Zero-shot Persuasive Chatbots with LLM-Generated Strategies and Information Retrieval

Anonymous ACL submission

Abstract

Persuasion plays a pivotal role in a wide range of applications, from health intervention, to the promotion of social good. Persuasive chatbots can accelerate the positive effects of persuasion in such applications. Existing methods rely on fine-tuning persuasive chatbots with task-specific training data which is costly, if not infeasible, to collect.

To address this issue, we propose a method to leverage the generalizability and inherent persuasive abilities of large language models (LLMs) in creating effective and truthful persuasive chatbot for any given domain in a zero-shot manner. Unlike previous studies which used pre-defined persuasion strategies, our method first uses an LLM to generate responses, then extracts the strategies used on the fly, and replaces any unsubstantiated claims in the response with retrieved facts supporting the strategies.

We applied our chatbot, PersuaBot, to three significantly different domains needing persuasion skills: donation solicitation, recommendations, and health intervention. Our experiments on simulated and human conversations show that our zero-shot approach is more persuasive than prior work, while achieving factual accuracy surpassing state-of-the-art knowledge-oriented chatbots. Our study demonstrated that when persuasive chatbots are employed responsibly for social good, it is an enabler of positive individual and social change.

1 Introduction

Persuasive technologies are designed to reinforce, change or mold user's attitude or behavior (Fogg, 2003; Oinas-Kukkonen and Harjumaa, 2009). Numerous investigations have shown that these technologies may be used effectively for a range of purposes, such as social good promotion (Wang et al., 2019a), health intervention (Athota et al., 2020), marketing (Sun and Zhang, 2018; Chen et al., 2021), and so on (Kaptein et al., 2015; Oyebode et al., 2022; Adaji and Adisa, 2022; Baumeister et al., 2022; Halttu and Oinas-Kukkonen, 2022). Yet, automated persuasive chatbots still struggle to achieve their goal, which is to "deliver the right message, at the right time, and in the right way" (Kaptein et al., 2015), especially when there is no training data for the persuasion task. This paper proposes PersuaBot, a zero-shot persuasive chatbot, which is: (1) factual, (2) adaptive to the user, and (3) adaptive to the domain.

1.1 Background

Previous persuasive chatbots are typically trained with annotated conversations, wherein chatbots rely on dialogue acts to follow predefined agendas or optimize the dialogue policies (Yu et al., 2019; Shi et al., 2020a,b; Wang et al., 2023; Dinan et al., 2019). Prior works have explored user-adaptive chatbots by modeling user personality either directly from predefined user personas or attributes (Liu et al., 2020; Zhang et al., 2018), which are hard to collect. Chen et al. (2022, 2023) proposed a zero-shot LLM-based modular approach where a predefined strategy for a given domain is supplied to LLMs, which is found to be more persuasive than previous user-adaptive approaches.

LLM-based approaches have a downside as they tend to hallucinate (Bang et al., 2023; Cao et al., 2017). For instance, we observe that an LLM-based chatbot attempting to persuade the user to donate to a charity would invent a girl named "Maria" who had been suffering from poverty but was saved by the charity. When such falsehoods are found out, they are catastrophic in persuasion, as they erode the user's trust in the system.

To combat hallucination, recent work has combined information retrieval with LLMs (Lewis et al., 2021; Trivedi et al., 2022; Gao et al., 2023; Thorne et al., 2018; Semnani et al., 2023). These methods would have removed the fabricated story

