixed-initiative dialogue,” *IEEE Transactions on Education*, vol. 48, no. 4, pp. 612–618, 2005.

- [100] J. F. Allen, D. K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, “Toward conversational human-computer interaction,” *AI magazine*, vol. 22, no. 4, p. 27, 2001.
- [101] J. Li, M. X. Zhou, H. Yang, and G. Mark, “Confiding in and listening to virtual agents: The effect of personality,” in *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*. ACM, 2017, pp. 275–286.
- [102] A. C. Williams, H. Kaur, G. Mark, A. L. Thompson, S. T. Iqbal, and J. Teevan, “Supporting workplace detachment and reattachment with conversational intelligence,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 88.
- [103] E. Luger and A. Sellen, “Like having a really bad pa: the gulf between user expectation and experience of conversational agents,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 5286–5297.
- [104] L. Sproull, M. Subramani, S. Kiesler, J. H. Walker, and K. Waters, “When the interface is a face,” *Human-computer interaction*, vol. 11, no. 2, pp. 97–124, 1996.
- [105] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. 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*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068.
- [106] E. Tallyn, H. Fried, R. Gianni, A. Isard, and C. Speed, “The ethnobot: Gathering ethnographies in the age of iot,” in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–13.
- [107] F. Ricci, L. Rokach, and B. Shapira, “Recommender systems: introduction and challenges,” in *Recommender systems handbook*. Springer, 2015, pp. 1–34.
- [108] F. Radlinski and N. Craswell, “A theoretical framework for conversational search,” in *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, 2017, pp. 117–126.
- [109] J. R. Trippas, D. Spina, L. Cavedon, H. Joho, and M. Sanderson, “Informing the design of spoken conversational search: Perspective paper,” in *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval*. ACM, 2018, pp. 32–41.
- [110] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, “The atis spoken language systems pilot corpus,” in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.
- [111] S. McGlashan, N. Fraser, N. Gilbert, E. Bilange, P. Heisterkamp, and N. Youd, “Dialogue management for telephone information systems,” in *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, 1992, pp. 245–246.

- [112] J. Cranshaw, E. Elwany, T. Newman, R. Kocielnik, B. Yu, S. Soni, J. Teevan, and A. Monroy-Hernández, “Calendar. help: Designing a workflow-based scheduling agent with humans in the loop,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 2382–2393.
- [113] D. Bohus and A. I. Rudnicky, “The ravenclaw dialog management framework: Architecture and systems,” *Computer Speech & Language*, vol. 23, no. 3, pp. 332–361, 2009.
- [114] J. Kang, K. Condiff, S. Chang, J. A. Konstan, L. Terveen, and F. M. Harper, “Understanding how people use natural language to ask for recommendations,” in *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 2017, pp. 229–237.
- [115] Y. Zhang, Q. V. Liao, and B. Srivastava, “Towards an optimal dialog strategy for information retrieval using both open-and close-ended questions,” in *23rd International Conference on Intelligent User Interfaces*. ACM, 2018, pp. 365–369.
- [116] A. Shamekhi, Q. V. Liao, D. Wang, R. K. Bellamy, and T. Erickson, “Face value? exploring the effects of embodiment for a group facilitation agent,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 391.
- [117] A. Følstad and P. B. Brandtzæg, “Chatbots and the new world of hci,” *interactions*, vol. 24, no. 4, pp. 38–42, 2017.
- [118] T. Bickmore, L. Pfeifer, and D. Schulman, “Relational agents improve engagement and learning in science museum visitors,” in *International Workshop on Intelligent Virtual Agents*. Springer, 2011, pp. 55–67.
- [119] S. S. Sundar and J. Kim, “Machine heuristic: When we trust computers more than humans with our personal information,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, p. 538.
- [120] S. Sannon, B. Stoll, D. DiFranzo, M. Jung, and N. N. Bazarova, “How personification and interactivity influence stress-related disclosures to conversational agents,” in *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2018, pp. 285–288.
- [121] M. McTear, Z. Callejas, and D. Griol, *The conversational interface: Talking to smart devices*. Springer, 2016.
- [122] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, “Paradise: A framework for evaluating spoken dialogue agents,” in *8th Conference of the European Chapter of the Association for Computational Linguistics*, 1997.
- [123] L. Dybkjaer, N. O. Bernsen, and W. Minker, “Evaluation and usability of multimodal spoken language dialogue systems,” *Speech Communication*, vol. 43, no. 1-2, pp. 33–54, 2004.

- [124] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, “How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2122–2132.
- [125] K. S. Hone and R. Graham, “Towards a tool for the subjective assessment of speech system interfaces (sassi),” *Natural Language Engineering*, vol. 6, no. 3-4, pp. 287–303, 2000.
- [126] J. Cassell and T. Bickmore, “External manifestations of trustworthiness in the interface,” *Communications of the ACM*, vol. 43, no. 12, pp. 50–56, 2000.
- [127] T. Bickmore and J. Cassell, “Relational agents: a model and implementation of building user trust,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2001, pp. 396–403.
- [128] D. Novick and I. Gris, “Building rapport between human and eca: A pilot study,” in *International Conference on Human-Computer Interaction*. Springer, 2014, pp. 472–480.
- [129] C. Bartneck, D. Kulic, E. Croft, and S. Zoghbi, “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots,” *International journal of social robotics*, vol. 1, no. 1, pp. 71–81, 2009.
- [130] J. Jiang, A. Hassan Awadallah, R. Jones, U. Ozertem, I. Zitouni, R. Gurunath Kulkarni, and O. Z. Khan, “Automatic online evaluation of intelligent assistants,” in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 506–516.
- [131] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, and M. Cieliebak, “Survey on evaluation methods for dialogue systems,” *Artificial Intelligence Review*, vol. 54, no. 1, pp. 755–810, 2021.
- [132] Y.-T. Yeh, M. Eskenazi, and S. Mehri, “A comprehensive assessment of dialog evaluation metrics,” in *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, 2021, pp. 15–33.
- [133] C. Tao, L. Mou, D. Zhao, and R. Yan, “Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [134] S. Ghazarian, J. Wei, A. Galstyan, and N. Peng, “Better automatic evaluation of open-domain dialogue systems with contextualized embeddings,” in *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 82–89.

- [135] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.
- [136] T. Lan, X.-L. Mao, W. Wei, X. Gao, and H. Huang, “Pone: A novel automatic evaluation metric for open-domain generative dialogue systems,” *ACM Transactions on Information Systems (TOIS)*, vol. 39, no. 1, pp. 1–37, 2020.
- [137] K. Sinha, P. Parthasarathi, J. Wang, R. Lowe, W. L. Hamilton, and J. Pineau, “Learning an unreferencec metric for online dialogue evaluation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2430–2441.
- [138] S. Mehri and M. Eskenazi, “Unsupervised evaluation of interactive dialog with Di-aloGPT,” in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, July 2020, pp. 225–235.
- [139] R. T. Lowe, N. Pow, I. V. Serban, L. Charlin, C.-W. Liu, and J. Pineau, “Training end-to-end dialogue systems with the ubuntu dialogue corpus,” *Dialogue & Discourse*, vol. 8, no. 1, pp. 31–65, 2017.
- [140] J. D. Williams, K. Asadi, and G. Zweig, “Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning,” *arXiv preprint arXiv:1702.03274*, 2017.
- [141] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma, “Topic aware neural response generation,” in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [142] T. Zhao, R. Zhao, and M. Eskenazi, “Learning discourse-level diversity for neural dialog models using conditional variational autoencoders,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 654–664.
- [143] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, “Emotional chatting machine: Emotional conversation generation with internal and external memory,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [144] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, “Building end-to-end dialogue systems using generative hierarchical neural network models,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [145] L. Shao, S. Gouws, D. Britz, A. Goldie, B. Strope, and R. Kurzweil, “Generating long and diverse responses with neural conversation models,” 2016.
- [146] J. Vincent, “Twitter taught microsoft’s friendly ai chatbot to be a racist asshole in less than a day,” Mar 2016. [Online]. Available: <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>

