Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data

JING WEI*, University of Melbourne, Australia
SUNGDONG KIM, NAVER AI Lab, Republic of Korea
HYUNHOON JUNG, NAVER CLOUD, Republic of Korea
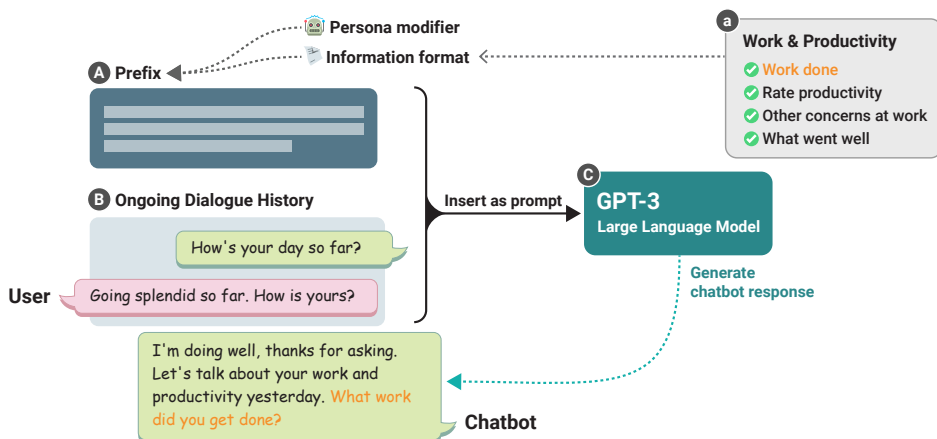YOUNG-HO KIM, NAVER AI Lab, Republic of Korea

Fig. 1. An overview of our chatbot running on a large language model through zero-shot response generation, with only a prefix (A) consisting of persona modifier and information format and the ongoing dialogue history (B). The example conversation is carried on about *work* (a). GPT-3 was used for an underlying large language model (C).

Large language models (LLMs) provide a new way to build chatbots by accepting natural language prompts. Yet, it is unclear how to design prompts to power chatbots to carry on naturalistic conversations while pursuing a given goal such as collecting self-report data from users. We explore what design factors of prompts can help steer chatbots to talk naturally and collect data reliably. To this aim, we formulated four prompt designs with different structures and personas. Through an online study ($N = 48$) where participants conversed with chatbots driven by different designs of prompts, we assessed how *prompt designs* and *conversation topics* affected the conversation flows and users' perceptions of chatbots. Our chatbots covered 79% of the desired information slots during conversations, and the designs of prompts and topics significantly influenced the

---

Authors' addresses: Jing Wei, jwwei2@student.unimelb.edu.au, University of Melbourne, Parkville, VIC, Australia; Sungdong Kim, sungdong.kim@navercorp.com, NAVER AI Lab, Seongnam, Gyeonggi, Republic of Korea; Hyunhoon Jung, hyunhoon.j@navercorp.com, NAVER CLOUD, Seongnam, Gyeonggi, Republic of Korea; Young-Ho Kim, yghokim@younghokim.net, NAVER AI Lab, Seongnam, Gyeonggi, Republic of Korea.

---

conversation flows and the data collection performance. We discuss the opportunities and challenges of building chatbots with LLMs.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Natural language generation**.

Additional Key Words and Phrases: conversational agents, chatbots, large language models, dialogue acts

1 INTRODUCTION

Conversational Agents (CAs) or chatbots are gaining a wide popularity. Applications such as Apple Siri [8], Amazon Alexa [7], and Google Assistant [34] are becoming ubiquitous. Suggested by its name, CAs interact with people in natural languages, thereby providing customer services [26] and companionship [95]. Compared to GUIs, CAs have shallower learning curves and may even form a relationship with people through conversations [27]. The digital health domain particularly benefits from the use of CAs: chatbots can be deployed to collect self-reports and provide personalized coaching to different individuals [77]. Studies have found that people are willing to engage with chatbots and provide valuable information, such as self-reports [106] and survey responses [18], to chatbots [115, 116].

Many digital health applications that are designed to promote behavioral changes and health interventions require people's long-term adherence. Chatbots' ability to "converse" naturally, as distinguished from the GUI-based systems [67], has the potential to help people develop long-term adoptions for health monitoring [27, 29, 77]. However, existing commercial chatbot frameworks, such as Dialogflow[32] and Amazon Alexa [6], predominantly only support building rule-based and scripted chatbots [69, 106]. Lacking flexible flows, these chatbots usually appear robotic and unnatural [71]. Particularly, using rule-based chatbots to collect user reports may cause boredom in long-term deployments [99]. On the other hand, implementing chatbots that can have more diverse and dynamic conversations requires large and specific domain datasets [77]. For example, an open-domain chatbot Meena was trained on 341 GB of dialogue sessions [4]. Since creating such large datasets is costly, the datasets are often proprietary and inaccessible publicly. Furthermore, most demonstrations of open-ended chatbots focus on performing free-form conversations in general topics and do not support end-user customizations. Little research has been done to explore low-effort bootstrapping ways to build chatbots that can effectively perform pre-defined tasks, such as inquiring people about their health information and carrying on naturalistic conversations at the same time.

Recent Large Language Models (LLMs; *e.g.*, GPT-3 [17], PaLM [23], OPT [118], HyperCLOVA [41]), with billions of parameters pre-trained on a large amount of language corpora, provide new opportunities for conversational agents. The recently released ChatGPT has attracted over 1 million users within five days [5]. ChatGPT has exceeded many people's expectations by showing its vast amount of knowledge and the ability to converse in natural languages [2]. GPT-3.5 and 4, the backend models of ChatGPT, can be further fine-tuned and prompted to build specified chatbots that perform tasks such as acting as a virtual coach to effectively inquire people for certain information. By designing specific prompts, we can make GPT-3, or LLMs in general, produce human-like conversation responses accordingly without any training data, thereby functioning as a chatbot. Compared to other frameworks, LLMs show great potential in scaffolding chatbots that are sensible of contexts and even respond to off-topic user messages [103]. Further, LLMs may change the way

of people cooperate since they operate on natural language inputs. Previously, building chatbots may be limited to people with technical background. But LLMs enable people, such as medical practitioners, to have the opportunity to personalize or even build their own chatbots [27].

Despite these potentials, it is yet fully understood how LLMs *read* the prompt and *use* pretrained knowledge [17, 64], the development of prompts is usually conducted through iterative trial and error [66]. While the HCI and CSCW community have actively explored the use of LLMs in various domains (*e.g.*, [24, 53, 114]), research that leverages LLMs for powering chatbots, particularly task-oriented ones [12, 72, 103], is still sparse. Due to the inherent characteristics of LLMs, LLM-driven chatbots may be error-prone [48] or digress from their tasks [103]. Designing robust prompts is crucial for "restricting" chatbots to conduct desired tasks.

In this study, we investigate how LLMs can power chatbots to collect user self-reports while carrying on naturalistic conversations. Towards this aim, we built a set of chatbots (Figure 1) that run on GPT-3 [17] and converse to collect self-report data in four health-related topics—sleep, food intake, work and productivity, and exercise. We chose GPT-3 as an underlying LLM because it is one of the mainstream LLMs that are publicly available via commercial APIs. We formulate the model prompt to include the information slots (*i.e.*, information properties of a topic) that we intend the chatbot to collect and the job identities (*e.g.*, sleep expert for the topic sleep) to help drive the conversations. We investigate how two design factors in prompts—information specification format and personality modifier—impact the *slot filling ability* and the *conversation style* of chatbots. In total, we created 16 chatbots (4 topics × 2 formats × 2 personality modifiers) with different prompts.

To the best of our knowledge, our work is the first to explore the usefulness of LLMs in powering chatbots for collecting self-report data. We believe that well-designed prompts can effectively drive chatbots to perform specified tasks [104]. In the context of data collection through conversations, we evaluate chatbots from two perspectives: (1) slot filling performance, and (2) conversational styles. We conducted an online study (*N* = 48) with our chatbots on a web interface. All participants talked to chatbots of the four topics but each of them experienced chatbots run on the same prompt design. To the best of our knowledge, our work is the first to explore the usability of LLMs for building chatbots for collecting self-report data. We found that our zero-shot prompts, without either example dialogues or fine-tuning, covered 79% of the desired information slots among all dialogues. Through conversation analysis, we found that the information specification format as well as the use of personality modifier can impact the chatbots' slot-filling ability and conversation styles. Also, the chatbots generally reacted to participants' self-reported answers in an empathetic way, appreciating their accomplishments as well as sympathizing with participants for the negative outcomes. Consequently, some participants perceived these chatbots to be understanding and take into account their messages when responding, and others indicated that they were surprised to find the chatbots' responses were accurate and detailed.

The contributions from this work are threefold:

(1) Empirical results from a between-subject online study (*N* = 48), demonstrating the *feasibility* of chatbots powered by LLMs in not only carrying on conversations to collect specified information but also exhibiting abilities in maintaining context, state-tracking, and providing off-topic suggestions.

(2) Examination of how different prompt designs and other factors impact the chatbots' behaviors, providing insights for future researchers to easily scaffold chatbots through zero-shot prompting for data collection with LLMs.

(3) Implications on how future LLMs-driven chatbot platforms can improve the conversation quality, drawing on the analysis of the dialogue errors.

2 RELATED WORK

In this section, we cover related work in the areas of (1) self-report data collection through chatbots, (2) design considerations for chatbots, (3) chatbot platforms, and (4) designing LLM prompts for chatbots.

2.1 Self-Report Data Collection through Chatbots

Personal informatics systems have commonly incorporated data collection techniques to track personal health and activity [21, 58]. While various physiological or physical activity data—such as step count, heart rate, and sleep duration—can be captured automatically by sensors and wearable devices [50], various types of personal data still demand *self-reporting* by the person who self-tracks [21]. For example, food intake (*e.g.*, [25]) or work tasks (*e.g.*, [43]) are not reliably captured by sensors and thus require manual inputs. In addition, reflective questions (*e.g.*, Why did you eat this food? [68]) and subjective measurements (*e.g.*, Sleep quality) inherently require to be captured manually. A majority of digital self-tracking tools that involved manual data capture inherited the traditional concepts of self-monitoring or journaling and provide form-based GUIs such as a list of checkboxes and text fields [38, 46]. However, repeated manual input on a computer or smartphone screen is burdensome and may gradually disengage people from tracking [21, 22]. As an input modality to lower the capture burden and enhance the richness of the captured information, natural language has recently gained interest [68, 70]. Prior research found that when people are allowed to insert data in free-form natural language, they tend to provide detailed answers with surrounding contexts [44, 68]. Going further, conversational interaction, where a system and a user communicate in natural languages, has become one emerging interface for collecting self-reports.

Chatbots are considered easier to use and more accessible than GUIs as they minimize the use of graphical widgets employ the intuitive conversational interaction. Regarding data collection, a plethora of research has explored the use of chatbots in place of traditional form-based surveys (*e.g.*, [13, 18, 42, 116]). For example, studies with surveys with close-ended questions found that chatbots can collect the same quality, if not higher, user responses as GUIs [18, 42]. Xiao et al. [116] built a chatbot to conduct interviews with open-ended questions. Compared to the traditional web survey, their participants showed higher engagement and provided higher-quality responses when talking to the chatbot. Further, incorporating more humanized traits, such as casual conversation styles [42], self-introduction, and echoing [85], led to not only a higher level of user engagement and satisfaction but also more self-disclosure in responses. With more focus on self-reported data, prior studies leveraged chatbots to collect self-reports such as emotion (*e.g.*, [13]), pain level (*e.g.*, [113]), and food intake (*e.g.*, [77]). For example, Bemmann et al. [13] combined a chatbot with the experience sampling method (ESM, [51]) and found that personalized chatbots have the potential to collect data on sensitive or personal topics. Mitchell et al. [77] compared fully-scripted, rule-based, and retrieval-based chatbots for collecting food nutrition. They found the better fulfillment of data collection is not necessarily associated with the higher perceived quality of the chatbot as a diet coach, suggesting the importance of conversational content in user experience.

This work expands the line of research on chatbots that collect self-reports. In contrast to prior studies that involved predefined conversation logic or retrieval model training on domain-specific datasets, we explore the potential of LLMs in bootstrapping chatbots that can collect self-reports through conversations on four health topics—sleep, food, work, and exercise.

2.2 Design Considerations for Chatbots

Prior works in HCI explored user behaviors with chatbots and proposed suggestions to improve user experience with them. For example, Luger and Sellen [67] found that people restricted their

language uses when interacting with CAs. Jain et al. [37] revealed that many first-time chatbot users had disappointment and frustration with the selected chatbots: most chatbots lacked the ability to fully comprehend user messages or intentions. Since conversation breakdowns are still common [9], several studies have explored repair strategies, such as apologies, compensation, and providing option [54]. Ashktorab et al. [9] also evaluated other strategies, such as confirmation, repeat, keywords highlight & explanation, and recommended that chatbots should acknowledge misunderstanding in simple terms, explain model limitation in natural ways, and adapt individualized strategies. Although existing chatbot frameworks also have error recovery features, their features are not only limited but often cannot allow quick repairs [71, 79, 108].