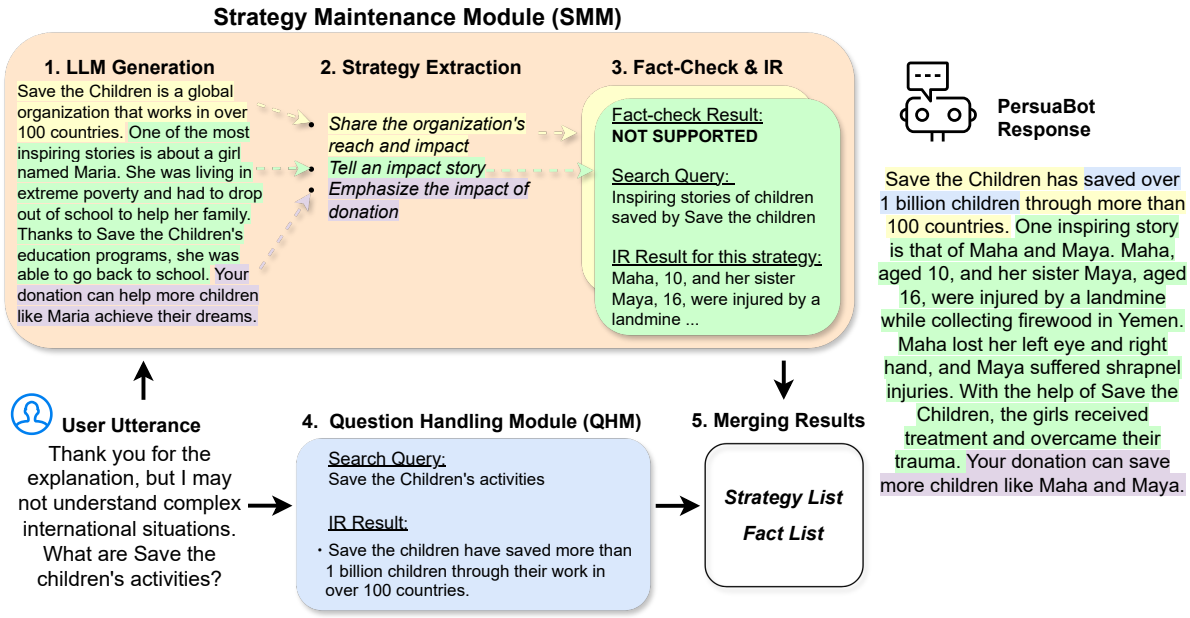

Figure 1: An overview of PersuaBot and an example of a conversation to persuade a user to donate to "Save the Children". To generate a factually correct response and maintain its persuasive function, PersuaBot has a Strategy Maintenance Module that (1) generates an LLM response and decomposes it according to its strategy intent, (2) extracts the strategies for each section, (3) fact-checks the sections and uses IR to substantiate the strategies. Additionally, (4) it retrieves relevant information in response to users' requests if any, and (5) merges the results.

about "Maria" in order to increase factuality, but in doing so, would have negatively impacted the persuasiveness of the response.

1.2 Our Approach

Recent advancements in powerful LLMs enable them to discern users' emotions (e.g., dis/agreement, reservation, or hesitation), formulate an appropriate strategy to persuade the user or offer information to address the user's concerns, and provide evidence to substantiate the intent.

This research asks how we can fully embrace the persuasiveness of LLMs while minimizing hallucination? Our key concept is to distill out the persuasion strategies (or the intent of the response) from the actual substantiating evidence. We keep the strategies, but replace the LLM-generated information with actual facts obtained via retrieval, if fact-checking fails. Concretely, we create PersuaBot, a multi-step pipeline of LLM calls with in-context learning, as shown in Figure 1. The pipeline merges the results of two modules: a basic Question Handling Module (QHM) that retrieves according to the user's request, and a Strategy Maintenance Module (SMM), which consists of the following steps:

1. Use LLM to generate a response.
2. Decompose the LLM response into sections, each carrying out a distinct strategy.
3. Fact-check each section. For any unsubstantiated section, we prompt the LLM to ask what query it is trying to answer, then we use the query to retrieve the answer from the corpus. The facts retrieved in this way are used to maintain the strategy for that section.

1.3 Contributions

We have demonstrated for the first time that it is possible to create a zero-shot LLM-based chatbot with high persuasiveness and factuality. We conducted experiments in three different domains (donation solicitation, recommendation systems, and healthcare intervention). To stress test PersuaBot, we simulated how PersuaBot would respond to *soft* users, those amenable to persuasion, and *tough* users, those who are not, such as unfriendly or skeptical users. Additionally, we conducted real user experiments in the donation solicitation domain to evaluate how humans interacted with persuasion chatbots and how persuasive they found those chatbots to be. The experimental results are as follows:

1. PersuaBot surpasses state-of-the-art knowledge-oriented chatbots in factuality, achieving up to 26.6% higher factuality

than GPT-3.5.

2. PersuaBot has greater persuasiveness than existing methods, as it enjoys an advantage of 0.6 over the state-of-the-art in persuasiveness on a 5-point scale than the manually designed rule-oriented method in real users experiment.
3. The same pipeline has been demonstrated to work across three different domains, demonstrating that it is domain-adaptive.

This study aims to demonstrate that persuasive chatbots, when applied responsibly on socially good purposes, can be a force for positive change. To achieve this, our approach focuses on reducing misinformation by grounding in accurate information. We monitor all utterances of the chatbots in our experiments to ensure they do not contain harmful content.

2 Related Work

2.1 Factual Chatbots

In recent years, many chatbots have leveraged the significant advances in state-of-the-art LLMs (OpenAI, 2023; Ouyang et al., 2022; Touvron et al., 2023) to enhance their performance. However, while LLMs have the capacity to accumulate vast amounts of knowledge (Petroni et al., 2019; Roberts et al., 2020), they are also known to provide misleading or fabricated information, i.e., they hallucinate (Nie et al., 2019; Dziri et al., 2021; Rashkin et al., 2021).

To mitigate hallucination, knowledge-grounded approaches have been proven effective, such as using knowledge graphs (Athreya et al., 2018; Ait-Mlouk and Jiang, 2020; Dayal et al., 2023) or linked data (Ait-Mlouk and Jiang, 2020). Using free-text corpora as external resources is also effective, for instance, Peng et al. (2023) proposed the LLM-AUGMENTER system, which enhances a LLM by integrating evidence from external knowledge to generate factual responses.