- [147] Z. Hu, X. Ma, Z. Liu, E. Hovy, and E. Xing, “Harnessing deep neural networks with logic rules,” *arXiv preprint arXiv:1603.06318*, 2016.
- [148] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proceedings of the 34th International Conference on Machine Learning- Volume 70*. JMLR. org, 2017, pp. 3319–3328.
- [149] M.-H. Su, C.-H. Wu, K.-Y. Huang, Q.-B. Hong, and H.-H. Huang, “Follow-up question generation using pattern-based seq2seq with a small corpus for interview coaching.” in *INTERSPEECH*, 2018.
- [150] M.-H. Su, C.-H. Wu, and Y. Chang, “Follow-up question generation using neural tensor network-based domain ontology population in an interview coaching system.” in *INTERSPEECH*, 2019.
- [151] P. R. SB, M. Agnihotri, and D. B. Jayagopi, “Automatic follow-up question generation for asynchronous interviews,” in *Proceedings of the Workshop on Intelligent Information Processing and Natural Language Generation*, 2020.
- [152] K. Inoue, K. Hara, D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara, “Job interviewer android with elaborate follow-up question generation,” in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020.
- [153] Y. Wang, C. Liu, M. Huang, and L. Nie, “Learning to ask questions in open-domain conversational systems with typed decoders,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2193–2203.
- [154] M. Aliannejadi, H. Zamani, F. Crestani, and W. B. Croft, “Asking clarifying questions in open-domain information-seeking conversations,” in *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval (SIGIR)*, 2019.
- [155] C. Nass, J. Steuer, and E. R. Tauber, “Computers are social actors,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1994, pp. 72–78.
- [156] M. Porcheron, J. E. Fischer, S. Reeves, and S. Sharples, “Voice interfaces in everyday life,” in *proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–12.
- [157] J. Cambre and C. Kulkarni, “One voice fits all? social implications and research challenges of designing voices for smart devices,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–19, 2019.
- [158] P. Khadpe, R. Krishna, L. Fei-Fei, J. Hancock, and M. Bernstein, “Conceptual metaphors impact perceptions of human-ai collaboration,” *arXiv preprint arXiv:2008.02311*, 2020.

- [159] M. Braun, A. Mainz, R. Chadowitz, B. Pfleging, and F. Alt, “At your service: Designing voice assistant personalities to improve automotive user interfaces,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–11.
- [160] R. K. Moore, “Appropriate voices for artefacts: some key insights,” in *1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*, 2017.
- [161] R. K. Moore, “Is spoken language all-or-nothing? implications for future speech-based human-machine interaction,” in *Dialogues with Social Robots*. Springer, 2017, pp. 281–291.
- [162] E. Luger and A. Sellen, “”like having a really bad pa”: The gulf between user expectation and experience of conversational agents,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: <https://doi.org/10.1145/2858036.2858288> p. 5286–5297.
- [163] D. R. Large, L. Clark, G. Burnett, K. Harrington, J. Luton, P. Thomas, and P. Bennett, “” it’s small talk, jim, but not as we know it.” engendering trust through human-agent conversation in an autonomous, self-driving car,” in *Proceedings of the 1st International Conference on Conversational User Interfaces*, 2019, pp. 1–7.
- [164] J. Cambre, Y. Liu, R. E. Taylor, and C. Kulkarni, “Vitro: Designing a voice assistant for the scientific lab workplace,” in *Proceedings of the 2019 on Designing Interactive Systems Conference*, 2019, pp. 1531–1542.
- [165] C. M. Myers, A. Furqan, and J. Zhu, “The impact of user characteristics and preferences on performance with an unfamiliar voice user interface,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: <https://doi.org/10.1145/3290605.3300277> p. 1–9.
- [166] A. Kim, J.-M. Park, and U. Lee, “Interruptibility for in-vehicle multitasking: Influence of voice task demands and adaptive behaviors,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–22, 2020.
- [167] J. Wirfs-Brock, S. Mennicken, and J. Thom, “Giving voice to silent data: Designing with personal music listening history,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–11.
- [168] S. R. Porter, “Raising response rates: What works?” *New directions for institutional research*, vol. 2004, no. 121, pp. 5–21, 2004.
- [169] E. Singer and M. P. Couper, “Some methodological uses of responses to open questions and other verbatim comments in quantitative surveys,” *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, vol. 11, no. 2, pp. 115–134, 2017.

- [170] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini, “Understanding malicious behavior in crowdsourcing platforms: The case of online surveys,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015, pp. 1631–1640.
- [171] J. A. Krosnick, “Survey research,” *Annual review of psychology*, vol. 50, no. 1, pp. 537–567, 1999.
- [172] A. L. Thomaz and C. Breazeal, “Teachable robots: Understanding human teaching behavior to build more effective robot learners,” *Artificial Intelligence*, vol. 172, no. 6-7, pp. 716–737, 2008.
- [173] Z. Xiao, M. X. Zhou, Q. V. Liao, G. Mark, C. Chi, W. Chen, and H. Yang, “Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys with open-ended questions,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 27, no. 3, pp. 1–37, 2020.
- [174] Z. Xiao, M. X. Zhou, and W.-T. Fu, “Who should be my teammates: Using a conversational agent to understand individuals and help teaming,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI ’19. New York, NY, USA: ACM, 2019. [Online]. Available: <http://doi.acm.org/10.1145/3301275.3302264> pp. 437–447.
- [175] C. Bauer, K. Figl, and R. Motschnig-Pitrik, “Introducing ‘active listening’ to instant messaging and e-mail: Benefits and limitations,” vol. 7, no. 2, pp. 1–17, 2010.
- [176] S. Louw, R. W. Todd, and P. Jimarkon, “Active listening in qualitative research interviews,” in *Proceedings of the International Conference: Research in Applied Linguistics, April*, 2011.
- [177] S. Kim, J. Lee, and G. 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*, ser. CHI ’19. New York, NY, USA: ACM, 2019. [Online]. Available: <https://doi.org/10.1145/3290605.3300316>
- [178] S. Egelman, E. Chi, and S. Dow, “Crowdsourcing in hci research,” in *Ways of Knowing in HCI*, J. Olson and W. Kellogg, Eds. Springer, 2014, pp. 267–289.
- [179] M. X. Zhou, G. Mark, J. Li, and H. Yang, “Trusting virtual agents: the effect of personality,” *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 9, no. 2-3, pp. 1–36, 2019.
- [180] H. P. Grice, “Logic and conversation,” in *Speech acts*. Brill, 1975, pp. 41–58.
- [181] L. Dybkjaer, N. O. Bernsen, and H. Dybkjaer, “Grice incorporated: Cooperativity in spoken dialogue,” in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996. [Online]. Available: <https://www.aclweb.org/anthology/C96-1056>

- [182] D. S. Jones, *Elementary information theory*. Clarendon Press, 1979.
- [183] G. N. Leech, “100 million words of english: the british national corpus (bnc),” *Language Research*, vol. 28, no. 1, pp. 1–13, 1992.
- [184] K. Hofland and S. Johansson, *Word frequencies in british and american english*. Norwegian computing centre for the Humanities, 1982.
- [185] T. Parviainen, “teropa/nlp,” Sep 2010. [Online]. Available: <https://github.com/teropa/nlp/tree/master/resources/corpora/webtext>
- [186] E. N. Forsythand and C. H. Martell, “Lexical and discourse analysis of online chat dialog,” in *Semantic Computing, 2007. ICSC 2007. International Conference on*. IEEE, 2007, pp. 19–26.
- [187] I. Altman and D. A. Taylor, *Social penetration: The development of interpersonal relationships*. Holt, Rinehart & Winston, 1973.
- [188] K. Krippendorff, “Computing krippendorff’s alpha-reliability,” Tech. Rep., 2011.
- [189] P. H. Rossi, J. D. Wright, and A. B. Anderson, Eds., *Handbook of survey research*. Academic Press., 2013.
- [190] D. A. Aaker, “Design and analysis: A researcher’s handbook,” *JMR, Journal of Marketing Research (pre-1986)*, vol. 13, no. 000003, p. 318, 1976.
- [191] V. Venkatesh, M. G. Morris, and P. L. Ackerman, “A longitudinal field investigation of gender differences in individual technology adoption decision-making processes,” *Organizational behavior and human decision processes*, vol. 83, no. 1, pp. 33–60, 2000.
- [192] M. G. Morris and V. Venkatesh, “Age differences in technology adoption decisions: Implications for a changing work force,” *Personnel psychology*, vol. 53, no. 2, pp. 375–403, 2000.
- [193] E. Hayes and M. Ohrnberger, “The gamer generation teaches school: The gaming practices and attitudes towards technology of pre-service teachers,” *Journal of Technology and Teacher Education*, vol. 21, no. 2, pp. 154–177, 2013.
- [194] A. Carstens and J. Beck, “Get ready for the gamer generation,” *TechTrends*, vol. 49, no. 3, pp. 22–25, 2005.
- [195] D. Schulman and T. Bickmore, “Persuading users through counseling dialogue with a conversational agent,” in *Proceedings of the 4th international conference on persuasive technology*. ACM, 2009, p. 25.
- [196] T. Bickmore, D. Schulman, and L. Yin, “Maintaining engagement in long-term interventions with relational agents,” *Applied Artificial Intelligence*, vol. 24, no. 6, pp. 648–666, 2010.