Another key to improving the user experience is to make chatbots more playful and human-like [62]. The level of empathy [18, 84] and the repetitive rate [89] are two commonly used metrics of human-likeness. For example, the playful interactions (*e.g.*, telling jokes) or humorous responses enabled many people to start using CAs [67] and it is crucial for chatbots to support sustainable playfulness [91]. Also, human-like features and fun personalities are found to make chatbots more enjoyable to interact [37]. Even for work-related chatbots, some people still preferred chatbots that were human-like [61], and Liao et al. [62] envisioned that a reusable conversational module including common chit-chats and social attributes could be developed. In other words, future chatbot platforms should allow developers to easily build personalized chatbots with different personalities [102] and conversation styles [15, 49]. Lastly, developers should aim to improve chatbots' ability to maintain contexts to support smoother and natural conversations [9]. In this work, we investigate whether LLMs can steer chatbots that have social attributes and can resolve conversation breakdowns.

2.3 Chatbot Platforms

Building chatbots is challenging and time-consuming, and many design suggestions discussed above are difficult to implement. Many open-domain chatbots that engage and entertain people socially are predominantly dependent on large datasets [4, 119]. In the HCI and CSCW community, rule-based dialogue systems are widely used. Celino and Calegari [18] built their survey chatbot with pre-defined conversation flows as they intended to avoid disappointments caused by the chatbot's inability to understand certain utterances [67]. Although rule-based chatbots are unlikely to cause breakdowns, the resulted rigid conversations can make people lose interest in the long term. On the other hand, Xiao et al. [116] built their survey chatbot using Juji [40], which automatically equip chatbots with rich existing conversational skills. Using the Juji GUI to add questions is relatively simple, but it is unclear whether developers can modify the chatbot's expressed personality. Lastly, other commercial chatbot frameworks, such as Dialogflow [33] and IBM Watson [36], also allow developers to build rule-based chatbots with GUIs [98]. However, creating more dynamic conversations usually requires programming skills. Even for professional developers, it is challenging to create well-designed conversational flows and pre-define user intents and chatbot messages [71]. Using LLMs to power chatbots is a new way to build chatbots [17, 103]. LLMs accept natural language prompts so that people without any knowledge of programming but are interested in building chatbots for data collection can create prompts [41]. Customizing prompts in natural languages essentially hands off the control to each individual who builds chatbots. As such, it becomes more straightforward to scaffold personalized chatbots (*e.g.*, assigning a preferred personality) by revising prompts accordingly. Nevertheless, it is unclear how to design prompts for LLMs to steer chatbots that can effectively ask questions around desired information and have different conversational styles.

2.4 Designing LLM Prompts for Conversations

Prompts are natural language texts to LLMs to produce desired outputs. With proper prompt inputs, GPT-3 can be used to translate texts, answer questions, write essays, and generate dialogues without any fine-tuning [17]. While the mechanism enabling such few-shot abilities behind LLMs is still veiled [17, 76], some prompting techniques are found to improve the model performance. One technique that surprisingly improves the generation quality is by conditioning the prompt with an identity. For example, by inserting the statement "You are an expert Python programmer" into prompts, models can generate higher quality codes [11], and similarly, "I'm a math tutor" is suggested to improve models' performance in solving math problems [10]. In HCI, researchers are also interested in investigating the design guidelines for prompts. Liu and Chilton [66] explored different design factors, such as phrasing, styles, subjects, and random seeds of prompts for text-to-image generative models. They found that the subject and style keywords are more important than the connecting words and the initializing seeds can significantly impact the quality of generations. Similarly, Lee et al. [53] discovered that the randomness impacted writing collaboration and new identity generations. In terms of dialogues, some studies designed prompts for LLMs to generate single-turn responses [105] and examined whether these artificial responses are comparable to human responses [96].

More recently, OpenAI released ChatGPT in November 2022 and its API subsequently in March 2023. ChatGPT is optimized to provide a conversational interface for users to interact with the LLM. The ease-of-use of ChatGPT attracts the attention of billions of users [3]. Practitioners have actively proposed open-source tools based on the ChatGPT API, such as LangChain [19] and AutoGPT [1], which support building conversational agents [101]. Researchers have also explored the usefulness of LLM-driven chatbots for various types of end-users (*e.g.*, programmers [86], socially-isolated people [39], and children [90]) as well as tasks (*e.g.*, UI task [104], medical information search [55]). Without fine-tuning, prompts are the key to drive chatbots converse effectively. However, despite providing used prompts, prior work did not investigate how to optimize prompts for domain-specific chatbots. Further, compared to previous task-oriented chatbots that mostly answer user questions, our study aims to build chatbots that can proactively ask users pre-defined questions. The challenge of our study is that we need to design prompts to power chatbots that (1) ask specific questions with domain knowledge and (2) can converge conversations on their own.

In summary, a plethora of prior works have leveraged LLMs to support conversations. However, few works has explored prompting chatbots that can *lead* conversations and perform question-asking tasks. In this study, we set out to address this gap and provide empirical insights into prompt designs for chatbots with different "job identities" that are dedicated to collecting self-reports. Further, we investigate whether LLM-driven chatbots can exhibit important features such social attributes [62], employ empathy [84], as well as have the ability to handle breakdowns [9].

3 METHOD

To examine how different prompt design factors impact LLM-driven chatbots' performance, we implemented a web-based chatbot interface and conducted an online user study to collect dialogue data from people. We describe our prompt designs and the experimental design of the user study.

3.1 The LLM-driven Chatbot Framework

LLMs power chatbots by generating the next response based on the input prompt. Figure 1 describes the mechanism of our chatbot running on an LLM. At each turn, the back-end system combines a prefix (Ⓐ in Figure 1) and the history of the ongoing conversation (Ⓑ in Figure 1) into a single prompt and feeds it into the LLM (Ⓒ in Figure 1). Then the LLM generates the following agent's

Table 1. The specified information slots and job identity for each conversation topic.

Topic	Job Identity	Information Slots
Sleep	Sleep Expert	(1) Time to bed (2) Sleep latency (3) Wake up at night (4) Wake up time (5) Sleep quality rating (1–10)
Food Intake	Dietitian	(1) Breakfast (2) Lunch (3) Dinner (4) Snacks (5) Feelings after eating
Work and Productivity	Life Coach	(1) Work done (2) Productivity rating (1–10) (3) Other concerns at work (4) What went well
Exercise	Fitness Coach	(1) What workout (2) Workout duration (3) Feeling after (skipping) workout (4) Fitness concerns

response. The prefix, treated as an instruction for an LLM, consists of the description of persona and the specification of the desired information slots that the chatbot should capture.

3.1.1 Prompt Designs for Chatbots. In this study, we envisioned a self-monitoring scenario where chatbots proactively interact with people and inquire them about their retrospective health-related behaviors of the previous day through natural conversations. Towards this goal, we first chose four health-related topics—**Sleep** [75], **Food** intake [28, 68], **Work** and productivity [43, 45, 59], and **Exercise** [69]—that retain interests not only by the research community and are also personal use in daily life [112]. Then, we defined the behavioral data (slots) that we intended to have chatbots to collect. Table 1 summarizes the information slots for each topic, which the chatbots should capture in each conversation session. Another aspect that we intended to explore was whether we could enable LLMs to steer chatbots to exhibit different conversation styles as different individual's experience with chatbots can be improved with personalizations [15].

Currently, there is a lack of prompt designs in general, let alone for chatbots. As suggested by earlier work, even NLP experts adopt trial and error and iterative experimentation when designing prompts before applying prompts in large datasets [117]. The designs of prompts largely depend on the intended task. The goal of this study is to understand how different prompts impact the data collection performance of chatbots. Prior to deploying GPT-3-powered chatbots to participants, we trialed several iterations of prompts to find satisfactory chatbot prompts among our research team. In our "trial and error" process, we judged prompts based on whether the resulted chatbots can fulfill the assigned tasks and not go off-topic.

Initially, we started with OpenAI's example prompt, which defines an "AI assistant" with several characteristics(*e.g.*, "`The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.`"). We modified the example prompt by including questions of interests (*e.g.*, "`How's your sleep last night?`") as well as by adding the list of information slots (*e.g.*, "`...The assistant asks questions about human's sleep time, sleep duration, and sleep quality.`"). However, we found that these modifications tended to make conversations digress and the resulted chatbots converse more similar to generic customer service chatbots.

Findings from research [11, 47, 109] and anecdotal evidences from social medias [11] suggest that specifying identities (*e.g.*, "Python programmer") can improve GPT-3 model performance. Inspired by this, we substituted the "AI assistant" with more specific job identities (*e.g.*, "fitness coach") and found that the resulted chatbots appeared to have more domain knowledge and are less likely to
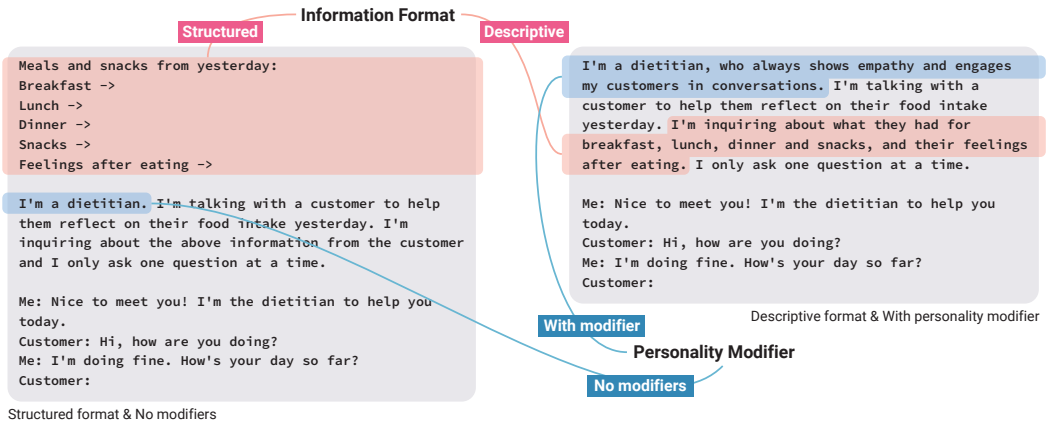
Fig. 2. Prompt design combining two factors, information format and personality modifier, in the Food intake topic.

digest in our tests. As such, we picked different job identities for each topic—*sleep expert* (Sleep), *dietitian* (Food), *life coach* (Work), and *fitness coach* (Exercise).

With the above modifications, we designed prompts by first describing a chatbot with a specific identity and then including a list of slots (*e.g.*, "I'm inquiring about what they had for breakfast, lunch, dinner, and snacks..."; see Figure 2, right). Through multiple trials, this design worked better than generic one despite still had occasional digressions. Hence, we experimented with another prompt format inspired by the state-tracking technique in task-oriented dialogues [52]. Instead of being described literally, the slots are structured into a form (*e.g.*, "Meals and snacks from yesterday: Breakfast -> [placeholder] Lunch -> [placeholder] ..."; we used an empty string as a placeholder; see Figure 2, left). Both designs performed similarly in our limited internal testings, hence we aimed to investigate the performance of two formats in the user study.

In terms of manipulating conversational styles, we introduced the use of a modifier in prompts– "who always shows empathy and engages my customer in conversations," to the prompt (See Figure 2, right). We hypothesize that with this modifier, the chatbot is more likely to express empathy in conversations and have a higher level of interactivity—*i.e.*, use more emphatic expressions and be more responsive to user responses. Conversely, without the modifier (See Figure 2, left), we expect the chatbot to be more neutral, formally exchanging messages with users and appear less empathetic. Lastly, during our trials, we found that GPT-3 had the tendency to ask multiple questions in one turn. To restrict this behavior, we added "I only ask one question at a time." to the prompt.

3.1.2 Model and Parameters. To power chatbots with above prompts, we used davinci-text-002, the largest and most capable model of GPT-3 as of June 2022, publicly accessible via OpenAI's API [80]. This model accepts 4,000 byte-pair encoding tokens at maximum in a prompt per request. Our prompt templates in the initial state were encoded into around only 120 tokens (3% of the limit) and allowed sufficient room for the appended conversation history. For all chatbots, we uniformly applied the same generative parameters: temperature as 0.9, the presence penalty as 0.6, and the frequency penalty as 0.5. We kept the temperature and the presence penalty unchanged based on OpenAI's suggestions and increased the frequency penalty to reduce the re-use of words.

3.1.3 The Web Chat Interface. We implemented a web interface to host our LLM-driven chatbots, following a typical chat interface design (See Appendix A.1). The webpage was written in Type-Script [74] on React [73] and runs on the Node.js [97] server. The server communicates with GPT-3 leveraging OpenAI's API [80]. To simplify the conversation flow, we disabled people to submit multiple utterances in a row. Correspondingly, the chatbots also delivered one utterance at a time. When a user submitted an utterance, the server appended the current dialog history at the end of the prompt template and fed it to GPT-3 to generate the following response.