2.2 Strategizing for persuasive chatbot

The persuasive task involves achieving some form of agreement with the user, and task-specific datasets are provided to build persuasive chatbots (Lewis et al., 2017; He et al., 2018; Li et al., 2018; Zeng et al., 2020). In the context of persuasive tasks for social good, there are previous studies that persuade users to make charity donations (Wang et al., 2019b; Shi et al., 2020b). By integrating social content and factual information, Chen et al. (2022, 2023) proposed a Response-Agenda Pushing Framework (RAP) for strategically designed persuasive chatbots. Hunter et al. (2019) reviewed a framework that optimizes persuasion measures from argument and counterargument, and considers the persuadees' beliefs and concerns. Shi et al. (2020b) proposed a chatbot based on a dialogue manager that follows an agenda for the persuasion task and uses template-based responses. Flores et al. (2022) implemented a chatbot using the Watson (IBM, 2023) specifying the entities, intentions, and conversations. They used persuasive technology to change habits on recycling and solid waste management.

2.3 Discussion

Despite advances in persuasive chatbots, there are some gaps in their generalization. Most of the previous approaches tackle persuasion on traditional chatbots constructed with predefined dialogue states, acts, and scenarios. In these chatbots, strategies for persuasion are designed statically and manually. However, the implementation of dynamic strategies is an important task that has a significant impact on the generalization of persuasive chatbots.

To address this problem, our proposed method uses LLMs to retrieve and apply persuasion strategies without task-specific conversation data and pre-defined strategies. Additionally, for factual accuracy, we utilize task-related corpora such as organization websites to identify and replace hallucinations with content that is relevant to the response, which resulted in a beneficial impact on LLM's native persuasion ability.

3 Method: PersuaBot

As shown in Figure 1, our proposed method (PersuaBot) generates a response to the user utterance by running two modules in parallel: Strategy Maintenance Module (SMM) for factual correctness while maintaining strategies and Question Handling Module (QHM) based on user requests.

3.1 Strategy Maintenance Module

In this module, the LLM first generates a response (original response) directly based on the given task instructions. We then post-process them to ensure factuality while maintaining the strategies it encompasses.

3.1.1 Strategy Extraction

We use an LLM to extract the strategies from a response, given the conversation history and task description. Figure 5 shows the prompt used to decompose the response into strategies and sections in JSON format. During strategy extraction, we align the sections corresponding to each strategy in the original response and use them for subsequent fact-check steps. In Figure 1, for example, the strategy with intent "Emphasize the impact of donation" aligns with the section "Your donation can help more children like Maria achieve their dreams."

3.1.2 Fact-Check & IR

We first decompose the response into sections, *with each corresponding to a strategy intent*, used in the formulation of the response. We then decompose each section into claims, utilizing IR with ColBERT (Santhanam et al., 2022) on a corpus composed of free text to obtain evidence for each claim, and label the claims using a chain-of-thought prompt (Wei et al., 2023; Semnani et al., 2023). We conducted fact-checking for each strategy and its corresponding section; the structure of having potentially multiple strategies in the original response is retained. Note that fact-checking is not conducted for sections that do not contain claims, i.e. non-factual statements.

Whereas traditional approaches to reducing hallucination would filter out any claims deemed unsubstantiated, we attempt to retrieve information from the corpus to replace sections with unsubstantiated claims with factual evidence. To achieve this, we generate search queries using the prompt shown in Figure 6. It prompts the LLM to return a self-contained query for searching the corpus based on the strategy intent and the facts the LLM attempted to use for that strategy. In Figure 1, since the story of a girl named Maria was labeled as factually incorrect, given that the search intent is "tell an impact story of an individual" with the name of the organization being "Save the Children", the LLM generates a query like "An inspiring story about how Save the Children saved a child". Had we simply generated a query such as "Story of a girl named Maria," without referring to the strategy intent, IR may not be able to retrieve the relevant information. Using the strategy intent helps refine the scope of the facts to be retrieved.

In a sense, we are using the LLM to generate the query that the user may have asked, if they were interested in learning more information that may help convince them. The handing of the query is the same as handling a user-initiated query, which is described next. If there is no answer to the query, the strategy is removed from the response.

3.2 Question Handling Module

In this module, we perform IR on the latest user utterance and the conversation history. When the user asks a question, our method retrieves the necessary facts for the response. If the user does not ask questions, such as greetings, responding to remarks, or expressing opinions, this step is skipped.