- [197] Y. Moon, *Intimate self-disclosure exchanges: Using computers to build reciprocal relationships with consumers.* Division of Research, Harvard Business School, 1998.
- [198] R. Paul and L. Elder, *The thinker's guide to the art of Socratic questioning.* Foundation for Critical Thinking Dillon Beach, CA, 2006.
- [199] A. Mavletova, “Data quality in pc and mobile web surveys,” *Social Science Computer Review*, vol. 31, no. 6, pp. 725–743, 2013.
- [200] T. Wells, J. T. Bailey, and M. W. Link, “Comparison of smartphone and online computer survey administration,” *Social Science Computer Review*, vol. 32, no. 2, pp. 238–255, 2014.
- [201] C. Antoun, M. P. Couper, and F. G. Conrad, “Effects of mobile versus pc web on survey response quality: A crossover experiment in a probability web panel,” *Public Opinion Quarterly*, vol. 81, no. S1, pp. 280–306, 2017.
- [202] J. B. Hirsh, S. K. Kang, and G. V. Bodenhausen, “Personalized persuasion: Tailoring persuasive appeals to recipients’ personality traits,” *Psychological Science*, vol. 23, no. 6, pp. 578–581, 2012.
- [203] R. E. Petty, J. T. Cacioppo, and R. Goldman, “Personal involvement as a determinant of argument-based persuasion,” *Journal of personality and social psychology*, vol. 41, no. 5, p. 847, 1981.
- [204] W. K. Park, “Mobile phone addiction,” *Mobile Communications: Re-negotiation of the Social Sphere*, pp. 253–272, 2005.
- [205] A. Bianchi and J. G. Phillips, “Psychological predictors of problem mobile phone use,” *CyberPsychology & Behavior*, vol. 8, no. 1, pp. 39–51, 2005.
- [206] G. M. Lucas, J. Gratch, A. King, and L.-P. Morency, “It’s only a computer: Virtual humans increase willingness to disclose,” *Computers in Human Behavior*, vol. 37, pp. 94–100, 2014.
- [207] Y. Wang, L. Gou, A. Xu, M. X. Zhou, H. Yang, and H. Badenes, “Veilme: An interactive visualization tool for privacy configuration of using personality traits,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* ACM, 2015, pp. 817–826.
- [208] K. Vaccaro, D. Huang, M. Eslami, C. Sandvig, K. Hamilton, and K. Karahalios, “The illusion of control: Placebo effects of control settings,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* ACM, 2018, p. 16.
- [209] J. Dunn, “We put siri, alexa, google assistant, and cortana through a marathon of tests to see who’s winning the virtual assistant race—here’s what we found. business insider 2016 (7 november),” 2016.

- [210] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, and M. Zhou, “Superagent: A customer service chatbot for e-commerce websites,” *Proceedings of ACL 2017, System Demonstrations*, pp. 97–102, 2017.
- [211] S. J. Law, J. Bourdage, and T. A. O’Neill, “To fake or not to fake: antecedents to interview faking, warning instructions, and its impact on applicant reactions,” *Frontiers in psychology*, vol. 7, p. 1771, 2016.
- [212] J. C. R. Licklider, “Man-computer symbiosis,” *IRE Transactions on Human Factors in Electronics*, vol. HFE-1, no. 1, pp. 4–11, 1960.
- [213] Z. Xiao, M. X. Zhou, W. Chen, H. Yang, and C. Chi, “If i hear you correctly: Building and evaluating interview chatbots with active listening skills,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: <https://doi.org/10.1145/3313831.3376131> p. 1–14.
- [214] J. F. Nunamaker, D. C. Derrick, A. C. Elkins, J. K. Burgoon, and M. W. Patton, “Embodied conversational agent-based kiosk for automated interviewing,” *Journal of Management Information Systems*, vol. 28, no. 1, pp. 17–48, 2011.
- [215] T. Gordon, *Leader effectiveness training, LET: The no-lose way to release the productive potential of people*. Wyden Books New York, NY, 1977.
- [216] C. R. Rogers and R. E. Farson, “Active listening,” *Organizational Psychology, 4th Ed. Englewood Cliffs, NJ*, pp. 255–266, 1984.
- [217] S. Louw, R. W. Todd, and P. Jimarkon, “Active listening in qualitative research interviews,” in *Proceedings of the International Conference: Research in Applied Linguistics, April*, 2011.
- [218] A. Watson, T. Bickmore, A. Cange, A. Kulshreshtha, and J. Kvedar, “An internet-based virtual coach to promote physical activity adherence in overweight adults: randomized controlled trial,” *Journal of medical Internet research*, vol. 14, no. 1, p. e1, 2012.
- [219] C. Bauer, K. Figl, and R. Motschnig-Pitrik, “Introducing ‘active listening’ to instant messaging and e-mail: Benefits and limitations,” *IADIS International Journal on WWW/Internet*, vol. 7, no. 2, pp. 1–17, 2010.
- [220] C. Bauer and K. Figl, “Active listening” in written online communication-a case study in a course on “soft skills for computer scientists,” in *2008 38th Annual Frontiers in Education Conference*. IEEE, 2008, pp. F2C–1.
- [221] J. Weizenbaum et al., “Eliza—a computer program for the study of natural language communication between man and machine,” *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.
- [222] M. C. McCord, J. W. Murdock, and B. K. Boguraev, “Deep parsing in watson,” *IBM Journal of research and development*, vol. 56, no. 3.4, pp. 3–1, 2012.

- [223] P. Gebhard, T. Baur, I. Damian, G. Mehlmann, J. Wagner, and E. André, “Exploring interaction strategies for virtual characters to induce stress in simulated job interviews,” in *Proceedings of the 2014 international conference on Autonomous agents and multiagent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 661–668.
- [224] H. Weger Jr, G. R. Castle, and M. C. Emmett, “Active listening in peer interviews: The influence of message paraphrasing on perceptions of listening skill,” *The Intl. Journal of Listening*, vol. 24, no. 1, pp. 34–49, 2010.
- [225] B. Decker, *How to communicate effectively*. Kogan Page, 1989.
- [226] S. Gupta, K. Jagannath, N. Aggarwal, R. Sridar, S. Wilde, and Y. Chen, “Artificially intelligent (ai) tutors in the classroom: A need assessment study of designing chatbots to sup-port student learning,” in *Proceedings of the 2019 PACIS Pacific Asia Conference on Information Systems*. AIS, 2019.
- [227] T. Lee, K. Jagannath, N. Aggarwal, R. Sridar, S. Wilde, T. Hill, and Y. Chen, “Intelligent career advisers in your pocket? a need assessment study of chatbots for student career advising,” 2019.
- [228] Z. Xiao, M. X. Zhou, and W.-T. Fu, “Who should be my teammates: Using a conversational agent to understand individuals and help teaming,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 437–447.
- [229] T. Hu, A. Xu, Z. Liu, Q. You, Y. Guo, V. Sinha, J. Luo, and R. Akkiraju, “Touch your heart: A tone-aware chatbot for customer care on social media,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 415.
- [230] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [231] R. Řehřek and P. Sojka, “Gensim—statistical semantics in python,” *statistical semantics; gensim; Python; LDA; SVD*, 2011.
- [232] Y. Song, S. Pan, S. Liu, M. X. Zhou, and W. Qian, “Topic and keyword re-ranking for lda-based topic modeling,” in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 1757–1760.
- [233] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian, “Tiara: Interactive, topic-based visual text summarization and analysis,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, p. 25, 2012.
- [234] G. Erkan and D. R. Radev, “Lexrank: Graph-based lexical centrality as salience in text summarization,” *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.
- [235] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar et al., “Universal sentence encoder,” *arXiv preprint arXiv:1803.11175*, 2018.