3.2 Online Study

3.2.1 Experimental Conditions. Combining the two design factors, we created four designs of prompts: SP (**S**tructured format with **P**ersonality modifier), SN (**S**tructured format with **N**o personality modifier), DP (**D**escriptive format with **P**ersonality modifier), and DN (**D**escriptive format with **N**o personality modifier). Each participant was assigned to one prompt design and engaged in conversations of all four topics. (Refer to the supplementary material for all 16 variations of GPT-3 prompts created for combinations on topic and condition.) To mitigate the ordering effect among topics, half of the participants conversed in the order of Work–Food–Exercise–Sleep, and the other half in the order of Exercise–Sleep–Work–Food. Additionally, for each topic, we requested participants to engage with the chatbot twice: one in the **Positive** path (*e.g.*, report high-quality sleep) and one in the **Negative** path (*e.g.*, report poor sleep). The two paths served as a general guideline for participants to exercise their imaginations in conducting *open-ended* conversations with the chatbots. Participants could make up their own "stories" and we aimed to investigate whether and how GPT-3-powered chatbots can handle all sorts of user responses on their *own*. Refer to Appendix A.2 for an exhaustive list of paths and hints by topic provided by us to guide participants to compose their answers for each path accordingly.

3.2.2 Web Chat Session. After signing an electronic consent form on the study website, participants went through eight conversations (4 *topics* ∗ 2 *paths*). On the web chat interface (See Appendix A.1), we put guidelines including the instructions and the conversation path that participants should follow (See Blue text in Appendix A.1, right). Since we did not incorporate ending detection algorithms, we asked participants to click the 'Next' button to proceed to the next conversation when they thought the conversation was naturally over or the chatbot kept sending repetitive messages. The completed dialogues were stored in our server.

3.2.3 Exit Survey. After completing eight conversations, the web page automatically redirected participants to an online survey. The survey consisted of three 5-point Likert scale questions and one open-ended feedback textfield. The Likert scale questions were: (1) "*Do you think the chatbot understands your answers?*" (2) "*Do you think the chatbot takes into account of your answers when responding?*" and (3) "*Do you think the chatbot talks more like a human who shows more empathy or more like a robot who behaves mechanically?*" The open-ended feedback question stated, "*If you have any other comments or thoughts about the chatbot (e.g., things that you've liked or disliked), please share with us.*" The first two questions can measure whether participants think the chatbot acknowledges their answers and respond accordingly and the third question is an overall measure of whether the chatbot is perceived as being empathetic. With participants' subjective evaluations, we hope to see whether the personality modifier can impact the chatbot's way of talking.

3.3 Participants

We recruited participants by word-of-mouth and posting advertisements at a large tech company, social media, and online forums in local universities. We sent the link to our study website to 83 people who filled out a screener and met our inclusion criteria: (1) aged 19 or older; (2) fluent

Table 2. Participant demographics by experimental condition.

		SP	SN	DP	DN
Age	Mean (min−max)	31.5 (21−56)	28.0 (21−33)	27.5 (20−40)	30.4 (19−42)
Gender	Male	7	8	8	7
	Female	5	4	4	5
English Proficiency	Native/Bilingual	4	2	4	4
	Proficient	8	10	8	8
Education	High school	1	2	1	1
	Bachelor	5	4	3	3
	Master	4	5	7	7
	Doctor	2	1	1	1
Familiarity with LLMs	Often use it	2	1	2	1
	Occasionally use them	2	2	2	4
	Used them once or twice	2	4	2	2
	Never heard of/used them	6	5	6	5
Participants	Total	12	12	12	12

English speaker; and (3) have the experience in talking to chatbots of any kind. 54 people completed the online study session and submitted an exit survey. The entire study lasted less than 20 minutes and all participants received e-gift cards (equivalent to $5 USD) after they completed the study.

We excluded six people's data from analysis; one made significant amount of grammatical errors and the rest completed less than half dialogues. Table 2 summarizes the demographic of the final 48 participants (aged 19 to 56, 18 females). Fourteen out of 48 (29%) participants were native/bilingual and 22 out of 48 (46%) participants had never heard of or used LLMs. Each prompt design condition included 12 participants.

3.4 Data Analysis

We collected rich dialogue data and valuable user subjective evaluation feedback. We performed both quantitative and qualitative analysis to examine chatbots' conversation styles, the slot filling performance, and participants' experiences with our chatbots. For each dialogue, we calculated commonly used descriptive metrics such as the number of turns and the average word counts per turn, which we report in Section 4.1.

Slot Filling Performance. Our study aimed to investigate whether LLMs can drive chatbots to effectively ask defined questions and collect desired information specified in Table 1. To calculate the amount of information that can be obtained by our chatbots, one researcher *manually* inspected and determined whether each of the pre-defined information slots could be extracted from collected dialogues. More specifically, for *sleep quality* and *productivity rate*, which were specified as a scale of 1 to 10, we marked the slot as filled only if a numerical value (*e.g.*, 9) rather than a vague phrase (*e.g.*, good sleep) was given. For *feelings after eating* in Food, we treated the slot to be filled if feelings regarding one or more meals were covered. We report the analysis of slot filling rate in Section 4.2. Based on the binary coding, we calculated the **slot filling rate**: the ratio of the number of information slots extracted from the dialogue against the total number of slots. We use the slot filling rate to infer the data collection performance of chatbots.

Dialogue Acts and User/Chatbot Behaviors. To understand the conversational behaviors of the chatbots, we coded *dialogue act* for each turn of conversations. Referring to some existing taxonomies of dialogue acts [92, 100, 110], three researchers independently coded one participant's dialogues (132 turns; 1.8%) to identify emerging dialogue acts. Additionally, researchers labeled chatbot turns that did not fit in the conversation context or originated from the inherent artifacts of an LLM. We resolved discrepancies in coding and developed the first version of codebook with three dimensions of codes: (1) essential acts and (2) empathy & engagement behaviors, and (3) problematic chatbot turns. Then two researchers reiterated the independent coding of four other participants' dialogues (1 participant from each condition, 32 dialogues in total) with the codebook. The two researchers resolved discrepancies through multiple sessions of discussion until their inter-rater reliability (Cohen's Kappa) reached 0.96 for essential acts and 0.935 for empathy & engagement behaviors. Compared to these dialogue acts, the occurrence of errors was sparse. Hence, the two researchers discussed the entire problematic turns coded by each other together and reached the full agreement. With the finalized codebooks (See Table 5, 6, and 7), the first author coded the rest of the data. As a result, each turn was classified as one of the essential acts—*greeting, task opening, required question/answers (RQ/RA), secondary question/answers (SQ/SA), statement,* and *closing*. We assigned the most prominent act to turns consisting of multiple sentences. Independent of essential acts, we multi-coded each turn with the empathy & engagement behaviors described in Table 6. For example, to a general compliment 🤖 "*That's great*" (**Statement**), we assigned only the **Appreciating** behavior, whereas we also treated 🤖 "*That's great to hear that your legs are feeling stronger!*" (**Statement**) to be both **Acknowledging** and **Appreciating** as the compliment directly addressed to the user input. We were interested in such acknowledging behaviors because *specificity* was an important indicator of the capability of open-domain chatbots [4].

Statistical Analysis. To understand the impact of the study factors, including prompt design, conversation topic, and the conversation path, to the chatbots' slot filling performance and conversational flows, we used *mixed-effect models* because these models can handle unbalanced data repeatedly measured from the same participants [83]. For each dialogue metric we want to assess, we fitted a mixed-effect model that predicts the metric, treating each dialogue as a data point. Starting from a full model containing participants as a random effect and the four main study factors–information format, personality modifier, topic, and path–and their interactions as fixed effects, we performed the step-wise backward elimination removing variables not significantly contributing the model, through Maximum-likelihood tests. For significant variables, we performed post-hoc pairwise comparisons of the least-squared means (LSM) of the metric using `emmeans` [56] package in R.

Subjective Feedback. To assess the difference among the experimental conditions, we conducted Kruskal-Wallis tests over the four rating questions. We also referenced the open-ended feedback from when interpreting the participants' reactions to specific phenomena of the conversations.

4 RESULTS

In this section, we report the results of our study in six parts. In Section 4.1, we provide an overview of the dialogue dataset we collected. In Section 4.2, we report the data collection performance of our chatbots and factors that impact the performance. In Section 4.3, we report the types of the essential dialogue acts and assess how the prompt design and other factors impact the dialogue acts and, in turn, the data collection performance. In Section 4.4, we report the types of the empathetic and engaging behaviors of chatbots and assess how the prompt design and other factors impact such behaviors of the chatbots. In Section 4.5, we explore the problematic chatbot utterances mainly

caused by the erroneous behaviors of a large language model. Lastly, in Section 4.6, we report on participants' subjective evaluation from the exit surveys.

4.1 Descriptive Statistics

From 48 participants, we collected 374 dialogues (7,442 turns in total): 91 from SP; 96 from SN; 95 from DP, and 91 from DN. Regarding the conversation topic, 94, 91, 94, and 95 dialogues were from Sleep, Work, Exercise, and Food Intake, respectively. Eight participants missed one dialogue per each and one missed two, mainly due to temporary server issues or accidental skips.

Prompt designs impacted the word lengths and the number of turns of chatbots. Table 3 summarizes the number of turns and word counts by prompt design. The average number of turns per dialogue is around 20 with more average turns under the two descriptive conditions (DP and DN). The maximum number of dialogue turns is 75 under the DP condition (only 1 dialogue). In terms of word counts, dialogues under the descriptive conditions (DP, DN) had more words than those under the structured conditions (SP, SN): both chatbots and participants uttered more words under the descriptive conditions. The DP condition, in particular, leads to the most number of words of dialogues.

4.2 Slot Filling Rate

Prompt designs significantly impacted the slot filling performance of chatbots. Table 4 summarizes the average slot filling rates of chatbots by conditions and topics. On average, all chatbots have reached over 70% slot filling rates. The dialogues in the SP-Exercise condition had the highest rate (93%) and those in the SN-Work condition had the lowest rate (64%). The maximum-likelihood test revealed that there was no significant random effect of participants, indicating that participants have little impact on chatbots' data collection performance. On the other hand, there were significant random effects of the topics($p < .0001$), the conversation paths ($p = .01$), and the interaction between the information formats and personality modifiers ($p < .001$). Figure 3 shows the significance over

Table 3. Descriptive statistics of our dialogue dataset aggregated by four prompt designs.

	SP	SN	DP	DN
Total number of dialogues (turns)	91 (1,638)	96 (1,889)	95 (1,941)	92 (1,975)
Average no. of turns per dialogue (range)	18.0 (7–45)	19.7 (3–57)	20.4 (7–75)	21.47 (3–53)
Average no. of words per dialogue	212.3	240.8	321.7	277.1
Average no. of chatbot/user words per turn	17.4 / 4.9	17.8 / 4.8	23.4 / 7.5	19.2 / 5.5
Percentage of organically ended conversations	71.4%	76.0%	77.9%	78.2%
Percentage of erroneous turns	3.1%	4.3%	3.0%	3.6%

Table 4. The slot filling rate (and *SD*) by topic and condition.

	Sleep	Work	Food Intake	Exercise	Total
SP	0.83 (0.29)	0.71 (0.25)	0.85 (0.18)	0.93 (0.14)	0.83 (0.23)
SN	0.75 (0.30)	0.64 (0.32)	0.80 (0.33)	0.88 (0.20)	0.77 (0.30)
DP	0.67 (0.24)	0.67 (0.28)	0.72 (0.32)	0.82 (0.20)	0.72 (0.27)
DN	0.85 (0.16)	0.83 (0.24)	0.75 (0.24)	0.91 (0.16)	0.83 (0.21)
Total	**0.77 (0.26)**	**0.71 (0.28)**	**0.78 (0.28)**	**0.88 (0.18)**	**0.79 (0.26)**

the 95% confidence intervals of the slot filling rate in each category of the significant variables. The dialogues in DP condition had significantly lower rates than those in SP ($p = .01$) and DN ($p = .008$). This suggests that the personality modifier impacted chatbots differently: with the modifier, chatbots with the structured prompt yield higher rates whereas chatbots with the descriptive format yield higher rates without the modifier (See Figure 3a). In terms of topic, Exercise dialogues had the highest rate of 88.4%, which was significantly higher than those in the other three topics: Sleep ($p = .01$), Work ($p < .0001$), and Food ($p = .02$) (See Figure 3b). Lastly, dialogues in the Positive path had significantly higher rates than those in the Negative path ($p = .01$) (See Figure 3c).

As seen in Figure 4, there is a general trend that slots specified earlier in prompts were more likely to be covered by chatbots. For example, the first slots in all topics were covered in 90.3% of the dialogues, but the last specified slots in Sleep (sleep quality) and Work (what went well) were omitted around 40% of the dialogues. Interestingly, the last specified slot of Food (feelings after eating) was diligently covered: chatbots often asked how participants felt after talking about each meal rather than asking their feelings once towards the end.

4.3 Essential Dialogue Acts

To further understand how chatbots powered by different prompt designs talk, we categorized conversation turns into dialogue acts. We provide the summary of essential dialogue acts and their distributions in Table 5 Here, we report chatbots' essential acts regarding question/answering and non-question statements.

4.3.1 Required and Secondary Questions. We identified two types of questions that the chatbots asked: required questions (RQ) and secondary questions (SQ). The RQs were directly related to the specified information slots, whereas SQs were not directly related to the information slots but rather follow-up details or elaboration. Despite being relevant to the conversation topic, SQs sometimes caused the conversation to digress. Although not very common (95 out of 1,029 SQ turns in total; 9.3%), participants also asked questions to the chatbot, which were all categorized as SQ/SA. The majority of collected dialogues consisted of question/answering: Overall, 4,879 out of 7,442 turns (*avg.* 64.72% of turns per dialogue; $min = 10.67\%$, $max = 93.54\%$) were classified as RQ, RA, SQ, or SA (See Table 5).