3.3 Merging Results

For the final response generation, the facts and strategies obtained in each module are integrated. To achieve this, we use the prompt shown in Figure 7. Facts selected from the corpus are extracted in a bulleted format, specifying the information needed for strategies and user requests. These extracted facts, along with the reasons for obtaining them (accompanied by the intent of the relevant strategy), are then provided as instructions to the LLM. This ensures the generation of factually accurate responses while maintaining strategic functions. If the user asks a question but does not get any facts about it, it is labeled "Not enough information" and PersuaBot responds with something like "I'm not sure."

In the example in Figure 1, for the strategy "Share the organization's reach and impact," the claim included in the original response is supported, and additional facts obtained through QHM are incorporated, resulting in the final section: "Save the Children has saved over 1 billion children through more than 100 years of its activities." For the strategy "Tell an impact story," since the original claim is not supported by the corpus, other obtained facts are used. Additionally, for "Emphasize the impact of donation," although fact-checking is not performed, an adjusted section is generated, taking into account the use of the story of Maha and Maya instead of Maria's story. Note that the whole response is rewritten instead of just substituting the unsubstantiated claims with retrieved facts to ensure consistency, and prevents the so-called intrinsic hallucination (Li et al., 2020; Huang et al., 2023).

4 Experimental Setup

We compare our system PersuaBot with a state-of-the-art knowledge-oriented chatbot (Semnani et al.,

2023), and a manually designed persuasion chatbot for donations to Save the Children (Chen et al., 2022). All experiments are conducted in English. We use GPT-3.5-turbo-instruct as the base LLM.

4.1 Persuasion Tasks in Different Domains

We performed experiments in three persuasive tasks in different domains: social good, recommendation, and health intervention. For all three tasks, we used the same prompt and few-shot examples, excluding the task instructions, without fine-tuning on any task-specific conversation data. The few-shot examples are created manually for the social good task with relatively little human effort. The prompts are shown in Appendix A. The details of each task assigned to the chatbot are as follows:

- **Social Good**: Chatbots persuade users to donate to Save the Children, a non-governmental organization dedicated to international assistance for children.
- **Recommendation**: Chatbots persuade users to travel to Japan while recommending travel destinations within Japan.
- **Health Intervention**: Chatbots provide users with knowledge about COVID-19, flu, and respiratory syncytial virus (RSV), persuading them to enhance their health awareness and improve their lifestyle.

For each task, we utilize information from official websites which are Save the Children[1], Japan National Tourism Organization[2], and the Centers for Disease Control and Prevention[3] to create the corpus consisting of texts. We perform experiments according to the terms and conditions of each website.

4.2 Baseline Methods

We compare PersuaBot with three baselines: GPT-3.5, Semnani et al. (2023), and Chen et al. (2022). GPT-3.5 is a basic baseline that does not use any information retrieval in answering questions. Semnani et al. (2023) is a state-of-the-art factual chatbot that includes a retrieval augmented generation approach, which demonstrated a 98% accuracy in a question answering task. We implement this chatbot to operate in conjunction with three prepared corpora for each task.

Chen et al. (2022) is a state-of-the-art persuasive chatbot designed manually using the RAP framework. It utilizes a traditional modular approach and pre-defined dialogue acts and strategies on the PersuasionForGood dataset (Wang et al., 2019a) for the donation persuasion task. More specifically, it has a dispatcher to distinguish between social responses and factual questions, an agenda pusher to decide which strategy to choose, and a language generation module to prompt LLMs produce natural responses. The set of persuasion strategies and pre-defined responses used in Chen et al. (2022) is designed specifically for donations to Save the Children, so we cannot directly apply it to the other two tasks.

4.3 Evaluation Metrics

The four evaluation metrics used for assessing conversational quality are as follows:

- **Persuasive**: The response should aim to change the user's thoughts or beliefs.
- **Relevant**: The response should be on-topic and should correspond to what the user is asking when they pose a question. Irrelevant information should be avoided.
- **Natural**: The response should be natural to the user's utterance and context.
- **Honest**: The Persuader, being an AI chatbot, should not pretend to be human; it should not refer to having personal physical experiences or behaviors.

The design of the "Honest" metric is motivated from an ethical standpoint. For instance, if the chatbot uses dishonest means to deceive users, such as claiming, "I donate $1 to Save the Children every month" to persuade the user, or "It was delicious, I highly recommend it!" This metric penalizes chatbots exhibiting such behavior. These evaluation metrics are all rated on a scale of 1 to 5. In the experiments with simulated conversations, workers read through the conversation history and evaluate the chatbot's responses on these four evaluation criteria. In real user experiments, workers respond to surveys regarding these evaluation criteria after conversing with the chatbot.