- [236] T. Bickmore, “Relational agents for chronic disease self management,” *Health Informatics: A Patient-Centered Approach to Diabetes*, pp. 181–204, 2010.
- [237] G. Keppel, *Design and analysis: A researcher’s handbook*. Prentice-Hall, Inc, 1991.
- [238] Q. Wang, Z. Mao, B. Wang, and L. Guo, “Knowledge graph embedding: A survey of approaches and applications,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.
- [239] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, “A review of relational machine learning for knowledge graphs,” *Proceedings of the IEEE*, vol. 104, no. 1, pp. 11–33, 2015.
- [240] J. Araki, D. Rajagopal, S. Sankaranarayanan, S. Holm, Y. Yamakawa, and T. Mitamura, “Generating questions and multiple-choice answers using semantic analysis of texts,” in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1125–1136.
- [241] X. Du, J. Shao, and C. Cardie, “Learning to ask: Neural question generation for reading comprehension,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1342–1352.
- [242] R. J. Gallagher, K. Reing, D. Kale, and G. Ver Steeg, “Anchored correlation explanation: Topic modeling with minimal domain knowledge,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 529–542, 2017.
- [243] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, and L. Carin, “Joint embedding of words and labels for text classification,” *arXiv preprint arXiv:1805.04174*, 2018.
- [244] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, “Wizard of wikipedia: Knowledge-powered conversational agents,” in *International Conference on Learning Representations*, 2018.
- [245] H. Duan, Y. Cao, C.-Y. Lin, and Y. Yu, “Searching questions by identifying question topic and question focus,” in *Proceedings of Acl-08: HLT*, 2008, pp. 156–164.
- [246] S. Moon, P. Shah, A. Kumar, and R. Subba, “Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
- [247] H. Bast, F. Bäurle, B. Buchhold, and E. Haußmann, “Easy access to the freebase dataset,” in *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, 2014.
- [248] C. Wilson, *Interview techniques for UX practitioners: A user-centered design method*. Newnes, 2013.

- [249] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, “Translating embeddings for modeling multi-relational data,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.
- [250] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems (NIPS)*, 2017.
- [251] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, “DIALOGPT : Large-scale generative pre-training for conversational response generation,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*, 2020.
- [252] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever et al., “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [253] M. Eskritt, J. Whalen, and K. Lee, “Preschoolers can recognize violations of the gricean maxims,” *British Journal of Developmental Psychology*, vol. 26, no. 3, pp. 435–443, 2008.
- [254] S. Kleinke, “Speaker activity and grice’s maxims of conversation at the interface of pragmatics and cognitive linguistics,” *Journal of pragmatics*, vol. 42, no. 12, pp. 3345–3366, 2010.
- [255] R. Langevin, R. J. Lordon, T. Avrahami, B. R. Cowan, T. Hirsch, and G. Hsieh, “Heuristic evaluation of conversational agents,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–15.
- [256] R. Bommasani and C. Cardie, “Intrinsic evaluation of summarization datasets,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.
- [257] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, “Learning entity and relation embeddings for knowledge graph completion,” in *Twenty-ninth AAAI conference on artificial intelligence (AAAI)*, 2015.
- [258] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, “Knowledge graph embedding via dynamic mapping matrix,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 2015.
- [259] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020.
- [260] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.

- [261] Z. Xiao, T. W. Li, K. Karahalios, and H. Sundaram, “Inform the uninformed: Improving online informed consent reading with an ai-powered chatbot,” *arXiv preprint arXiv:2302.00832*, 2023.
- [262] A. Sherlock and S. Brownie, “Patients’ recollection and understanding of informed consent: a literature review,” *ANZ journal of surgery*, vol. 84, no. 4, pp. 207–210, 2014.
- [263] N. C. F. Moreira, C. Pacheco-Pereira, L. Keenan, G. Cummings, and C. Flores-Mir, “Informed consent comprehension and recollection in adult dental patients: a systematic review,” *The Journal of the American Dental Association*, vol. 147, no. 8, pp. 605–619, 2016.
- [264] C. M. Ferreira and S. Serpa, “Informed consent in social sciences research: Ethical challenges,” *Int'l J. Soc. Sci. Stud.*, vol. 6, p. 13, 2018.
- [265] L. Ghandour, R. Yasmine, and F. El-Kak, “Giving consent without getting informed: a cross-cultural issue in research ethics,” *Journal of Empirical Research on Human Research Ethics*, vol. 8, no. 3, pp. 12–21, 2013.
- [266] M. G. Del Carmen and S. Joffe, “Informed consent for medical treatment and research: a review,” *The oncologist*, vol. 10, no. 8, pp. 636–641, 2005.
- [267] B. R. Cassileth, R. V. Zupkis, K. Sutton-Smith, and V. March, “Informed consent—why are its goals imperfectly realized?” *New England journal of medicine*, vol. 302, no. 16, pp. 896–900, 1980.
- [268] J. D. Heider, J. L. Hartnett, E. J. Perez, and J. E. Edlund, “Perceptions and understanding of research situations as a function of consent form characteristics and experimenter instructions,” *Methods in Psychology*, vol. 2, p. 100015, 2020.
- [269] M. Assale, E. Barbero, and F. Cabitza, “Digitizing the informed consent: the challenges to design for practices,” in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2019, pp. 609–615.
- [270] J. Walkup and E. Bock, “What do prospective research participants want to know? what do they assume they know already?” *Journal of Empirical Research on Human Research Ethics*, vol. 4, no. 2, pp. 59–63, 2009.
- [271] C.-Y. Lin, A. Broström, P. Nilsen, M. D. Griffiths, and A. H. Pakpour, “Psychometric validation of the persian bergen social media addiction scale using classic test theory and rasch models,” *Journal of behavioral addictions*, vol. 6, no. 4, pp. 620–629, 2017.
- [272] Baylor University, “Are you addicted to social media? six questions,” *Science Daily*, Oct. 2016.
- [273] K. Mulligan, J. A. Krosnick, W. Smith, M. Green, and G. Bizer, “Nondifferentiation on attitude rating scales: A test of survey satisficing theory,” *Unpublished manuscript*, 2001.

- [274] Z. Xiao, Q. V. Liao, M. Zhou, T. Grandison, and Y. Li, “Powering an ai chatbot with expert sourcing to support credible health information access,” in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023, pp. 2–18.
- [275] T. W. Bickmore, H. Trinh, S. Olafsson, T. K. O’Leary, R. Asadi, N. M. Rickles, and R. Cruz, “Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant,” *Journal of medical Internet research*, vol. 20, no. 9, p. e11510, 2018.
- [276] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, “Language (technology) is power: A critical survey of “bias” in nlp,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5454–5476.
- [277] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, “Dialogpt: Large-scale generative pre-training for conversational response generation,” *arXiv preprint arXiv:1911.00536*, 2019.
- [278] Juji, “Juji document for chatbot designers,” <https://docs.juji.io/>, 2020, [Online; accessed 14-June-2020].
- [279] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray et al., “Training language models to follow instructions with human feedback,” *arXiv preprint arXiv:2203.02155*, 2022.
- [280] W. Grabe, *Reading in a second language: Moving from theory to practice*. Cambridge University Press, 2008.
- [281] S. Henderson, “Power imbalance between nurses and patients: a potential inhibitor of partnership in care,” *Journal of clinical nursing*, vol. 12, no. 4, pp. 501–508, 2003.
- [282] A. Tapal, E. Oren, R. Dar, and B. Eitam, “The sense of agency scale: A measure of consciously perceived control over one’s mind, body, and the immediate environment,” *Frontiers in psychology*, vol. 8, p. 1552, 2017.
- [283] Y. Kim, J. Dykema, J. Stevenson, P. Black, and D. P. Moberg, “Straightlining: overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys,” *Social Science Computer Review*, vol. 37, no. 2, pp. 214–233, 2019.
- [284] T. Yan, “Nondifferentiation,” *Encyclopedia of survey research methods*, vol. 2, pp. 520–521, 2008.
- [285] J. R. Lewis, “Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the asq,” *ACM Sigchi Bulletin*, vol. 23, no. 1, pp. 78–81, 1991.
- [286] M. Kay, G. L. Nelson, and E. B. Hekler, “Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of hci,” in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 4521–4532.