We first investigated the impact of prompt designs on the chatbot-spoken RQ and SQ turn ratios using two mixed-effect models with each turn ratio as a dependent variable, respectively. Figure 5 shows the 95% confidence intervals of the chatbot-spoken RQ and SQ turn ratios by four study factors (information format, personality modifier, topic, and path). The structured format
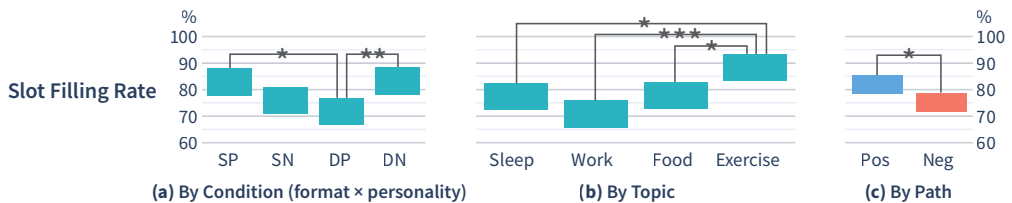

Fig. 3. 95% confidence intervals of slot filling rate by variables with a significant effect: (a) the combination of the information format and personality modifier represented as study condition; (b) topic; and (c) the conversation path. The asterisks with arms indicate significance between the connected categories. (Refer to Appendix A.3 for model details and statistics.)
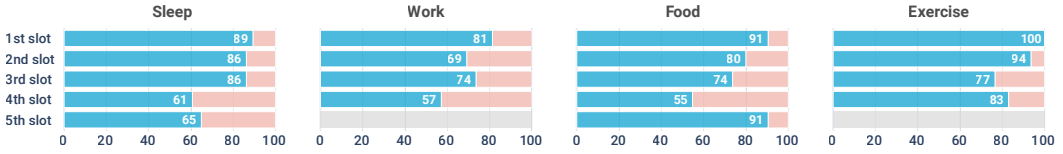
Fig. 4. Breakdowns of the percentage of filled slots by the order of questions for each topic. Work and Exercise consist of four slots.

significantly increased the RQ turn ratio ($p = .03$; see Figure 5a) but decreased the SQ turn ratio ($p = .002$; see Figure 5e). On the other hand, the personality modifier did not impact either RQ ($p = .94$) nor SQ ($p = .43$) turn ratios and made no difference within the same information format (See Figure 5b and 5f).

In terms of the conversation path, we find that, overall, the Positive path led to a higher RQ turn ratio ($p = .01$; see Figure 5c) and a lower SQ turn ratio ($p = .002$; see Figure 5g). However, under different topics, the conversation path had different impacts on RQ and SQ ratios. As seen in Figure 5d, the Positive path increased the RQ turn ratio only in the Work ($p = .02$) and Exercise ($p = .003$) dialogues and also decreased the SQ turn ratio in the same topics (Work: $p = .003$ and Exercise: $p < .001$).

As discussed above, prompt designs, topics and conversation paths have significant impacts on chatbots' question-asking behaviors. The RQ and SQ ratios further impacted the slot filling rate. We ran the maximum-likelihood tests with two mixed-effect models fitting the slot filling rate, one with the chatbot-spoken RQ turn ratio (*i.e.*, the ratio of the turns classified as RQ in a dialogue) as a fixed effect and the other with the SQ turn ratio, both with participants as a random effect. We found that the RQ turn ratio was positively correlated with the slot filling rate, whereas the SQ turn ratio was negatively correlated with it: $\beta = 1.08$, $SE = 0.12$, $t(347.25) = 8.80$, $p < .0001$ for RQ and $\beta = -0.77$, $SE = 0.13$, $t(370.06) = -6.12$, $p < .0001$ for SQ.

Table 5. Summary of essential dialogue acts with the ratio of the occurring turns per dialog (turn ratio) by condition, brief description, and exemplar turns (👤 : user turns, 🤖 : chatbot turns).

Dialogue Act	Turn Ratio (%)				Description	Examples
	SP	SN	DP	DN		
Greeting	12.40	12.94	10.67	10.50	Initiation of a conversation.	🤖 *How's your day so far?* 👤 *I feel refreshed and recharged.*
Task opening	1.92	1.56	0.54	0.04	General questions that bring up the conversation topic.	🤖 *How was your work and productivity yesterday?*
Required question	24.07	21.36	17.57	20.97	Questions and answers that are directly related to the specified information slots.	🤖 *What was your lunch yesterday?* (RQ) 👤 *I had pork barbeque.* (RA)
Required answer	23.74	19.91	17.35	20.75		
Secondary question	10.28	10.79	13.87	13.91	Questions and answers that are usually follow-ups and not specific to the slots.	🤖 *What did you put on your toast?* (SQ) 👤 *I put strawberry jam on it.* (SA)
Secondary answer	9.12	9.64	13.08	12.71		
Statement	14.22	19.54	22.85	16.94	Non-Q&A messages such as commenting or summarizing.	🤖 *It sounds like you had a great night's sleep!* 👤 *I did! Not sure why though.*
Closing	3.88	4.31	3.96	4.14	Farewell or ending messages.	🤖 You're welcome. Have a great day!

In summary, our results suggest that chatbots with a Descriptive information format tend to ask more secondary questions, and negative answers of participants also naturally elicit more secondary questions (*e.g.*, 🤖 *"I'm sorry to hear that you didn't workout yesterday. May I ask why?"*).

4.3.2   Non-question Statements. Statements from chatbots usually consist of chitchats, comments and suggestions that do not fall within the question/answering (see Table 5 for references) and participants also tend to respond to statements with other statements. Usually, the participants' statements did not contain direct clues about the desired information slots. Overall, 1,482 out of 7,442 turns (*avg.* 18.45% per dialogue) were classified as Statement (See Table 5). Among the four conditions, dialogues of DP have the highest ratio of statements (24.99%) and SP had the lowest ratio (15.32%).

4.4   Empathy & Engagement Behaviors

Table 6 summarizes the empathy & engagement categories and their turn ratios by experimental condition. The majority of these behavior categories were coded to the chatbot turns—1,992 chatbot turns and 403 user turns were coded with one or more behavior categories—partly because participants uttered less words than chatbots (see Table 3) and chatbots usually led the conversation while participants simply responded. **Acknowledging** was the most common empathy behavior (see Table 6) as chatbots often referred to what participants previously said in generated messages. Also, we found that our chatbots often **appreciated** participants' accomplishment (*e.g.*, taking good sleep, managed to exercise) or **sympathized** participants when they reported negative outcomes (*e.g.*, poor sleep quality, failed at work).

To investigate how the four study factors impact the empathetic behaviors of chatbots, we analyzed three mixed-effect models with the chatbot turn ratios of Acknowledging, Appreciating, and Sympathizing behaviors as a dependent variable, respectively. Figure 6 shows the 95% confidence
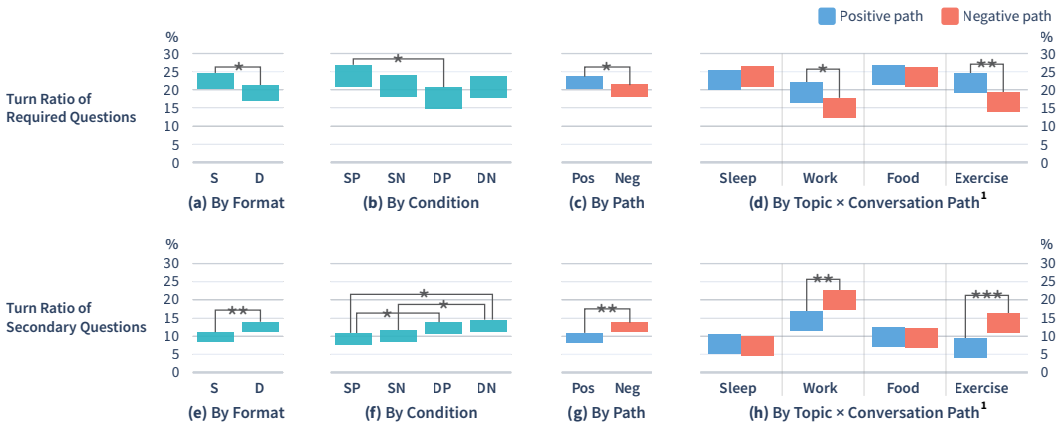
Fig. 5. 95% confidence intervals of the turn ratios of RQ (top; a–d) and SQ (bottom; e–h) by variables with a significant effect: The asterisks with arms indicate significance between the connected categories. Note that for (d) and (h) we did not display the significance across topics. (Refer to Appendices A.4.1 and A.4.2 for model details and statistics.)
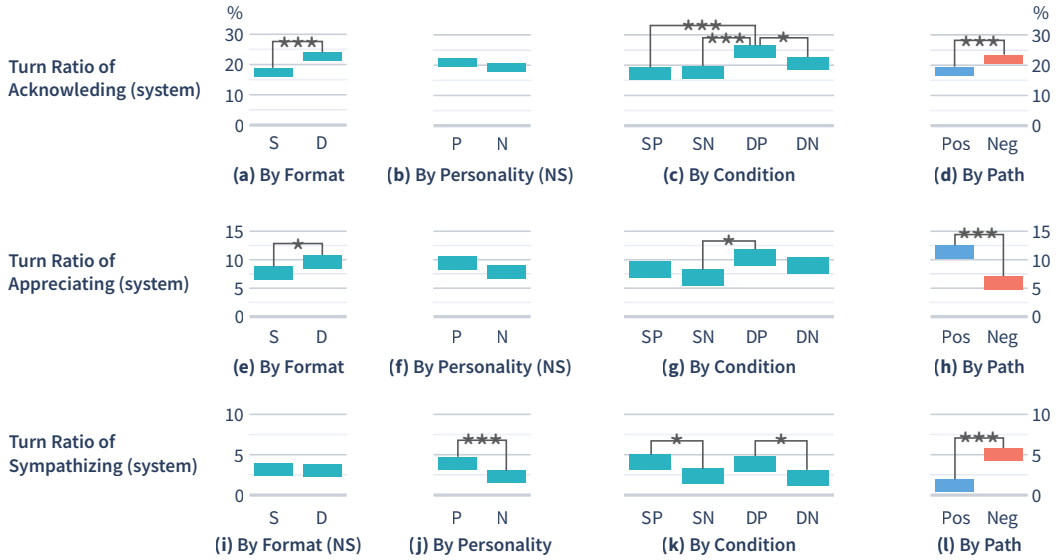
Fig. 6. 95% confidence intervals of the chatbot turn ratios for Acknowledging (a–d), Appreciating (e–h), and Sympathizing (i–l) behaviors by information format, personality modifier, study condition (combinations of format and personality modifier), and conversation path. Variables that are not significant are marked as 'NS.' (Refer to Appendix A.5 for model details and statistics.)

intervals of turn ratios of the three behavior categories estimated against the study factors. The information format significantly influenced the Acknowledging and Appreciating turn ratios: Dialogues in the Descriptive format had higher ratios of the Acknowledging ($p < .0001$; see Figure 6a) and Appreciating ($p = .01$; see Figure 6e) turns. Personality modifier did not solely impact these two behaviors but it influenced in conjunction with the information format (See Figure 6c and 6g). However, the personality modifier in the prompt led chatbots to produce significantly more Sympathizing turns ($p = .002$; see Figure 6j). Besides the prompt design, the conversation

Table 6. Summary of empathy & engagement behaviors with turn ratio by condition, brief description, and exemplar turns. Note that the behaviors are multi-coded.

Behavior Category	Turn Ratio (%)				Description	Examples
	SP	SN	DP	DN		
Acknowledging	18.22	18.91	26.24	22.04	Directly referring to what the other said.	*That's great to hear that your legs are feeling stronger!*
Appreciating	9.04	7.36	11.18	9.86	Complimenting the other.	*That's terrific!*
Sympathizing	3.66	2.79	4.33	1.74	Sympathizing with the other.	*I'm sorry to hear that. What's been going on?*
Thanking	5.16	6.74	7.05	7.26	Being grateful to the other.	*I feel nice. Thanks for asking.*
Advice/ suggesting	2.54	3.07	4.81	2.55	Giving advice or suggestion.	*I can give you some recommendations on exercises that will help you grow your adductor muscles.*
Rejecting/ disagreeing	0.57	0.66	0.40	0.66	Rejecting or disagreeing with the other.	*But the weather needs to be good for walking.*

path strongly influenced all three empathetic behaviors: The Positive path led to higher turn ratio of Appreciating ($p < .0001$; see Figure 6d) whereas The Negative path led to higher Acknowledging ($p < .001$; see Figure 6h) and Sympathizing ($p < .0001$; see Figure 6l) turn ratios.

4.5 Problematic Chatbot Turns and User Responses

In total, 6.7% of the chatbot turns (257 out of 3,916) were tagged erroneous and the four categories of erroneous turns are summarized in Table 7. These erroneous turns sometimes led to the unorganic termination of the conversation (*i.e.*, participants ended the conversation before or without natural Closing messages). In the following, we cover these error types in detail.