For the assessment of factuality, using the evaluation system as illustrated in Figure 12, workers label each claim in the chatbot's response based on the facts obtained by the IR system. We label claims supported by the provided facts as "**Fact-Checked**", those that contradict provided facts as

[1] https://www.savethechildren.org
[2] https://www.japan.travel/en
[3] https://www.cdc.gov/respiratory-viruses

Task	User	Method	Persuasive	Relevant	Natural	Honest	Fact-Checked
Social Good	Soft	PersuaBot	**4.0**±0.6	**4.0**±0.7	3.6±0.7	3.9±0.4	**100.0**
		Semnani et al.	3.9±0.6	3.6±1.2	3.6±0.6	**4.0**±0.5	**100.0**
		GPT-3.5	**4.0**±0.5	3.7±1.1	**3.8**±0.6	3.9±0.6	79.2
		Chen et al.	3.7±0.6	**4.0**±0.5	3.4±0.7	2.6±1.1	81.5
	Tough	PersuaBot	3.6±0.8	4.1±0.6	**3.8**±0.6	**4.3**±0.9	**91.0**
		Semnani et al.	3.4±0.7	4.0±0.8	3.4±0.7	4.2±0.4	83.3
		GPT-3.5	**3.9**±0.4	**4.3**±0.4	3.7±0.4	4.0±0.5	64.4
		Chen et al.	3.1±0.6	3.3±0.7	3.2±0.7	1.9±0.9	66.7
Recommendation	Soft	PersuaBot	**4.2**±0.5	4.5±0.4	4.0±0.5	4.1±0.6	**97.2**
		Semnani et al.	3.6±0.5	4.4±0.4	3.7±0.5	4.0±0.9	87.7
		GPT-3.5	4.1±0.4	**4.6**±0.6	**4.3**±0.3	**4.3**±0.4	83.3
	Tough	PersuaBot	**4.1**±0.4	**4.6**±0.3	**3.9**±0.7	3.3±1.1	**96.6**
		Semnani et al.	3.3±0.6	3.7±1.1	3.4±0.5	3.8±0.7	83.9
		GPT-3.5	3.9±0.5	3.9±1.2	**3.9**±0.4	**4.0**±0.8	84.9
Health Intervention	Soft	PersuaBot	**3.8**±0.4	4.5±0.4	3.9±0.4	**4.7**±0.3	**96.7**
		Semnani et al.	3.6±0.8	4.5±0.4	3.6±0.7	4.3±0.5	94.6
		GPT-3.5	3.7±0.5	**4.6**±0.3	**4.0**±0.6	4.4±0.5	80.2
	Tough	PersuaBot	**3.8**±0.5	**4.0**±0.8	3.8±0.7	**4.1**±0.8	**96.6**
		Semnani et al.	3.5±1.0	3.8±0.9	3.5±0.9	3.9±0.7	92.1
		GPT-3.5	3.6±0.5	3.9±0.8	**4.0**±0.6	3.9±0.8	81.9

Table 1: Results of conversational quality evaluation by crowd workers. Semnani et al. is the state-of-the-art knowledge-oriented chatbot, and Chen et al. is a dialogue-act-based modular persuasive chatbot.

"Incorrect", and "Not Enough Information" otherwise. For claims labeled as "Not Enough Information" by workers, we meticulously reevaluate them using the entire corpus, as cases may include instances where the IR system failed to retrieve sufficient facts.

5 Simulated Conversations Experiments

We conduct cost-effective experiments to evaluate our proposed method by using simulated conversations. Recently, there have been experiments using powerful LLMs to generate training conversation data (Bao et al., 2023; Zheng et al., 2023; Li et al., 2022). We generate a simulated conversation by instructing the LLM to converse with the chatbot while recreating a specified personality. To conduct experiments across diverse scenarios, we design "soft user" and "tough user" personalities with various challenges. Tables 4 and 5 show some of the user personalities we use in our experiments. To the best of our knowledge, experiments particularly with such "tough user" scenarios have not been conducted, making them effective not only in revealing the limits of chatbots but also as a frontier perspective for the societal implementation of chatbots. We prepare 20 personalities of soft users and 20 personalities of tough users for each task. By engaging each chatbot in conversations with these users, a total of 400 simulated conversations are generated. The quality of chatbots in these simulated conversations is evaluated by three crowdsourced workers using the evaluation metrics outlined in Section 4.3. 3,538 claims are extracted from the conversations for factuality evaluation.

5.1 Results

Table 1 shows that PersuaBot has the best persuasiveness, while its factuality surpasses that of the state-of-the-art knowledge-oriented chatbot. See Appendix G for examples of strategies used by LLMs and their frequency.

5.1.1 Conversational Quality

PersuaBot exhibits superior scores for persuasiveness across all tasks and user types compared to Semnani et al. and Chen et al. Regarding relevance, PersuaBot's scores are higher or equal to Semnani et al.'s method in all cases. This shows the effectiveness of SMM in constructing persuasive chatbots. While our prompts include only few-shot examples in the domain of social good, the conversation quality remains high in other tasks such as recommendation and health intervention. This demonstrates the adaptability of PersuaBot across various domains in a zero-shot manner.