- [287] H. Chen, P. Cohen, and S. Chen, “How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies,” *Communications in Statistics—simulation and Computation(R)*, vol. 39, no. 4, pp. 860–864, 2010.
- [288] G. R. Gibbs, “Thematic coding and categorizing,” *Analyzing qualitative data*, vol. 703, pp. 38–56, 2007.
- [289] R. McElreath, *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2015.
- [290] R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers, “Robust misinterpretation of confidence intervals,” *Psychonomic bulletin & review*, vol. 21, no. 5, pp. 1157–1164, 2014.
- [291] K. Y. Tam and S. Y. Ho, “Web personalization as a persuasion strategy: An elaboration likelihood model perspective,” *Information systems research*, vol. 16, no. 3, pp. 271–291, 2005.
- [292] K. R. Ripley, M. A. Hance, S. A. Kerr, L. E. Brewer, and K. E. Conlon, “Uninformed consent? the effect of participant characteristics and delivery format on informed consent,” *Ethics & Behavior*, vol. 28, no. 7, pp. 517–543, 2018.
- [293] S. Ruan, L. Jiang, J. Xu, B. J.-K. Tham, Z. Qiu, Y. Zhu, E. L. Murnane, E. Brunskill, and J. A. Landay, “Quizbot: A dialogue-based adaptive learning system for factual knowledge,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, p. 357.
- [294] F. Colace, M. De Santo, M. Lombardi, F. Pascale, A. Pietrosanto, and S. Lemma, “Chatbot for e-learning: A case of study,” *International Journal of Mechanical Engineering and Robotics Research*, vol. 7, no. 5, pp. 528–533, 2018.
- [295] Y. Ge, Z. Xiao, J. Diesner, H. Ji, K. Karahalios, and H. Sundaram, “What should i ask: A knowledge-driven approach for follow-up questions generation in conversational surveys,” *arXiv preprint arXiv:2205.10977*, 2022.
- [296] B. Barry and G. L. Stewart, “Composition, process, and performance in self-managed groups: The role of personality.” *Journal of Applied psychology*, vol. 82, no. 1, p. 62, 1997.
- [297] S. T. Bell, “Deep-level composition variables as predictors of team performance: a meta-analysis.” *Journal of applied psychology*, vol. 92, no. 3, p. 595, 2007.
- [298] T. Halfhill, T. M. Nielsen, E. Sundstrom, and A. Weilbaecher, “Group personality composition and performance in military service teams.” *Military Psychology*, vol. 17, no. 1, p. 41, 2005.

- [299] I. Lykourentzou, A. Antoniou, Y. Naudet, and S. P. Dow, “Personality matters: Balancing for personality types leads to better outcomes for crowd teams,” in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 2016, pp. 260–273.
- [300] E. F. Barkley, K. P. Cross, and C. H. Major, *Collaborative learning techniques: A handbook for college faculty*. John Wiley & Sons, 2014.
- [301] D. W. Johnson, R. T. Johnson, and K. Smith, “The state of cooperative learning in postsecondary and professional settings,” *Educational Psychology Review*, vol. 19, no. 1, pp. 15–29, 2007.
- [302] M. Ziegler, C. MacCann, and R. Roberts, *New perspectives on faking in personality assessment*. Oxford University Press, 2011.
- [303] S. E. Humphrey, J. R. Hollenbeck, C. J. Meyer, and D. R. Ilgen, “Trait configurations in self-managed teams: A conceptual examination of the use of seeding for maximizing and minimizing trait variance in teams.” *Journal of Applied Psychology*, vol. 92, no. 3, p. 885, 2007.
- [304] D. L. Paulhus, “Measurement and control of response bias.” 1991.
- [305] L. Uziel, “Impression management (“lie”) scales are associated with interpersonally oriented self-control, not other-deception,” *Journal of personality*, vol. 82, no. 3, pp. 200–212, 2014.
- [306] T. J. Whelan, L. Aiman-Smith, C. Kimbrough, and L. Taylor, “Group personality composition, satisfaction and performance in virtual teams,” in *24th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA*, 2009.
- [307] J. B. Bear and A. W. Woolley, “The role of gender in team collaboration and performance,” *Interdisciplinary science reviews*, vol. 36, no. 2, pp. 146–153, 2011.
- [308] K. A. Jehn, G. B. Northcraft, and M. A. Neale, “Why differences make a difference: A field study of diversity, conflict and performance in workgroups,” *Administrative science quarterly*, vol. 44, no. 4, pp. 741–763, 1999.
- [309] M. A. Burress, “The relationship between team leader behaviors and team performance and satisfaction,” Ph.D. dissertation, University of North Texas, 1996.
- [310] G. L. Stewart, “A meta-analytic review of relationships between team design features and team performance,” *Journal of management*, vol. 32, no. 1, pp. 29–55, 2006.
- [311] A. C. Edmondson and Z. Lei, “Psychological safety: The history, renaissance, and future of an interpersonal construct,” *Annu. Rev. Organ. Psychol. Organ. Behav.*, vol. 1, no. 1, pp. 23–43, 2014.

- [312] S. J. Zaccaro and M. C. McCoy, “The effects of task and interpersonal cohesiveness on performance of a disjunctive group task,” *Journal of applied social psychology*, vol. 18, no. 10, pp. 837–851, 1988.
- [313] C. D. Georges and E. V. Susteren, “5 ways to get the survey data you want,” 2018. [Online]. Available: <https://www.surveymonkey.com/curiosity/5-best-ways-to-get-survey-data/>
- [314] C. Nass, Y. Moon, B. J. Fogg, B. Reeves, and C. Dryer, “Can computer personalities be human personalities?” in *Conference companion on Human factors in computing systems*. ACM, 1995, pp. 228–229.
- [315] D. Oertig, J. Schüler, V. Brandstätter, and A. A. Augustine, “The influence of avoidance temperament and avoidance-based achievement goals on flow,” *Journal of personality*, vol. 82, no. 3, pp. 171–181, 2014.
- [316] V. U. Druskat and S. B. Wolff, “Building the emotional intelligence of groups,” *Harvard business review*, vol. 79, no. 3, pp. 80–91, 2001.
- [317] F. Jahanbakhsh, W.-T. Fu, K. Karahalios, D. Marinov, and B. Bailey, “You want me to work with who?: Stakeholder perceptions of automated team formation in project-based courses,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 3201–3212.
- [318] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr, “A very brief measure of the big-five personality domains,” *Journal of Research in personality*, vol. 37, no. 6, pp. 504–528, 2003.
- [319] M. Pennock-Roman, “Interpreting test performance in selective admissions for hispanic students.” 1992.
- [320] N. Ratanawongsa, P. A. Thomas, S. S. Marinopoulos, T. Dorman, L. M. Wilson, B. H. Ashar, J. L. Magaziner, R. G. Miller, G. P. Prokopowicz, R. Qayyum et al., “The reported validity and reliability of methods for evaluating continuing medical education: a systematic review,” *Academic Medicine*, vol. 83, no. 3, pp. 274–283, 2008.
- [321] W. Shen, P. R. Sackett, N. R. Kuncel, A. S. Beatty, J. L. Rigdon, and T. B. Kiger, “All validities are not created equal: Determinants of variation in sat validity across schools,” *Applied Measurement in Education*, vol. 25, no. 3, pp. 197–219, 2012.
- [322] E. Meyer, *The culture map: Breaking through the invisible boundaries of global business*. PublicAffairs, 2014.
- [323] X.-H. Lin, R. Bian, R. Zhu, and H.-S. Che, “Team personality composition and team effectiveness: The mediating effects of team process,” *Acta Psychologica Sinica*, vol. 40, no. 4, pp. 437–447, 2008.