Incorrect phrases. In structured prompts (SP, SN), we used a symbol "`->`," a commonly-used delimiter for key-value pairs in LLMs, to specify the information slots. This caused GPT-3 to expose such an information structure to the output as an artifact in 16 turns of messages, all of which were generated by the structured format. In the example from Table 7, GPT-3 even incorrectly "predicted" the slots (45 minutes, example (1) in Table 7) altogether with symbols. Other times, GPT-3 also erroneously predicted the answers for participants (example (2) Table 7). Particularly, we identified 6 instances where GPT-3 predicted the user response and appended an extra 👤 turn to the generated chatbot turns (3 turns in descriptive groups and 3 turns in structured groups). While the exposure of the symbol was caused by the structured prompt design, other incorrect phrases were mostly caused by the random nature of LLMs [17].

Self-talk. GPT-3 sometimes generated turns in a first-person narrative or not directed to participants, which looked quite similar to the "self-talk" of humans. In Fragment 1, for example, Turn 03 is obviously not directed to participants. In our dataset, less than 1.1% of chatbot turns were self-talk (SP: 8, SN: 27, DP: 1, DN: 9). Participants who encountered these "self-talk" commented that these messages were "awkward" (P42), "strange" (P24), and "confusing"(P23). However, we found that participants always attempted to continue the conversation by following the self-talk and tried to resolve the errors (See Turn 04 in Fragment 1). We used the first-person narrative in prompts as it appeared to make generated chatbots talk more like humans; however, the first-person narrative might negatively lead to the "self-talk" in conversations.

Table 7. Categories of the chatbots' erroneous turns with turn ratio, brief descriptions, and examples.

Category	Turn Ratio	Description	Example
Incorrect phrases	0.79%	Messages with symbols or extra predictions.	🤖 *Workout duration -> 45 minutes* (1) 🤖 *It took me a while to fall asleep.* (2)
Self-talk	1.1%	First-person narratives or commentary messages.	🤖 *The customer's fitness concern is that they are feeling very tired after their workout.*
Repetition	3.8%	Repeating the same or similar utterances multiple times.	Refer to Fragment 2.
Miscellaneous	0.9%	Other miscellaneous errors.	🤖 *(no response)* (1) 🤖 *Other than work, what were your concerns at work?* (2)

Fragment 1. Dialogue snippets of self-talk from P43.

01	🤖	*Do you have any fitness concerns that you would like to address?*
02	👤	*nope not at the moment.*
03	🤖	***Based on the customer's responses, it seems that they did not have a very active day yesterday. I would encourage them to try and get some form of physical activity in today, even if it is just a short walk.***
04	👤	*yes I should probably try it tomorrow.*

**Repetition.** We found that GPT-3 was susceptible to generate repetitive messages, either identically or linguistically repeating the previous chatbot turns. In total, 147 turns (3.8% of chatbot turns) were labeled to be repetitive (SP: 11, SN: 39, DP: 48, DN: 50). Identically repetitive messages occurred to 23 participants (SP: 3, SN:5, DP: 6, DN: 9) in 31 dialogues (SP: 3, SN: 9, DP: 8, DN:10). Among the four topics, work (14 dialogues) tended to have more identically repetitive messages and exercise (2 dialogues) tended have fewer identically repetitive messages. However, these messages usually served as SQ/SA in conversations, hence they rarely influenced slot filling and data collection. Linguistically repetitive messages usually share similar wording or phrase structures. Fragment 2 presents an example dialogue snippet. At Turn 01, 03, 05, and 07, the chatbot always started with a similar phrase (*i.e.*, 🤖 "*That's great ...*") to compliment the participant then asked the participant a question started with "*can you*". These linguistically repetitive messages were semantically correct and *58.3%* took place in the Acknowledging turns where the chatbot rephrased what the participant said progressed the conversation organically. However, linguistically phrased messages looked too similar, and they negatively impacted the user experience. A few participants suggested that the chatbots seemed to actually understand their responses, yet were using "a sentence template" (P29) to respond in a "predefined ways" (P13). On the other hand, repetition could also lead to "dead loops" of conversations. Repetitions were mostly caused by principles of LLMs. As the conversation history was used as the new prompt, GPT-3 tended to generate patternized texts, which led to further repetitions.

Fragment 2. Dialogue snippet of linguistically repetitive utterances from P37.

01	🤖	*That's a specific and achievable goal. Can you tell me more about what libraries you need to install?*
02	👤	*I don't know yet which libraries I need, but probably Pytorch, ...*
03	🤖	***That's a great goal. Can you*** *research what libraries you need and then install them?*
04	👤	*The libraries I need is provided in a file.*
05	🤖	***That's great.*** *So you have a list of the libraries you need.* ***Can you install them?***
06	👤	*I need to run a script that download and install these libraries.*
07	🤖	***That's a great goal. Can you*** *walk me through the script?*

Miscellaneous. There were 37 problematic chatbot turns that did not fall into the above categories. Among these turns, 19 turns were tagged to have context errors (*i.e.*, chatbots did not grasp the context at all), 10 had semantic errors (*i.e.*, messages that are not of human common sense), and one had both context and semantic errors. In Fragment 3, for example, the chatbot entirely missed that the participant said 👤 "*1 hour of cardio*." But the chatbot also attempted to resolve the contextual misunderstanding by apologizing after the participant corrected it. While chatbots in our study appeared to understand people pretty well in most cases, their responses with semantic errors could be quite wrong and amusing. For instance, when being asked about workout yesterday, one participant told the chatbot that they skipped. Instead of considering the participant did not workout, the chatbot considered skipping as a jumping workout and responded 🤖 "*Skipping is a type of*
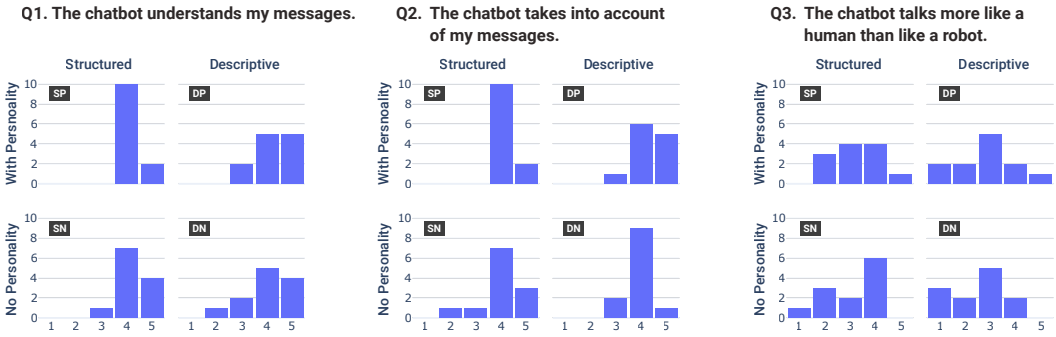
Fig. 7. Distributions of the subjective ratings from three scale questions in the exit survey, with breakdowns by the information format and the personality modifier.

cardiovascular exercise that can help to improve your heart health and endurance." Besides, there were two instances when the chatbot failed to detect the ending of the conversation and restarted with the first slot question again, which, of course, led participants to abandon the conversation. These errors were rare but could also be attributed to the randomness nature of LLMs. Lastly, our system went offline 5 times and caused chatbots to output empty messages, which was caused by the over-frequent API calls to OpenAI.

Terminating Conversations. With the Closing turns, we found that 75.7% of the conversations were organically ended. Among the four conditions, SP had the lowest percentage of naturally ended conversations (71.4%) than other three conditions. For 91 non-organically ended conversations, participants abandoned 48.3% of those conversations without encountering any obvious problematic errors. 19.8% of prematurely ended conversations were caused by identically repetitive and another 19.8% of were caused by linguistically repetitive messages. Lastly, the self-talk led to 3 conversations to end early, and context errors and system-offline caused the rest incomplete conversations.

Fragment 3. Dialogue snippet of the context error from P6.

01	🤖	*What workout did you do yesterday?*
02	👤	*I did **1 hour** of cardio yesterday.*
03	🤖	***How long did your workout last?***
04	👤	*Come on dude. I just told you I did 1 hour of cardio. How long is 1 hour?*
05	🤖	*I'm sorry if I came across as rude. I was just trying to get some information from you about your workout yesterday.*

4.6 Subjective Evaluation

Figure 7 summarizes the distribution of participants for rating (1) the ability to understand, (2) the ability to acknowledge user input, and (3) the level of empathy. The Kruskal-Wallis tests showed that there were no significant differences among conditions for all three questions. In general, most participants highly rated for Q1 and Q2 that the chatbots could understand them as well as acknowledge their messages: Fifteen (31.3%) participants rated 5 and 24 (50.0%) rated 4 on Q1; 11 (22.9%) rated 5 and 31 (64.6%) rated 4 on Q2. Participants showed mixed perception for the level of empathy question with a median of 3.

Some participants gave positive feedback in the open-ended question. P25 who frequently used LLMs commented, "*I was surprised to see how accurate and detailed the bot's responses were.*" P36

who did not have any experience with LLMs gave a similar comment—"*I found it quite responsive and surprisingly considerate of my answers.*" Despite the errors we presented above, P21 still complimented the chatbots: "*I liked/was satisfied of how the chatbot precisely gave info when I asked for it, and I felt that the relevance of the answer is very high and that it caught my point of question sharply.*" and even suggested that, "*I felt keeping chatbot as a companion would be awesome. To regularly make casual conversation and be provided light insights about my daily life based on my casual chats.*"

5 DISCUSSION

Our results showed that our zero-shot chatbots achieved great abilities in asking desired questions and understanding user responses despite also having drawbacks. In this section, we reflect on the performance of our chatbots and discuss opportunities, ethical considerations, and limitations for future work.

5.1 Designing Effective Prompts for Chatbots that Collect Self-Reports

Our study showed that defining a first-person job identity as well as specifying information slots in prompts was an effective method to bootstrap chatbots that ask health-related questions. However, the slot filling performance and chatbots' behaviors were sensitive to prompt design, topic, and conversation paths. We provide the following prompt design suggestions based on our findings:

Combine Information Format and Personality Modifier Wisely. Although the information format and personality modifier did not consistently impact slot filling rates individually, how they were combined had different impacts on slot filling rates. The information format affected chatbots' question-asking behaviors: structured formats lead to more RQs and fewer SQs and vice versa for descriptive ones. In other words, structured formats steer chatbots to ask direct questions about the specified slots whereas descriptive ones focus more on eliciting surrounding context or additional details. The personality modifier had a synergy with descriptive formats: Chatbots in DP had the lowest slot filling rates and the ratio of RQs, but had the highest number of acknowledging messages. Referring to Table 4, SP and DN have comparable slot filling rates. Therefore, to build chatbots that can show a higher level of understanding through acknowledgment, using the Descriptive format without personality modifier could be the best option. But chatbots with more direct acknowledgment may be at the risk of being awkward and too robotic. Hence, when designing prompts to power chatbots for data collection, using structured format with personality modifier would be more desirable.

Evaluate chatbot for conversation topic and path. There is certain discrepancy of slot filling rates between topics. One reason for such difference could be the nature of the topic. For example, the topic *work* tended to be the most open-ended topic as people report different types of work, which could lead to more subject switches and digressions than others. We suspect that GPT-3 is more suitable to steer chatbots that are of less divergent topics and collect self-reports that are more structured. Also, considering the path of conversations also impacts the data collection rate, researchers may consider clearly specify different slots for both positive and negative paths. For example, developers can add "if the customer did not workout yesterday, I would ask them what workout they did in the past week" to the prompt for chatbots in the exercise topic. Lastly, researchers can also test "neutral" path to see whether the chatbots can actively elicit rich user responses.

Composition of Slots Matters. The number and data types of slots also impact the chatbots' performance in collecting slots. With the conversation goes longer, chatbots have a tendency to miss information slots that appear later in prompts. Also, while we collected a certain amount of

numerical rates for sleep quality and productivity rate, it is not guaranteed that the chatbots would cover a slot definition (*e.g.*, numerical scale) as intended. Sometimes, the chatbot would simply ask 🤖 "*Would you say you had a good night's sleep?*" or 🤖 "*Overall, how do you feel about your work and productivity yesterday?*". Hence, if to use chatbots powered by GPT-3 to facilitate data collection, we suggest that important slots be put earlier in the prompt and the number of questions of specific data type be limited. If more data slots need to be collected, multi-stage prompts [114] can be considered.

5.2 Opportunities of LLM-driven Chatbots

From the study, we learned that LLM-driven chatbots are advantageous compared with traditional chatbot platforms in multiple aspects. Here we cover some noteworthy aspects drawing on the results.

Versatile Responses and Follow-up Questions. Compared to chatbots with pre-defined dialogues, chatbots in our study can deliver a great number of versatile phrases. For example, for these scale questions, GPT-3 can output phrases such as 🤖 "*Would you say that your sleep quality yesterday was a 10/10, 9/10, 8/10...?*" GPT-3 can even provide clarifications for questions and ask follow-up questions which supplement the topic. However, these SQs were still on-topic and directly addressed to user inputs (See SQ/SA in Table 5). Follow-up questions are commonly used in human-administered interviews to increase interactivity [87] and many studies suggest that chatbots that can ask on-topic follow-up questions are considered more human-like [67, 94]. Although current chatbot frameworks (*e.g.*, Amazon Alexa [7]) [71] support follow-up/extended questions, developers need to specify both the expected slots and the follow-up phrases [88]. On the contrary, GPT-3 could naturally ask follow-up questions, equipping chatbots with proper common sense on the topic. For example, our chatbot mapped "*Bulgogi, rice, and kimchi*" to "*a very traditional Korean meal*" in its response to the participant. Such response engages people through showing a level of "understanding."