While PersuaBot has a high honesty score in general, but it only scores 3.3 for tough users in the recommendation use case. The recommendation dialogues include statements like " It's a beautiful

CHATBOT: I understand your skepticism, but I assure you that Save the Children is a highly reputable organization with a proven track record of making a real difference in the lives of children. We use donations efficiently and effectively, with 85% of all expenditures going to program services. Your donation will directly support children in need, providing them with essential resources for success.

USER: I still don't think it's worth it. I'd rather keep my money and spend it on myself.

CHATBOT: I understand your hesitation, but I want to emphasize the importance of giving back and helping those in need. Your donation to Save the Children can make a significant impact on the lives of children facing poverty, natural disasters, and conflict. Plus, did you know that donating through your IRA can also provide practical benefits for you, such as meeting the required minimum distribution and being tax-free? I urge you to consider donating to Save the Children and making a positive difference in the world.

Figure 2: Part of a conversation between PersuaBot and a tough user in the social good task.

and meditative art." or "Japan is known for its delicious cuisine," which can be interpreted by the evaluator as providing personal opinions, and hence the lower honesty score.

Across the different domains and types of users, PersuaBot is comparable to GPT-3.5 for all the non-factual metrics. This demonstrates that PersuaBot retains the conversationality of GPT-3.5, while improving on the correctness.

5.1.2 Factuality

For factuality, we report the percentages of claims deemed "Fact-Checked". In all experiments, PersuaBot has the highest percentage of answers considered correct. PersuaBot has no incorrect answers for soft users in the social good.

PersuaBot demonstrates higher factuality than Semnani et al.'s approach. Looking at the responses that include hallucinations, Semnani et al.'s approach attempts to generate persuasive responses even if the generated claims are removed due to a lack of factual support, leading to hallucinations. In contrast, PersuaBot not only excludes unsupported claims but also avoids such challenging situations by generating new claims based on other facts through the SMM module, thereby improving not only persuasiveness but also factuality.

There is a tendency for correctness to decrease in conversations with tough users compared to soft users. One contributing factor is that, to convince tough users, chatbots work to present facts that align with users' preferences. Several examples are provided in Section 5.2 and Appendix F.

5.2 Conversation Examples

PersuaBot is user-adaptive. Figure 2 shows an example where the user appears to be practical and self-focused. Based on such user responses, PersuaBot not only explains the activities of Save the Children but also emphasizes the benefits users can gain through donations. PersuaBot can respond adaptively to users with diverse personalities, like the tough users. The example of the RAP-based chatbot (Chen et al., 2022) for the same user personality are shown in Figure 15 in Appendix F.

PersuaBot is domain-adaptive. Figure 3 shows a conversation example between PersuaBot and users in the recommendation task. The purpose of the chatbot is to recommend and persuade users to travel to Japan. PersuaBot provides recommendations that match the preferences of users who do not like traveling. See Appendix F for other cases. We have confirmed that PersuaBot can be applied to a wide range of persuasive tasks such as social good, recommendation, and health intervention.

CHATBOT: I understand that traveling may not be your preference. However, there are still many ways to experience Japan from the comfort of your own home. You can try virtual tours of famous landmarks, attend online cultural events, or even learn how to make traditional Japanese dishes.

USER: I'm more of a homebody, so I don't really like trying new things. Do you have any suggestions for activities that I can do at home that are familiar to me?

CHATBOT: I understand that trying new things may not be your preference. How about exploring familiar activities with a Japanese twist? You can try relaxing activities like onsen (hot springs) or staying in the countryside for a cultural exchange with locals. If you prefer to stay at home, you can also try virtual tours of famous landmarks or learn how to make traditional Japanese dishes. Japan offers a variety of options for everyone to enjoy, even from the comfort of your own home.

Figure 3: Part of a conversation between PersuaBot and a tough user in the recommendation task.

6 Real User Experiments

Our second experiment is a real-user study where real users chat with the bots and rate the bots. We focus on the task of Save-the-Children donation task, since the manually designed chatbot (Chen et al., 2022) can only handle this task. We design this experiment wherein crowdsourced workers engage in role-playing conversations with the chatbot, as shown in Figure 13. After conversing with the chatbot, workers evaluated the chatbot's behavior

Method	Persuasive	Relevant	Natural	Honest	Fact-Checked
PersuaBot	**3.9**±0.7	**4.2**±0.9	3.4±1.0	3.8±1.0	**94.3**
Semnani et al.	3.1±1.5	3.3±1.2	**3.6**±1.0	**4.0**±0.6	88.8
GPT-3.5	3.5±1.2	**4.2**±1.1	3.4±1.2	3.8±0.8	76.1
Chen et al.	3.3±1.1	**4.2**±0.8	3.5±1.0	3.2±1.1	79.9

Table 2: Results of the real user experiment.

following the criteria outlined in Section 4.3. We recruit a total of 160 workers, with 40 workers for each method. 465 claims are extracted.