- [324] Z. Xiao, S. Mennicken, B. Huber, A. Shonkoff, and J. Thom, “Let me ask you this: How can a voice assistant elicit explicit user feedback?” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–24, 2021.
- [325] A. Maedche, C. Legner, A. Benlian, B. Berger, H. Gimpel, T. Hess, O. Hinz, S. Morana, and M. Söllner, “Ai-based digital assistants,” *Business & Information Systems Engineering*, vol. 61, no. 4, pp. 535–544, 2019.
- [326] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, “Accurately interpreting clickthrough data as implicit feedback,” in *ACM SIGIR Forum*, vol. 51, no. 1. Acm New York, NY, USA, 2017, pp. 4–11.
- [327] F. M. Harper, X. Li, Y. Chen, and J. A. Konstan, “An economic model of user rating in an online recommender system,” in *International conference on user modeling*. Springer, 2005, pp. 307–316.
- [328] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, “Policy shaping: Integrating human feedback with reinforcement learning,” in *Advances in neural information processing systems*, 2013, pp. 2625–2633.
- [329] J. Y. Kim, J. Teevan, and N. Craswell, “Explicit in situ user feedback for web search results,” in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 829–832.
- [330] S. N. Patel, J. A. Kientz, G. R. Hayes, S. Bhat, and G. D. Abowd, “Farther than you may think: An empirical investigation of the proximity of users to their mobile phones,” in *International conference on ubiquitous computing*. Springer, 2006, pp. 123–140.
- [331] G. Adomavicius and A. Tuzhilin, “Context-aware recommender systems,” in *Recommender systems handbook*. Springer, 2011, pp. 217–253.
- [332] M. H. Cohen, M. H. Cohen, J. P. Giangola, and J. Balogh, *Voice user interface design*. Addison-Wesley Professional, 2004.
- [333] C. Pearl, *Designing voice user interfaces: principles of conversational experiences*. ” O'Reilly Media, Inc.”, 2016.
- [334] M. Zanker and M. Jessenitschnig, “Collaborative feature-combination recommender exploiting explicit and implicit user feedback,” in *2009 IEEE Conference on Commerce and Enterprise Computing*. IEEE, 2009, pp. 49–56.
- [335] H. Egermann, M. E. Sutherland, O. Grewe, F. Nagel, R. Kopiez, and E. Altenmüller, “Does music listening in a social context alter experience? a physiological and psychological perspective on emotion,” *Musicae Scientiae*, vol. 15, no. 3, pp. 307–323, 2011.
- [336] J. Parson, D. Braga, M. Tjalve, and J. Oh, “Evaluating voice quality and speech synthesis using crowdsourcing,” in *International Conference on Text, Speech and Dialogue*. Springer, 2013, pp. 233–240.

- [337] S. Gabriele and S. Chiasson, “Understanding fitness tracker users’ security and privacy knowledge, attitudes and behaviours,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.
- [338] S. Davidoff, M. K. Lee, A. K. Dey, and J. Zimmerman, “Rapidly exploring application design through speed dating,” in *International Conference on Ubiquitous Computing*. Springer, 2007, pp. 429–446.
- [339] C. Nass, B. Fogg, and Y. Moon, “Can computers be teammates?” *International Journal of Human-Computer Studies*, vol. 45, no. 6, pp. 669–678, 1996.
- [340] D. Tao, “Intention to use and actual use of electronic information resources: further exploring technology acceptance model (tam),” in *AMIA Annual Symposium Proceedings*, vol. 2009. American Medical Informatics Association, 2009, p. 629.
- [341] D. R. Ilgen, C. D. Fisher, and M. S. Taylor, “Consequences of individual feedback on behavior in organizations.” *Journal of applied psychology*, vol. 64, no. 4, p. 349, 1979.
- [342] G. M. Sinatra, C. M. Kardash, G. Taasoobshirazi, and D. Lombardi, “Promoting attitude change and expressed willingness to take action toward climate change in college students,” *Instructional Science*, vol. 40, no. 1, pp. 1–17, 2012.
- [343] R. F. DeVellis, *Scale development: Theory and applications*. Sage publications, 2016, vol. 26.
- [344] J. Xiao, R. Catrambone, and J. T. Stasko, “Be quiet? evaluating proactive and reactive user interface assistants,” Georgia Institute of Technology, Tech. Rep., 2003.
- [345] R. A. Virzi, “Refining the test phase of usability evaluation: How many subjects is enough?” *Human factors*, vol. 34, no. 4, pp. 457–468, 1992.
- [346] M. G. Ehrhart, K. H. Ehrhart, S. C. Roesch, B. G. Chung-Herrera, K. Nadler, and K. Bradshaw, “Testing the latent factor structure and construct validity of the ten-item personality inventory,” *Personality and individual Differences*, vol. 47, no. 8, pp. 900–905, 2009.
- [347] J. D. Van Der Laan, A. Heino, and D. De Waard, “A simple procedure for the assessment of acceptance of advanced transport telematics,” *Transportation Research Part C: Emerging Technologies*, vol. 5, no. 1, pp. 1–10, 1997.
- [348] J. De Winter, M. Kyriakidis, D. Dodou, and R. Happee, “Using crowdflower to study the relationship between self-reported violations and traffic accidents,” *Procedia Manufacturing*, vol. 3, pp. 2518–2525, 2015.
- [349] C. Van Pelt and A. Sorokin, “Designing a scalable crowdsourcing platform,” in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 765–766.

- [350] N. Abdi, X. Zhan, K. M. Ramokapane, and J. Such, “Privacy norms for smart home personal assistants,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–14.
- [351] N. Abdi, K. M. Ramokapane, and J. M. Such, “More than smart speakers: security and privacy perceptions of smart home personal assistants,” in *Fifteenth Symposium on Usable Privacy and Security (SOUPS) 2019*, 2019.
- [352] S. Liao, C. Wilson, L. Cheng, H. Hu, and H. Deng, “Measuring the effectiveness of privacy policies for voice assistant applications,” in *Annual Computer Security Applications Conference*, 2020, pp. 856–869.
- [353] P. Klasnja, S. Consolvo, T. Choudhury, R. Beckwith, and J. Hightower, “Exploring privacy concerns about personal sensing,” in *International Conference on Pervasive Computing*. Springer, 2009, pp. 176–183.
- [354] S. J. Ko, C. M. Judd, and I. V. Blair, “What the voice reveals: Within-and between-category stereotyping on the basis of voice,” *Personality and Social Psychology Bulletin*, vol. 32, no. 6, pp. 806–819, 2006.
- [355] J. J. Faraway, *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press, 2016.
- [356] I. H. Kuo, J. M. Rabindran, E. Broadbent, Y. I. Lee, N. Kerse, R. Stafford, and B. A. MacDonald, “Age and gender factors in user acceptance of healthcare robots,” in *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2009, pp. 214–219.
- [357] A. Easwara Moorthy and K.-P. L. Vu, “Privacy concerns for use of voice activated personal assistant in the public space,” *International Journal of Human-Computer Interaction*, vol. 31, no. 4, pp. 307–335, 2015.
- [358] J. Lau, B. Zimmerman, and F. Schaub, “Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–31, 2018.
- [359] E. Cho, S. S. Sundar, S. Abdullah, and N. Motalebi, “Will deleting history make alexa more trustworthy? effects of privacy and content customization on user experience of smart speakers,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.
- [360] J. Broekens, A. Pommeranz, P. Wiggers, and C. M. Jonker, “Factors influencing user motivation for giving online preference feedback,” in *5th Multidisciplinary Workshop on Advances in Preference Handling (MPREF’10)*, 2010.
- [361] F. Bentley, C. Luvogt, M. Silverman, R. Wirasinghe, B. White, and D. Lottridge, “Understanding the long-term use of smart speaker assistants,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 3, Sep. 2018. [Online]. Available: <https://doi.org/10.1145/3264901>