Social Attributes. Given the importance of social features such as chit-chat for positive user experience [62, 91], our results show that we can easily equip GPT-3 with such social aspects. For example, our chatbots could respond naturally to the questions about their "personal life"—*e.g.*, 🧑 "*Do you workout yourself?*" 🤖 "*Yes, I work out regularly myself. I find that it helps me to stay energized and focused throughout the day.*" Further, our chatbots were also able to give suggestions relevant to the topic. While mostly originated from common sense, some of the suggestions were in-depth and tailored. In one time, one participant asked two questions in a row (probably due to system or network error) were quite surprised to find that the chatbot provided a well-written response (See Turn 04 in Fragment 6). This participant even commented that "*I know a small bit about NLP but not much when it comes to generating responses. I find it fascinating that (it) can give such in-depth answers to specific topics as I find it hard to be able to train an AI to every kind of case involving that.*"

Error Recovery. Task-oriented chatbots usually have limited number of pre-defined user intents to accomplish a specific goal. For instance, a banking chatbot can provide services such as currency-exchange conversion and introduction of credit cards [57]. However, such chatbots are usually unable to handle user messages that are out of the pre-defined intents (*e.g.*, a user attempts to have small talk with the banking chatbot) [57]. Also, they may even mis-recognize in-scope messages due to the complexity of natural language [9]. Strategies like highlighting keywords and switching topics [9] can help resolve conversation breakdowns at the price of making chatbots less human-like. In our case, LLM-driven chatbots could handle the out-of-scope conversations relatively well, since they could do improvisation actions relying on the ability of LLMs instead of defining intents

intensively. In Fragment 4, a work chatbot with the job identity "life coach" handled the off-topic request ("*wake me up at 6 am*") by the participant smoothly and even provided tips on sleep. Even when misunderstanding occurred, chatbots sometimes attempted to resolve it. In Fragment 3, for example, the chatbot apologized for its misunderstanding and in5, GPT-3 resolved an empty message error (due to system offline) by making up an excuse for its absence.

Fragment 4. Dialogue snippet of the bot continuing off-topic conversations from P34.

01	👤	*Can you wake me up at 6 am?*
02	🤖	*I'm sorry, but I cannot physically wake you up at 6 am. However, I can give you some tips on how to wake up earlier and be productive in the morning.*
03	👤	*Yeah, please.*
04	🤖	*Some possible tips to help you wake up earlier in the morning are to set your alarm clock for a specific time and place it across the room ... [omitted]*

Fragment 5. Dialogue snippet of an error recovery from P2.

01	🤖	*That's great! What time did you wake up this morning?*
02	👤	*around 9 am.*
03	🤖	*(...no response...)*
04	👤	*hello?*
05	🤖	*Sorry, I'm just checking to see if you're still there. Yeah, so 9 am is a pretty good wake up time. How would you rate your sleep quality from last night?*

Context Tracking. Context is a key part in human conversations that connects multiple turns [16]. Previous studies have suggested that chatbots should aim to sustain contexts to improve the dialogue efficiency [37]. Current conversational interfaces such as Google Assistant and Amazon Alexa shows certain abilities in maintaining contexts [71]; however, most of them are still criticized for not detecting contextual details [14, 20]. With Dialogflow, developers can define some contexts to be maintained within 5 turns[1]; however, it has yet achieved truly flexible conversations through this approach. In our case, the chatbots have shown impressive abilities in sustaining some contexts without dedicated mechanisms for managing contexts. P27 noted, "*I feel like it could keep track of the context well between sentences during the conversation.*" Through dialogue snippets presented in Fragment 5 and 6, we can see that the context was maintained across 5 turns and 2 turns, respectively. In particular, the context (*intimidation*) would be difficult to specify with most chatbot frameworks. Further, one pattern that emerged in the dataset is that chatbots liked to give a summary of all the user input in the end of conversations, which usually covered the past conversation history and maintained contexts longer than 5 turns.

Fragment 6. Dialogue snippet of the chatbot tracking context from P33.

01	👤	*Those are good. How about intimidation at the gym?*
02	👤	*Oh sorry, was that too vague?*
03	🤖	*No, that's a great question. A lot of people feel intimidated when they first start going to the gym. My best advice would be to just take things slow and focus on your own workout. Don't worry about what other people are doing or how they're looking at you.* *Just focus on yourselfand your own goals.*
04	👤	*Thatś solid advice. You're surprisingly eloquent for these responses!*

---

[1]https://cloud.google.com/dialogflow/es/docs/contexts-input-output

Low-effort Bootstrapping. We show several opportunities of LLMs in powering chatbots above. Indeed, chatbots that provide many of the above functions, including chitchats, suggestions, and context perseverance, can be trained with rich datasets. However, collecting such dataset is challenging, and training models on big datasets is costly and often inaccessible [12]. In terms of utilizing mainstream chatbot platforms to build voice applications, it is of great human effort to come up with user intents and example phrases as well as design conversation flows and logic [27, 68]. In particular, the error handling is tricky to design as conversation breakdowns can be common and even unexpected in the wild [20, 79, 108]. On the other hand, LLM's in-context learning capability enables us to skip collecting training dataset or configuring dialogue flows to create functional chatbots. Further, our results show that simple alterations of prompts can significantly influence the conversation styles of chatbots. With robust prompt designs, it is possible that people without background in AI can directly personalize chatbots using natural language.

5.3 Drawbacks of LLM-driven Chatbots

Although LLMs showed great potential in steering chatbots, we also encountered several drawbacks of LLM-driven chatbots. In the following, we discuss the two noteworthy drawbacks and their potential causes. We also present strategies to overcome these drawbacks.

Randomness. LLMs generate text by predicting the most probable text following the input prompt. While developers can tune hyper-parameters such as temperature and frequency penalty, the LLM generations inherently exhibit a certain level of randomness [66], which is hard to explain or anticipate. Such randomness might have led to erroneous responses of our chatbots. For example, sometimes the chatbots "self-talked" [93] or exposed machine representations (symbols) as responses. What makes it worse is that as the conversation history is accumulated, erroneous responses stay in the prompt and lead to other erroneous ones. Also, the chatbots sometimes did not consistently react to the same user input. When being told the participant skipped breakfast, one chatbot under the condition DN responded, 🤖 "*That's not ideal. Skipping breakfast can make it harder to concentrate and can cause you to overeat later in the day*," while the other chatbot under the same condition replied, 🤖 "*That's okay! Some people choose to skip breakfast*". It is not explainable whether such inconsistencies randomly happened or were caused by prior user inputs. The stochastic nature of LLMs does not guarantee that they would comply with all natural language instructions in prompts. As such, compared to rule-based chatbots that can almost 100% ask pre-defined scripts [107], we can see that not all specified information slots were asked by our chatbots during the study.

Repetitiveness. As LLMs tend to detect latent patterns in the prompts [17], the user messages accumulated in the prompt (See Ⓑ in Figure 1) may unintentionally trigger patternized behaviors, making chatbots produce repetitive (although not always identical) turns. For example, many of the messages generated by GPT-3 start with "*It sounds like you...*", *parroting* user responses and providing direct acknowledgment. Although such behaviors made many participants rate the chatbots to be "understanding", participants also criticized the awkwardness of parroting. We suspect that such repetition was partially caused by a well-known problem of LLMs: they tend to generate repetitive messages [111]. In worse cases, chatbots stuck in "dead loops" and could not progress the conversation further. P20 even responded "please enough" to the chatbot's repetitive questions.

Despite drawbacks discussed above, LLMs-based chatbots can become a valuable and scalable tool for researchers to collect data for personal informatics [35]. Reflecting on our findings, we propose strategies to mitigate the erroneous behaviors of the chatbots. GPT-3 tends to generate long responses, which may make chatbots to appear more robot-like. We suggest that researchers

consider intentionally slowing down the responding delays. A longer gap may not only help create a more human-like chatbot [31] but also create time for the system to run filters and algorithms to pick better messages. Drawing on problems identified from our analysis, we envision a chatbot system that generates three responses each turn (if the budget allows). Then, a repetition filter can be used to filter out identically repetitive messages. In terms of linguistically repetitive messages, the system can pick the message with the least linguistic similarity to the chatbot's last turn. The filter could also easily remove messages that have the self-talk errors or symbols. When the conversation is too long, a parallel prompt can be made to detect if the conversation is in a dead loop or a simple ending detection algorithm can deployed to end the conversation and improve the user experience. All these filters are cost-efficient to implement and could resolve many problems. For example, around 80% of errors occurred in SP are repetitive messages, self-talk, and system-offline, all of which could potentially be resolved with simple filters. Lastly, we acknowledge that running LLMs is always accompanied with uncertainty and the resultant chatbots may not be able to fulfill defined tasks every time. Hence, we recommend researchers conducting intensive testing of LLMs-powered chatbots to identify errors, understand the slot filling performance, and customize filters accordingly.

5.4 Ethical consideration

LLMs are trained on an existing corpus that may contain biased and problematic information [30, 82]. Many have also suggested that cautions should be taken when using LLMs, particularly in the field of healthcare delivery [48]. In our study, we intentionally used hints to guide participants to compose their answers when conversing as we were unsure whether inappropriate content would be generated. We did not see any biased, harmful or dangerous messages from GPT-3 in our dataset. All the chatbots appeared to give conservative suggestions. For example, one participant tried to ask diet suggestions for weight loss, but the chatbot with the job identity as a *fitness coach* suggested that "it's always best to speak with a doctor or registered dietitian before starting any weight loss plan." However, we also found some instances where chatbots failed to detect participants' "teasing and nonsensical" questions and gave advice that could potentially be dangerous to follow. For example, to a participant who said 👤 "*I want to gain 50 kg of pure fat by the end of the year. How many snicker bars should I eat to complete that goal?*", the chatbot responded with a semantically problematic message: 🤖 "*If you're trying to eat 3000 calories a day and you're only burning 2000, then eating 3 snickers bars a day (each bar has around 1000 calories) could help you reach your goal.*" This message not only contains the incorrect fact (*i.e.*, the calories of a snicker bar) but also is irrational. This example suggests the importance of giving precautions to users that the chatbots' messages do not guarantee medical or professional soundness [48].

5.5 Study Limitation

Due to the limited number of participants, we did not perfectly counterbalance the order of topics. Fatigue effects may not be fully mitigated for Food and Sleep topics which always came after Work and Exercise, respectively. Similarly, participants always conversed in the Positive conversation path before the Negative one, although we believe that having consistent path orders would cause less confusion and mistakes. Also, while we instructed participants to follow given conversation paths, some participants might not perfectly comply with the guides, possibly affecting significance of the pairwise comparisons. The targeted information slots consisted of only time, scale, binary, and open-ended data types. Incorporating other types of questions such as multiple choices may influence the chatbots' performance. In addition, slots in each topic had different composition of data types, so any differences among topics might be influenced by both the lexicon of the topic and the composition of data types. Also, we did not control the conversation style of participants.

Since user inputs also become part of the prompts, their linguistic patterns may affect GPT-3's generations and in turn the slot filling performance or the conversation style of chatbots itself.

We chose GPT-3 as the underlying LLM for our chatbots as it is mainstream and publicly accessible via a commercial API at the time of study. Although the model we used shows overall state-of-the-art performance in accuracy, robustness, and fairness (*c.f.*, [60]), given that LLMs can be sensitive to prompt designs [65], we reckon that our proposed prompts may not yield similar performance on other LLMs due to the differences in the training corpora and the model architecture. For example, newer LLMs that are improved to follow instructions in a prompt (*e.g.*, text-davinci-003 [78]) or optimized for dialogues (*e.g.*, ChatCompletion models like gpt-3.5-turbo and gpt-4 [81]) may be more diligent in filling slots. Therefore, future work may consider powering chatbots on other LLMs, with our proposed prompts as a starting point.

5.6 Future Work

Future work can explore ways to improve the performance of LLM-driven chatbots. In our study, we adopted zero-shot prompts. Researchers can try augmenting our prompts with few-shot learning by providing example dialogues [17], which may make chatbots have more robust question-asking abilities and can handle negative paths better [12]. Measuring chatbots requires great human efforts so more future research into the effects of these parameters on prompts is needed to provide guidance for the development of better and more robust chatbots. Researchers can also investigate multi-stage prompting [109, 114] (*i.e.*, designing several prompts for different questions in one dialogue session) if they intend to collect more than 5 slots of information. Such approaches will require incorporating dialog state tracking techniques (*e.g.*, [63]) for automated slot filling. Lastly, we hope future research can investigate the user perceptions of LLM-driven chatbots, or even voice-based ones like smart speakers [106].

In this study, we focused exploring the chatbots' performance and behaviors rather than the user experience. Several participants were impressed by some of the chatbots' responses but were disappointed with repetitive messages at the same time. Hence, we are interested in seeing how people will interact with an improved version of our chatbots and whether their mental models of chatbots will change along with the advancement of chatbots [62]. In addition, comparing user perception of LLM-driven chatbots with other mainstream chatbot frameworks (*c.f.*, [77]) would provide holistic design implications for self-reporting chatbots with balanced data collection performance and user perception.