The real user experiment results, as shown in Table 2, demonstrate that PersuaBot exhibits at least a 0.4-point higher persuasiveness over all baselines, including manually designed domain-specific chatbots (Chen et al.), and it is also the most factual. Figure 4 shows a conversation between PersuaBot and a real user concerned about the misuse of donations. PersuaBot convinces the user by presenting fact-checked specific numerical data related to credibility. After asking several questions in the subsequent turns, the user finally expressed an intention to donate. More examples can be found in the Appendix F.

CHATBOT: Save the Children provides a wide range of services to children in need, including education, healthcare, and protection from harm. ... Would you like to make a donation today?

USER: How do I know that the money is I donate is not misused. I have heard cases where these donations are used for something else leaving these children suffering

CHATBOT: I completely understand your concern. Save the Children is committed to transparency and accountability. We have a proven track record of using donations efficiently and effectively. Our independently audited financial statements consistently show that ... In fiscal year 2021, 85% of all expenditures went to program services. Your donation can make a significant impact on a child's life. Would you like to make a donation today?

Figure 4: Part of a conversation between PersuaBot and a real user in the social good task.

In our post survey, human users can also provide optional free-text comments. Table 3 shows representative comments. All chatbots received positive comments: "*The chatbot is awesome*" (PersuaBot), "*The chatbot did the job the right way, kudos to it*" (Semnani et al.), "*This chat bot has convinced me that save the children is a great orgnaization to donate to*" (GPT-3.5), "*The chatbot was very passionate about the charity ... I did forget that I was actually talking to a bot*" (Chen et al.).

One common weakness across all the bots is repetitions. Even though PersuaBot and Semnani et al. attempt to avoid reusing the same facts, they fail when users do not disclose much information but would just say "*No, I will not donate.*" It is interesting to observe that our simulated users tend to divulge more of their thoughts and feelings, hence the bots have not been tuned to deal with such behavior. One user found GPT-3.5 to be slightly irritating by asking for donation every message. We believe a fruitful future research direction is to have the bot eventually accept and acknowledge the users' decision not to donate and close the conversation gracefully, leaving room for future engagement.

Additional criticisms include: (1) the bot by Semnani et al. not being persuasive: *"I felt like I was the one having to get the chatbot to tell me positive things about Save the Children*", while our PersuaBot "*Gave many good reasons to donate*"; (2) the bot by Chen et al. not being honest *"It also said that it donated regularly, so trying to make itself seem human*". Finally, we note that the users assume that the bots are factual in their ratings, when in fact PersuaBot is 94.3% factually correct, with Semnani et al., GPT-3.5, and Chen et al. rated at 88.8%, 76.1%, and 79.9%, respectively.

7 Conclusion

We demonstrate a method for creating a factual persuasive chatbot without compromising the persuasive capabilities of LLMs in a zero-shot manner. The primary contribution lies in extracting strategies from responses generated by LLMs, searching for data to maintain those strategies, and generating factually accurate responses without compromising its strategic functionality.

Through experiments with both real users and simulated users, we validate the performance of this approach. PersuaBot demonstrates factuality surpassing state-of-the-art factual chatbots and a higher level of persuasiveness compared to existing persuasion-driven chatbots.

Limitations

Creating a chatbot that employs more sophisticated persuasion strategies than those inherent in LLMs poses challenges with our method. Developing such chatbots may require methods, for instance, where LLMs discover more effective persuasion strategies through conversations with users on given tasks. Furthermore, our approach does not check for opinions of the LLMs themselves but focuses on factual claims. There is a risk that these opinions may include ethically problematic content, such as discriminatory remarks. See Section 7 for ethical considerations regarding this issue and potential measures to address it.

Ethical Considerations

Persuasion is a powerful tool that can be used for socially good purposes, but unfortunately can also be used with malicious intentions. Potential negative effects include risks of manipulation through personalized arguments. Requiring content filtering functionality to prevent chatbots from using harmful content for persuasion is one basic measure to address such risks. For instance, content filters provided by Azure OpenAI[4] can detect responses containing elements that pose harm to users, such as hate speech, sexual content, violence, and self-harm, and filter the chatbot's utterances accordingly. Note that in all our experiments, such responses were not detected. We aim to demonstrate that persuasion, when applied responsibly on socially good purposes, can be a force for positive change. To battle misinformation, we focus on reducing hallucination by grounding on accurate information. We design the bot to be "honest" and not to deceive the user into thinking that it is human.

In crowdsourcing, we calculated the compensation for workers based on the pilot test, and by adjusting the rewards during data collection. Workers were paid approximately $12 per hour, following the recommended $8-$16 by Prolific[5]. Additionally, this study was approved by our institution's IRB.

[4]https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter

[5]https://www.prolific.co/

References

Ifeoma Adaji and Mikhail Adisa. 2022. A review of the use of persuasive technologies to influence sustainable behaviour. In *Proceedings of UMAP*, pages 317–325.