- [362] Y. Koren, “Collaborative filtering with temporal dynamics,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 447–456.
- [363] A. Z. Jacobs and H. Wallach, “Measurement and fairness,” in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 375–385.
- [364] A. Gatt and E. Krahmer, “Survey of the state of the art in natural language generation: Core tasks, applications and evaluation,” *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.
- [365] X. Puig, T. Shu, S. Li, Z. Wang, Y.-H. Liao, J. B. Tenenbaum, S. Fidler, and A. Torralba, “Watch-and-help: A challenge for social perception and human-ai collaboration,” *arXiv preprint arXiv:2010.09890*, 2020.
- [366] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, “Sim-to-real transfer of robotic control with dynamics randomization,” in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3803–3810.
- [367] B. E. Goodstadt and L. A. Hjelle, “Power to the powerless: Locus of control and the use of power.” *Journal of Personality and Social Psychology*, vol. 27, no. 2, p. 190, 1973.
- [368] U. B. Karli, S. Cao, and C.-M. Huang, ““what if it is wrong”: Effects of power dynamics and trust repair strategy on trust and compliance in hri,” in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI ’23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: <https://doi.org/10.1145/3568162.3576964> p. 271–280.
- [369] A. M. Aroyo, T. Kyohei, T. Koyama, H. Takahashi, F. Rea, A. Sciutti, Y. Yoshikawa, H. Ishiguro, and G. Sandini, “Will people morally crack under the authority of a famous wicked robot?” in *2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2018, pp. 35–42.
- [370] D. P. Farrington, “Longitudinal research strategies: Advantages, problems, and prospects,” *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 30, no. 3, pp. 369–374, 1991.
- [371] S. Menard, *Handbook of longitudinal research: Design, measurement, and analysis*. Elsevier, 2007.
- [372] V. Bellotti and A. Sellen, “Design for privacy in ubiquitous computing environments,” in *Proceedings of the Third European Conference on Computer-Supported Cooperative Work 13–17 September 1993, Milan, Italy ECSCW’93*. Springer, 1993, pp. 77–92.
- [373] A. Fadhil and S. Gabrielli, “Addressing challenges in promoting healthy lifestyles: the al-chatbot approach,” in *Proceedings of the 11th EAI international conference on pervasive computing technologies for healthcare*, 2017, pp. 261–265.

- [374] V. M. Kumar, A. Keerthana, M. Madhumitha, S. Valliammai, and V. Vinithasri, “Sana-tive chatbot for health seekers,” *International Journal Of Engineering And Computer Science*, vol. 5, no. 03, pp. 16 022–16 025, 2016.
- [375] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [376] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, Oct. 2020.
- [377] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations (ICLR)*, 2018.

## APPENDIX A: APPENDIX FOR A KNOWLEDGE-DRIVEN APPROACH FOR FOLLOW-UP QUESTIONS GENERATION

### A.1 IMPLEMENTATION DETAILS

All the knowledge selection models use BERT<sub>large</sub> as the backbone encoder which is initialized with the pre-trained uncased weights. We use OpenKE<sup>40</sup> to pre-train all knowledge embeddings. Models are implemented by Pytorch framework [375] and Huggingface transformers [376]. We tune the parameters of each knowledge selection model with the following search space (bold indicate the choice for our final model according to the performance on the validation set): KG embeddings size: {100, 200, 300, **400**, 500}, hidden states: {100, 200, 300, 400, 500, **600**, 700}. The selection models are optimized by AdamW [377] with the learning rate of  $4e - 5$  and the linear learning scheduler. A default setting trains for 50 epochs, using a batch size of 20. Early stopping is adopted if performance on the validation set doesn't increase for consecutive 10 epochs.

As for the generative model, we initialize it with DialoGPT<sub>large</sub> and also use AdamW with the linear learning scheduler to fine-tune models. The learning rate is set to  $5e - 5$ . Similarly, we train the model for 5 epochs with the batch size 1 by default, and use early stopping to stop the training when the performance on the validation set doesn't improve for consecutive 5 epochs. During the generation, we use beam search with beam size 2.

### A.2 DATA ANNOTATION PROCESS

We built a web interface to annotate the dataset. Before the annotation task, we first showed a tutorial to explain the purpose of the task, what information is required, and how to complete the task. To facilitate the creation of high-quality follow-up questions, we asked annotators first to select the most interesting and meaningful entity mentioned in the dialogue context that they may want to ask in the follow-up question. And then, annotators need to specify what relation of the selected entity they want to collect in the follow-up questions. As the last step, we instruct annotators to write a follow-up question in a clear and coherent manner based on the above information. At the end of the tutorial, we ask all annotators to complete a tutorial task to ensure their understanding of the process.

---

<sup>40</sup><http://openke.thunlp.org/>

Datasets	Size	Domain	Knowledge Base
MHMC-IV [149]	3.4k	School Admission	
Interview Coaching [150]	1.2k	School Admission	(ConceptNet)
FQG [151]	1k	Job Interviews	
OURS	10k	General	(Freebase)

Table A.1: Statistics of our dataset and other previous datasets for follow-up question generation.

	Train	Validation	Test
Number of Dialogues	8032	1004	1004
Number of Utterances	311298	38739	38002
Number of Unique Entities	7429	1910	1920
Avg. Entities per Dialogue	2.44	2.45	2.48

Table A.2: Dataset statistics of the Knowledge-Driven Follow-up Question Generation Task.

### A.3 DATASET STATISTICS

The final dataset we collect consists of 10040 dialogues, which we divide into 8032 for train, 1004 for validation, and 1004 for the test, and we present a comparison between ours and related datasets in Table A.1. A total of 8165 unique entities are mentioned in the question-answer pairs. On average, each question-answer pair mentioned 2.45 unique entities. And for each mentioned entity, the average number of connected relations is 9.44. More details are shown in Table A.2. Regarding the annotated follow-up question for each question-answer pair, there are a total of 3917 'What' questions, 81 'How' questions, 341 'When' questions, 2619 'Who' questions, 12 'Why' questions, 806 'Which' questions, 716 'Where' Questions, and 1548 closed-ended questions.

## APPENDIX B: APPENDIX FOR CONVERSATIONAL AI FOR ONLINE INFORMED CONSENT

### B.1 MANIPULATION CHECKS

To validate our problematic social media use survey design with different risk levels, we asked participants to rate their perceived sensitivity and risk regarding the survey upon completion. Both ratings are on 5-point Likert scales. The results confirm our design. Participants in the Low-risk condition felt the survey was the least sensitive, whereas the participants in the High-risk condition perceived the survey as the most sensitive (Low:  $M = 1.64$ ,  $SD = 0.83$ ; Medium:  $M = 1.92$ ,  $SD = 1.27$ ; High:  $M = 3.13$ ,  $SD = 1.31$ ). The ANOVA analysis with the Tukey method suggests that, for the perceived data sensitivity, the differences between the three risk levels are statistically significant (High-Low:  $p < 0.01$ ; Medium-Low:  $p < 0.01$ ; High-Medium:  $p < 0.01$ ).

Participants in the Low-risk condition report the lowest level of perceived risk for the survey, whereas participants in the High-risk condition report the highest (Low:  $M = 1.66$ ,  $SD = 0.67$ ; Medium:  $M = 2.23$ ,  $SD = 1.11$ ; High:  $M = 2.71$ ,  $SD = 1.11$ ). Through ANOVA analysis with Tukey method, we found the differences between Low and High and between High and Medium are statistically significant, but the difference between Low and Medium is not (High-Low:  $p < 0.01$ ; Medium-Low:  $p = 0.25$ ; High-Medium:  $p < 0.01$ ). Both manipulation checks suggested our manipulation of the survey's perceived level of risk was successful.

### B.2 BAYESIAN MODEL SPECIFICATIONS

In this section, we provide details of the set up of our Bayesian models. In a Bayesian formulation, we define a likelihood function to model the dependent variable that represents the modeler's view of the data, not a claim about the world. The likelihood function is parametric, meaning that we treat each model parameter as a random variable drawn from another distribution with parameters (its prior distribution). We use weakly informative priors to make conservative estimations and encode skepticism but not impossibility towards large effect sizes. For all the Bayesian models discussed in this section, all parameters achieved a Gelman-Rubin statistic (a measure of MCMC convergence) of 1.0, indicating that the multiple sampling chains converged. Traceplots that visualize sampling results are available in the supplementary materials.

All of our Bayesian models (except the SEM model) share the same predictor (consent method) and covariates (risk level and demographics). We use partially pooled varying intercepts to describe the effect of the consent method and gender, making the models hierarchical. We describe the parameters for the variables below, with  $i$  representing the  $i$ -th participant.