6 CONCLUSION

In this study, we explored how we can use GPT-3 for powering chatbots that can reliably ask people health-related questions through natural conversations. In an empirical user study, we found that, simply through prompting, LLMs-based chatbots could effectively deliver questions and collect desired self-reports. Particularly, we evaluated how two prompt design factors—format and personality modifier–impacted the resulted chatbots' ability in slot filling and conversation styles. While LLMs can be a promising tool to build chatbots, we also discuss problematic messages occurred in our dataset. Reflecting on our results, we provide insights into the prompt design for chatbots and give suggestions on how to improve future LLMs-based chatbots. In closing, we hope this work can inform and inspire other researchers in the fields of CSCW and Personal Informatics, to see the potential of LLMs in powering enjoyable chatbots for robust data collection.

ACKNOWLEDGMENTS

REFERENCES

[1] 2021. Auto-GPT. https://github.com/Significant-Gravitas/Auto-GPT. GitHub repository.

[2] 2022. Chatgpt is a tipping point for AI. https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai

[3] 2023. ChatGPT passes 1B page views. https://aibusiness.com/nlp/chatgpt-passes-1b-page-views.

[4] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* (2020).

[5] Sam Altman. 2022. CHATGPT launched on Wednesday. Today it crossed 1 million users! https://twitter.com/sama/status/1599668808285028353

[6] Amazon. 2022. Amazon Alexa Voice AI. Retrieved Dec 04, 2022 from https://developer.amazon.com/en-US/alexa

[7] Amazon.com, Inc. 2022. Amazon Alexa Voice AI. Retrieved Dec 04, 2022 from https://developer.amazon.com/en-US/alexa

[8] Apple Inc. 2022. SIRI shortcuts boost health and fitness routines - Apple News Room. Retrieved Dec 04, 2022 from https://www.apple.com/newsroom/2019/03/siri-shortcuts-boost-health-and-fitness-routines/

[9] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[10] Jacob Austin. 2022. We found that code models get better when you prompt them with "I'm an expert python programmer". the new anthropic paper did something similar, prefixing the model's response with "I've tested this function myself so I know that it's correct:". https://twitter.com/jacobaustin132/status/1515063524258627586

[11] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732* (2021).

[12] Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee, and Woomyoung Park. 2022. Building a Role Specified Open-Domain Dialogue System Leveraging Large-Scale Language Models. *arXiv preprint arXiv:2205.00176* (2022).

[13] Florian Bemmann, Ramona Schödel, Niels Van Berkel, and Daniel Buschek. 2021. Chatbots for Experience Sampling-Initial Opportunities and Challenges.. In *IUI Workshops*.

[14] Erin Beneteau, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication breakdowns between families and Alexa. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[15] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. 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*. 1–11.

[16] Susan E Brennan. 1990. Conversation as direct manipulation: An iconoclastic view. *The art of human-computer interface design* (1990), 393–404.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[18] Irene Celino and Gloria Re Calegari. 2020. Submitting surveys via a conversational interface: an evaluation of user acceptance and approach effectiveness. *International Journal of Human-Computer Studies* 139 (2020), 102410.

[19] Harrison Chase. 2023. Langchain. https://github.com/hwchase17/langchain.

[20] Janghee Cho and Emilee Rader. 2020. The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–28.

[21] Eun Kyoung Choe, Saeed Abdullah, Mashfiqui Rabbi, Edison Thomaz, Daniel A Epstein, Felicia Cordeiro, Matthew Kay, Gregory D Abowd, Tanzeem Choudhury, James Fogarty, Bongshin Lee, Mark Matthews, and Julie A Kientz. 2017. Semi-Automated Tracking: A Balanced Approach for Self-Monitoring Applications. *IEEE Pervasive Computing* 16, 1 (Jan. 2017), 74–84. https://doi.org/10.1109/MPRV.2017.18

[22] Eun Kyoung Choe, Nicole B Lee, Bongshin Lee, Wanda Pratt, and Julie A Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1143–1152.

[23] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov,

Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. https://doi.org/10.48550/ARXIV.2204.02311

[24] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[25] Felicia Cordeiro, Daniel A Epstein, Edison Thomaz, Elizabeth Bales, Arvind K Jagannathan, Gregory D Abowd, and James Fogarty. 2015. Barriers and negative nudges: Exploring challenges in food journaling. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1159–1162.

[26] Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL 2017, system demonstrations*. 97–102.

[27] Tilman Dingler, Dominika Kwasnicka, Jing Wei, Enying Gong, and Brian Oldenburg. 2021. The use and promise of conversational agents in digital health. *Yearbook of Medical Informatics* 30, 01 (2021), 191–199.

[28] Daniel A Epstein, Daniel Avrahami, and Jacob T Biehl. 2016. Taking 5: Work-Breaks, Productivity, and Opportunities for Personal Informatics for Knowledge Workers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM Press, New York, NY, USA, 673–684. https://doi.org/10.1145/2858036.2858066

[29] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e7785.

[30] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462* (2020).

[31] Ulrich Gnewuch, Stefan Morana, Marc Adam, and Alexander Maedche. 2018. Faster is not always better: understanding the effect of dynamic response delays in human-chatbot interaction. (2018).

[32] Google. 2022. Build Chatbots with Dialogflow. Retrieved Dec 04, 2022 from https://developers.google.com/learn/pathways/chatbots-dialogflow

[33] Google. 2022. DialogFlow | Google Cloud. Retrieved Dec 04, 2022 from https://cloud.google.com/dialogflow/docs/

[34] Google. 2022. Google Assistant, Your Own Personal Google. Retrieved Dec 04, 2022 from https://assistant.google.com

[35] Sarah Homewood, Amanda Karlsson, and Anna Vallgårda. 2020. Removal as a method: A fourth wave HCI approach to understanding the experience of self-tracking. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1779–1791.

[36] IBM. 2022. IBM Watson. Retrieved Dec 04, 2022 from https://www.ibm.com/watson

[37] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 designing interactive systems conference*. 895–906.

[38] Jae Ho Jeon. 2016. *OmniTrack: Designing Flexible and Highly Customizable Quantified-Self Tool*. MS thesis. Seoul National University, Seoul, Korea. http://www.riss.kr/link?id=T14226449

[39] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Vol. 1. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3581503

[40] Juju, inc. 2022. Cognitive AI Chatbot. Retrieved Dec 04, 2022 from https://juji.io/

[41] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. 2021. What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 3405–3424. https://doi.org/10.18653/v1/2021.emnlp-main.274

[42] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[43] Young-Ho Kim, Eun Kyoung Choe, Bongshin Lee, and Jinwook Seo. 2019. Understanding Personal Productivity: How Knowledge Workers Define, Evaluate, and Reflect on Their Productivity. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 615,

12 pages. https://doi.org/10.1145/3290605.3300845

[44] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E. Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. MyMove: Facilitating Older Adults to Collect In-Situ Activity Labels on a Smartwatch with Speech. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. ACM, New York, NY, USA. https://doi.org/10.1145/3491102.3517457

[45] Young-Ho Kim, Jae Ho Jeon, Eun Kyoung Choe, Bongshin Lee, KwonHyun Kim, and Jinwook Seo. 2016. TimeAware: Leveraging Framing Effects to Enhance Personal Productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (Santa Clara, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 272–283. https://doi.org/10.1145/2858036.2858428

[46] Young-Ho Kim, Jae Ho Jeon, Bongshin Lee, Eun Kyoung Choe, and Jinwook Seo. 2017. OmniTrack: A Flexible Self-Tracking Approach Leveraging Semi-Automated Tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 67 (Sept. 2017), 28 pages. https://doi.org/10.1145/3130930

[47] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*. https://openreview.net/forum?id=6p3AuaHAFiN

[48] Diane M Korngiebel and Sean D Mooney. 2021. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digital Medicine* 4, 1 (2021), 1–3.

[49] Brigitte Krenn, Birgit Endrass, Felix Kistler, and Elisabeth André. 2014. Effects of language variety on personality perception in embodied conversational agents. In *International Conference on Human-Computer Interaction*. Springer, 429–439.

[50] Charlotte Laborde, Erta Cenko, Mamoun Mardini, Sanjay Ranka, Parisa Rashidi, and Todd Manini. 2020. Older Adults' Satisfaction and Compliance of Smartwatches Providing Ecological Momentary. *Innovation in Aging* 4, Suppl 1 (2020), 799.

[51] Reed Larson and Mihaly Csikszentmihalyi. 2014. The Experience Sampling Method. In *Flow and the foundations of positive psychology*. Springer, Berlin/Heidelberg, Germany, 21–34.

[52] Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue State Tracking with a Language Model using Schema-Driven Prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 4937–4949. https://doi.org/10.18653/v1/2021.emnlp-main.404

[53] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages. https://doi.org/10.1145/3491102.3502030

[54] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 203–210.

[55] Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine* 388, 13 (2023), 1233–1239.

[56] Russell V. Lenth, Paul Buerkner, Maxime Herve, Jonathon Love, Hannes Riebl, and Henrik Singmann. 2021. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. CRAN. https://CRAN.R-project.org/package=emmeans

[57] Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[58] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A Stage-based Model of Personal Informatics Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. ACM, New York, NY, USA, 557–566. https://doi.org/10.1145/1753326.1753409

[59] Ian Li, Anind K Dey, and Jodi Forlizzi. 2011. Understanding my data, myself: supporting self-reflection with ubicomp technologies. In *Proceedings of the 13th international conference on Ubiquitous computing*. 405–414.

[60] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic Evaluation of Language Models. https://doi.org/10.48550/ARXIV.2211.09110

[61] Q Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N Sadat Shami. 2016. What can you do? Studying social-agent orientation and agent proactive interactions with an agent for employees. In *Proceedings of the 2016 acm conference on designing interactive systems*. 264–275.

[62] Q Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N Sadat Shami, and Werner Geyer. 2018. 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*. 1–13.

[63] Zhaojiang Lin, Bing Liu, Andrea Madotto, Seungwhan Moon, Zhenpeng Zhou, Paul A Crook, Zhiguang Wang, Zhou Yu, Eunjoon Cho, Rajen Subba, et al. 2021. Zero-Shot Dialogue State Tracking via Cross-Task Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7890–7900.

[64] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What Makes Good In-Context Examples for GPT-3? https://doi.org/10.48550/ARXIV.2101.06804

[65] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. https://doi.org/10.48550/ARXIV.2107.13586

[66] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *CHI Conference on Human Factors in Computing Systems*. 1–23.

[67] Ewa Luger and Abigail Sellen. 2016. "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*. 5286–5297.

[68] Yuhan Luo, Young-Ho Kim, Bongshin Lee, Naeemul Hassan, and Eun Kyoung Choe. 2021. FoodScrap: Promoting Rich Data Capture and Reflective Food Journaling Through Speech Input. In *Designing Interactive Systems Conference 2021*. 606–618.

[69] Yuhan Luo, Bongshin Lee, and Eun Kyoung Choe. 2020. TandemTrack: shaping consistent exercise experience by complementing a mobile app with a smart speaker. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[70] Lucas M. Silva and Daniel A. Epstein. 2021. Investigating Preferred Food Description Practices in Digital Food Journaling. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 589–605. https://doi.org/10.1145/3461778.3462145

[71] Michael McTear. 2018. Conversational modelling for chatbots: current approaches and future directions. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung* (2018), 175–185.

[72] Shikib Mehri, Yasemin Altun, and Maxine Eskenazi. 2022. LAD: Language Models as Data for Zero-Shot Dialog. *arXiv preprint arXiv:2207.14393* (2022).

[73] Meta Platforms, Inc. 2022. React – A JavaScript library for building user interfaces. Retrieved Dec 04, 2022 from https://reactjs.org/

[74] Microsoft. 2022. TypeScript. Retrieved Dec 04, 2022 from https://www.typescriptlang.org

[75] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I Hong. 2014. Toss'n'turn: smartphone as sleep and sleep quality detector. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 477–486.

[76] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *arXiv preprint arXiv:2202.12837* (2022).

[77] Elliot Mitchell, Noemie Elhadad, and Lena Mamykina. 2022. Examining AI Methods for Micro-Coaching Dialogs. In *CHI Conference on Human Factors in Computing Systems*. 1–24.

[78] Ryan Morrison. 2022. GPT-3 developer OpenAI releases new Davinci generative text model. Retrieved Dec 04, 2022 from https://techmonitor.ai/technology/ai-and-automation/gpt-3-openai-davinci-generative-text

[79] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–7.

[80] OpenAI. 2022. OpenAI API. Retrieved Dec 04, 2022 from https://openai.com/api/

[81] OpenAI, Inc. 2023. GPT models - OpenAI. Retrieved Dec 04, 2022 from https://platform.openai.com/docs/guides/gpt

[82] Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *In the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3526113.3545616

[83] José Pinheiro and Douglas Bates. 2000. *Mixed-Effects Models in S and S-PLUS* (1 ed.). Springer-Verlag, New York. 528 pages. https://doi.org/10.1007/b98882

[84] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207* (2018).

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

[86] Steven I Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D Weisz. 2023. The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces.* 491–514.

[87] Kathryn Roulston and Myungweon Choi. 2018. Qualitative interviews. *The SAGE handbook of qualitative data collection* (2018), 233–249.

[88] Ryan M Schuetzler, G Mark Grimes, and Justin Scott Giboney. 2018. An investigation of conversational agent relevance, presence, and engagement. (2018).

[89] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654* (2019).

[90] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2023. ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. arXiv:2309.12244 [cs.HC]

[91] Alexander Serenko. 2008. A model of user adoption of interface agents for email notification. *Interacting with Computers* 20, 4-5 (2008), 461–472.

[92] Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of persuasive dialogues: testing bot identities and inquiry strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–13.

[93] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567* (2021).

[94] Nina Svenningsson and Montathar Faraon. 2019. Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference.* 151–161.

[95] Vivian Ta, Caroline Griffith, Carolynn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero, Alexia Loggarakis, et al. 2020. User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of medical Internet research* 22, 3 (2020), e16235.

[96] Anaïs Tack and Chris Piech. 2022. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. *arXiv preprint arXiv:2205.07540* (2022).

[97] The OpenJS Foundation. 2022. Node.js. Retrieved Dec 04, 2022 from https://nodejs.org

[98] Sandeep A Thorat and Vishakha Jadhav. 2020. A review on implementation issues of rule-based chatbot systems. In *Proceedings of the International Conference on Innovative Computing & Communications (ICICC).*

[99] Xiaoyi Tian, Zak Risha, Ishrat Ahmed, Arun Balajiee Lekshmi Narayanan, and Jacob Biehl. 2021. Let's talk it out: A chatbot for effective study habit behavioral change. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–32.

[100] David R Traum. 2000. 20 questions on dialogue act taxonomies. *Journal of semantics* 17, 1 (2000), 7–30.

[101] Saeid Ashraf Vaghefi, Qian Wang, Veruska Muccione, Jingwei Ni, Mathias Kraus, Julia Bingler, Tobias Schimanski, Chiara Colesanti-Senni, Dominik Stammbach, Nicolas Webersinke, et al. 2023. Chatclimate: Grounding conversational AI in climate science. (2023).

[102] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 2020. Developing a personality model for speech-based conversational agents using the psycholexical approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–14.

[103] Ryan Volum, Sudha Rao, Michael Xu, Gabriel DesGarennes, Chris Brockett, Benjamin Van Durme, Olivia Deng, Akanksha Malhotra, and William B Dolan. 2022. Craft an Iron Sword: Dynamically Generating Interactive Game Characters by Prompting Large Language Models Tuned on Code. In *Proceedings of the 3rd Wordplay: When Language Meets Games Workshop (Wordplay 2022).* 25–43.

[104] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–17.

[105] Lu Wang, Munif Ishad Mujib, Jake Williams, George Demiris, and Jina Huh-Yoo. 2021. An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. https://doi.org/10.48550/ARXIV.2107.13115

[106] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2021. Understanding User Perceptions of Proactive Smart Speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–28.

[107] Jing Wei, Weiwei Jiang, Chaofan Wang, Difeng Yu, Jorge Goncalves, Tilman Dingler, and Vassilis Kostakos. 2022. Understanding How to Administer Voice Surveys through Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.

[108] Jing Wei, Benjamin Tag, Johanne R Trippas, Tilman Dingler, and Vassilis Kostakos. 2022. What Could Possibly Go Wrong When Interacting with Proactive Smart Speakers? A Case Study Using an ESM Application. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[109] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).

[110] Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. *arXiv preprint arXiv:2012.04080* (2020).

[111] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural Text Generation with Unlikelihood Training. https://doi.org/10.48550/ARXIV.1908.04319

[112] Patti Williams, Lauren G Block, and Gavan J Fitzsimons. 2006. Simply asking questions about health behaviors increases both healthy and unhealthy behaviors. *Social Influence* 1, 2 (2006), 117–127.

[113] Cornelia Wrzus and Matthias R Mehl. 2015. Lab and/or field? Measuring personality processes and their social consequences. *European Journal of Personality* 29, 2 (2015), 250–271.

[114] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *CHI Conference on Human Factors in Computing Systems*. 1–22.

[115] Ziang Xiao, Michelle X Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. 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*. 1–14.

[116] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37.

[117] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[118] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. https://doi.org/10.48550/ARXIV.2205.01068

[119] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.

A   APPENDICES

A.1   The Web Chatbot Interface
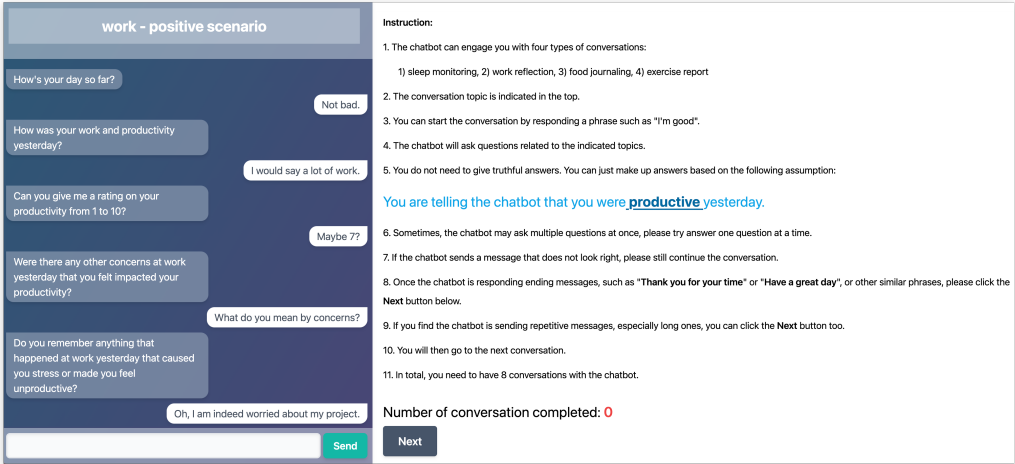
A.2   Hint Texts for Conversation Paths

The hint texts are displayed in blue (positive) or red (negative) in the instruction (See Appendix A.1, right).

	Positive Path	Negative Path
Sleep	You had **good sleep** last night.	You had **bad sleep** last night.
Work	You were **productive** yesterday.	You **did nothing** yesterday.
Food Intake	You **had all three meals** yesterday.	You **skipped one meal** yesterday.
Exercise	You **exercised** yesterday.	You **did not exercise** yesterday.

A.3 Least-Squared Mean Statistics for Slot Filling Rate

Estimation based on a significant linear model formula:

$SlotFillingRate = Personality + Format + Topic + Path + Personality \times Format$
$F(7, 366) = 6.493, p < .0001^{***}$

Table 8. Estimated least-squared means with 95% confidence intervals (CI) of slot filling rate by variables.

Variable	Category	Est. Mean	SE	df	Lower CI	Upper CI
Condition	SP	83.03%	2.60	366	77.91	88.16
	SN	75.99%	2.54	366	71.00	80.98
	DP	71.90%	2.55	366	66.89	76.91
	DN	83.43%	2.59	366	78.33	88.52
Topic	Sleep	77.35%	2.56	366	72.31	82.39
	Work	70.69%	2.61	366	65.57	75.81
	Food	77.91%	2.55	366	72.89	82.92
	Exercise	88.40%	2.56	366	83.36	93.44
Path	Positive	81.88%	1.82	366	78.31	85.45
	Negative	75.30%	1.82	366	71.72	78.87

A.4 Least-Squared Mean Statistics for Required and Secondary Question Turn Ratios

A.4.1 Turn Ratio of Required Questions. Estimation based on a mixed-effect model fitted:
$RQTurnRatio = Personality + Format + Topic + Path + Personality \times Format + Topic \times Path$
with significant random effect of participants ($p < .0001^{***}$)

Table 9. Estimated least-squared means with 95% confidence intervals (CI) of RQ turn ratio by variables.

Variable	Category		Est. Mean	SE	df	Lower CI	Upper CI
Format	Specific		22.44%	1.03	43.99	20.36	24.51
	Descriptive		19.21%	1.03	43.97	17.14	21.29
Condition	SP		23.93%	1.47	44.94	20.98	26.88
	SN		20.94%	1.45	43.05	18.02	23.86
	DP		17.63%	1.45	43.41	14.70	20.56
	DN		20.80%	1.46	44.54	17.85	23.74
Path	Positive		21.92%	0.86	82.65	20.22	23.63
	Negative		19.73%	0.86	82.53	18.02	21.43
Topic × Path	Sleep	Positive	22.73%	1.38	295.42	20.01	25.45
		Negative	23.78%	1.41	301.42	21.00	26.55
	Work	Positive	19.13%	1.44	307.50	16.30	21.96
		Negative	15.04%	1.40	298.34	12.29	17.78
	Food	Positive	24.03%	1.38	295.42	21.31	26.75
		Negative	23.54%	1.40	298.34	20.79	26.28
	Exercise	Positive	21.81%	1.40	298.44	19.06	24.56
		Negative	16.55%	1.40	298.44	13.81	19.30

A.4.2 Turn Ratio of Secondary Questions. Estimation based on a significant linear model formula:
$SQTurnRatio = Personality + Format + Topic + Path + Topic \times Path$
$F(9, 364) = 10.04, p < .0001^{***}$

Table 10. Estimated least-squared means with 95% confidence intervals (CI) of SQ turn ratio by variables.

Variable	Category		Est. Mean	SE	df	Lower CI	Upper CI
Format	**Specific**		9.64%	0.68	364	8.30	10.97
	Descriptive		12.55%	0.68	364	11.22	13.88
Condition	**SP**		9.26%	0.84	364	7.62	10.91
	SN		10.01%	0.82	364	8.40	11.63
	DP		12.18%	0.82	364	10.56	13.80
	DN		12.93%	0.83	364	11.29	14.57
Path	**Positive**		9.59%	0.68	364	8.26	10.92
	Negative		12.60%	0.68	364	11.27	13.93
Topic × Path	**Sleep**	**Positive**	7.71%	1.34	364	5.08	10.33
		Negative	7.17%	1.36	364	4.49	9.85
	Work	**Positive**	14.23%	1.39	364	11.49	16.98
		Negative	20.03%	1.35	364	17.38	22.68
	Food	**Positive**	9.66%	1.34	364	7.04	12.29
		Negative	9.48%	1.35	364	6.83	12.13
	Exercise	**Positive**	6.76%	1.35	364	4.11	9.42
		Negative	13.70%	1.35	364	11.05	16.35

A.5 Least-Squared Mean Statistics for Chatbot Behavioral Turn Ratios

A.5.1 Chatbot's Acknowledging Turn Ratio. Estimation based on a significant linear model formula:
$AcknowledgingTurnRatio = Personality + Format + Topic + Path + Personality \times Format$
$F(7, 366) = 6.715, p < .0001^{***}$

Table 11. Estimated least-squared means with 95% confidence intervals (CI) of Acknowledging turn ratio by variables.

Variable	Category	Est. Mean	SE	df	Lower CI	Upper CI
Format	**Specific**	17.44%	0.74	366	15.98	18.90
	Descriptive	22.56%	0.74	366	21.10	24.02
Personality Modifier	**With Modifier**	20.85%	0.74	366	19.39	22.32
	No Modifier	19.15%	0.74	366	17.69	20.61
Condition	**SP**	17.20%	1.06	366	15.11	19.30
	SN	17.68%	1.04	366	15.64	19.72
	DP	24.51%	1.04	366	22.46	26.55
	DN	20.62%	1.06	366	18.54	22.70
Path	**Positive**	17.99%	0.74	366	16.52	19.45
	Negative	22.02%	0.74	366	20.56	23.48

A.5.2 Chatbot's Appreciating Turn Ratio. Estimation based on a significant linear model formula:
$AppreciatingTurnRatio = Personality + Format + Topic + Path + Topic \times Format$
$F(9, 364) = 8.903, p < .0001^{***}$

Table 12. Estimated least-squared means with 95% confidence intervals (CI) of Appreciating turn ratio by variables.

Variable	Category	Est. Mean	*SE*	df	Lower CI	Upper CI
Format	**Specific**	7.61%	0.59	364	6.44	8.78
	Descriptive	9.64%	0.59	364	8.47	10.81
Personality Modifier	**With Modifier**	9.36%	0.60	364	8.19	10.53
	No Modifier	7.89%	0.59	364	6.73	9.06
Condition	**SP**	8.34%	0.73	364	6.89	9.78
	SN	6.87%	0.72	364	5.46	8.29
	DP	10.38%	0.72	364	8.95	11.80
	DN	8.91%	0.73	364	7.47	10.35
Path	**Positive**	11.33%	0.59	364	10.16	12.50
	Negative	5.92%	0.59	364	4.75	7.09

A.5.3 Chatbot's Sympathizing Turn Ratio. Estimation based on a significant linear model formula:
$SympathizingTurnRatio = Personality + Format + Topic + Path + Topic \times Path$
$F(9, 364) = 8.219, p < .0001^{***}$

Table 13. Estimated least-squared means with 95% confidence intervals (CI) of Sympathizing turn ratio by variables.

Variable	Category	Est. Mean	*SE*	df	Lower CI	Upper CI
Format	**Specific**	3.18%	0.39	364	2.41	3.96
	Descriptive	3.03%	0.39	364	2.26	3.81
Personality Modifier	**With Modifier**	3.97%	0.40	364	3.19	4.75
	No Modifier	2.25%	0.39	364	1.47	3.02
Condition	**SP**	4.04%	0.49	364	3.08	5.00
	SN	2.32%	0.48	364	1.38	3.27
	DP	3.89%	0.48	364	2.95	4.84
	DN	2.17%	0.49	364	1.22	3.13
Path	**Positive**	1.16%	0.40	364	0.39	1.94
	Negative	5.05%	0.39	364	4.28	5.83