$C_i$ : Varying intercept for the effect of the consent method used by participant  $i$ .

$\beta_R \sum_{n=0}^{R[i]-1} \delta_R$ : Slope for risk level of participant  $i$  as an ordinal variable.

$G_i$ : Varying intercept for the effect of gender of participant  $i$ .

$\beta_A \sum_{n=0}^{A[i]-1} \delta_A$ : Slope for the age level of participant  $i$  as an ordinal variable.

$\beta_E \sum_{n=0}^{E[i]-1} \delta_E$ : Slope for the education level of participant  $i$  as an ordinal variable.

$\beta_I \sum_{n=0}^{I[i]-1} \delta_I$ : Slope for the income level of participant  $i$  as an ordinal variable.

We define the linear model made up of the predictor and covariates as follows:

$$LM_i = C_{[i]} + \beta_R \sum_{n=0}^{R[i]-1} \delta_R + G_{[i]} + \beta_A \sum_{n=0}^{A[i]-1} \delta_A + \beta_E \sum_{n=0}^{E[i]-1} \delta_E + \beta_I \sum_{n=0}^{I[i]-1} \delta_I \quad (\text{B.1})$$

The priors we selected for the parameters are as follows. We will specify the hyper priors within each model in the upcoming sections.

$C_i \sim N(\bar{\mu}, \bar{\sigma})$	Prior for each consent method
$\beta_R \sim N(\mu_\beta, \sigma_\beta)$	Prior for slope of risk level
$\delta_R \sim Dirichlet(2)$	Prior for the cutpoints of ordinal risk levels
$G_i \sim N(\bar{\mu}, \bar{\sigma})$	Prior for each gender
$\beta_A \sim N(\mu_\beta, \sigma_\beta)$	Prior for slope of age level
$\delta_A \sim Dirichlet(2)$	Prior for the cutpoints of ordinal age levels
$\beta_E \sim N(\mu_\beta, \sigma_\beta)$	Prior for slope of education level
$\delta_E \sim Dirichlet(2)$	Prior for the cutpoints of ordinal education levels
$\beta_I \sim N(\mu_\beta, \sigma_\beta)$	Prior for slope of income level
$\delta_I \sim Dirichlet(2)$	Prior for the cutpoints of ordinal income levels

As described in Section 6.2.7, we built two kinds of Bayesian models, a linear regression model and an ordered logistic regression model. Next, we detail the formulation of these two models.

### B.2.1 Linear Regression Model

We model the dependent variable as a Normal distribution.

$$\begin{aligned} DV_i &\sim N(\mu_i, \sigma_i) \\ \mu_i &= \alpha_\mu + LM_i \\ \sigma_i &= \alpha_\sigma + C_{\sigma_i} \end{aligned} \tag{B.2}$$

The additional priors for parameters new in the equations above are as follows:

$\alpha_\mu \sim N(\mu_{DV}, \sigma_{DV} * 0.1)$	Prior for the intercept of $\mu$
$\alpha_\sigma \sim N(\sigma_{DV}, \sigma_{DV} * 0.1)$	Prior for the intercept of $\sigma$
$C_{\sigma_i} \sim N(\bar{\mu}, \bar{\sigma})$	Prior for the effect of each consent method on $\sigma$

For hyper priors, we set  $\bar{\mu} \sim N(0, 1)$  and  $\bar{\sigma} \sim Exp(1)$ . For the priors of slopes  $\beta$ , we set  $\mu_\beta = 0$  and  $\sigma_\beta = 1$ . For the priors of intercepts  $\alpha$ ,  $\mu_{DV}$  and  $\sigma_{DV}$  represents the mean and standard deviation of the dependent variable observed in our data.

### B.2.2 Ordered Logistic Model

In an ordered logistic model, we model the logit of the cumulative probability function of the dependent variable as a linear regression. Let  $q_k$  be the cumulative probability of getting a value of  $k$  or lower on the ordinal scale.

$$\begin{aligned} logit(q_k) &= \alpha_k - \phi_i \\ \phi_i &= LM_i \\ \alpha_k &\sim N(0, 1) \end{aligned} \tag{B.3}$$

For hyper priors, we set  $\bar{\mu} \sim N(0, 0.25)$  and  $\bar{\sigma} \sim Exp(1)$ . For the priors of slopes  $\beta$ , we set  $\mu_\beta = 0$  and  $\sigma_\beta = 0.25$ .

### B.3 MODEL CONVERGENCE

We applied non-centered parameterization to all of the models in order to increase convergence. For all of the Bayesian models described, the Gelman-Rubin statistic (a measure of MCMC convergence) for all parameters was around 1, indicating that the multiple sampling chains converged. Traceplots for the MCMC chains in all of the models are shown in the following figures. We only show the traceplots for coefficients of Consent Method, since these were the most significant variables in our models. The traceplots for the other parameters show that the models are equally well-behaved.

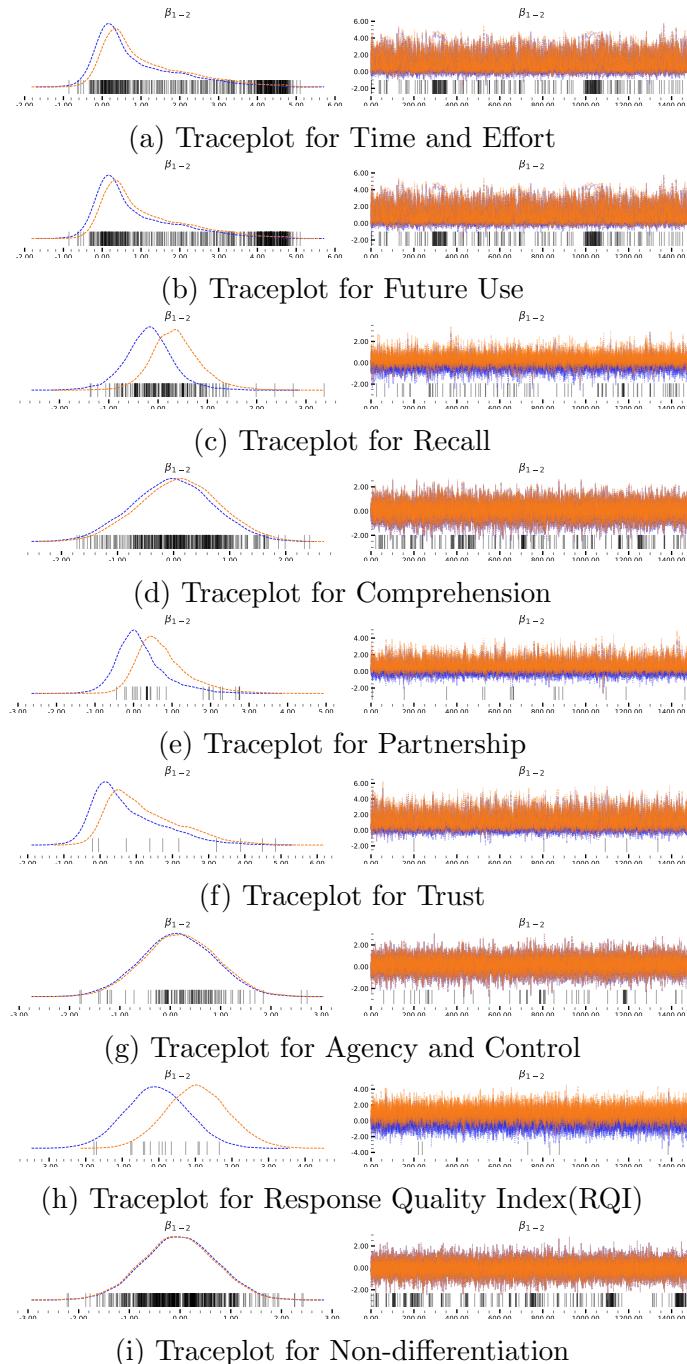


Figure B.1: Traceplots showing the results of the MCMC estimation for our models. The left column is the posterior distribution for  $\beta_{1-2}$ , the consent condition. The right column shows the corresponding sampling traces. The color mappings for the  $\beta_{1-2}$  plots are: blue - Form Condition, orange - Chatbot Condition.