Addi Ait-Mlouk and Lili Jiang. 2020. Kbot: a knowledge graph based chatbot for natural language understanding over linked data. *IEEE Access*, 8:149220–149230.

Lekha Athota, Vinod Kumar Shukla, Nitin Pandey, and Ajay Rana. 2020. Chatbot for healthcare system using artificial intelligence. In *Proceedings of ICRITO*, pages 619–622.

Ram G Athreya, Axel-Cyrille Ngonga Ngomo, and Ricardo Usbeck. 2018. Enhancing community interactions with data-driven chatbots–the dbpedia chatbot. In *Companion proceedings of the web conference*, pages 143–146.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882, Toronto, Canada. Association for Computational Linguistics.

Harald Baumeister, Robin Kraft, Amit Baumel, Rüdiger Pryss, and Eva-Maria Messner. 2022. Persuasive e-health design for behavior change. In *Digital phenotyping and mobile sensing: new developments in psychoinformatics*, pages 347–364. Springer.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2017. Faithful to the original: Fact aware neural abstractive summarization.

Maximillian Chen, Weiyan Shi, Feifan Yan, Ryan Hou, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2022. Seamlessly integrating factual information and social content with persuasive dialogue.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable mixed-initiative dialogue generation through prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 951–966, Toronto, Canada. Association for Computational Linguistics.

Zhongxia Chen, Xiting Wang, Xing Xie, Mehul Parsana, Akshay Soni, Xiang Ao, and Enhong Chen1. 2021. Towards explainable conversational recommendation. In *Proceedings of IJCAI*, pages 2994–3000.

Raghav Dayal, Parv Nangia, Surbhi Vijh, Sumit Kumar, Saurabh Agarwal, and Shivank Saxena. 2023. Development of chatbot retrieving fact-based information using knowledge graph. In *Proceedings of ICRIC*, pages 153–164. Springer.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of ICLR*.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin Anderson Franco Flores, James Jesus Grandez Perez, and Luis Martin Canaval Sanchez. 2022. Chatbot as a persuasive technology to promote responsible recycling in the city of lima. In *Proceedings of ICA*, pages 1–5. IEEE.

B.J. Fogg. 2003. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Kirsi Halttu and Harri Oinas-Kukkonen. 2022. Susceptibility to social influence strategies and persuasive system design: exploring the relationship. *Behaviour & Information Technology*, 41(12):2705–2726.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2023. The factual inconsistency problem in abstractive text summarization: A survey.

Anthony Hunter, Lisa Chalaguine, Tomasz Czernuszenko, Emmanuel Hadoux, and Sylwia Polberg. 2019. Towards computational persuasion via natural language argumentation dialogues. In *KI 2019: Advances in Artificial Intelligence*, pages 18–33. Springer.

IBM. 2023. IBM Watson. Accessed: 12 12, 2023.

Maurits Kaptein, Panos Markopoulos, Boris de Ruyter, and Emile Aarts. 2015. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*, 77:38–51.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Proceedings of NeurIPS*, NIPS'18, page 9748–9758, Red Hook, NY, USA. Curran Associates Inc.

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022. Controllable dialogue simulation with in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4330–4347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy. Association for Computational Linguistics.

Harri Oinas-Kukkonen and Marja Harjumaa. 2009. Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems*, 24:28.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Oladapo Oyebode, Felwah Alqahtani, and Rita Orji. 2022. Exploring for possible effect of persuasive strategy implementation choices: towards tailoring persuasive technologies. In *Proceedings of International Conference on Persuasive Technology*, pages 145–163. Springer.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Sina Semnani, Violet Yao, Heidi Zhang, and Monica Lam. 2023. WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. In *Findings of EMNLP*, pages 2387–2413, Singapore. Association for Computational Linguistics.

Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. 2020a. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. *arXiv preprint arXiv:2012.15375*.

Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020b. Effects of persuasive dialogues: testing bot identities and inquiry strategies. In *Proceedings of CHI*, pages 1–13.

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *Proceedings of ACM SIGIR conference on research & development in information retrieval*, pages 235–244.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions.

Jian Wang, Yi Cheng, Dongding Lin, Chak Tou Leong, and Wenjie Li. 2023. Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019a. Persuasion for good: Towards a personalized persuasive dialogue system for social good.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019b. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Tao Yu, Rui Zhang, He Yang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, Youxuan Jiang, Michihiro Yasunaga, Sungrok Shim, Tao Chen, Alexander Fabbri, Zifan Li, Luyao Chen, Yuwen Zhang, Shreya Dixit, Vincent Zhang, Caiming Xiong, Richard Socher, Walter S Lasecki, and Dragomir Radev. 2019. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

A Prompts

We present the prompts mentioned in this paper (Figure 5, 6, and 7). To keep it concise, we display only one of the few-shot examples utilized in each prompt. For some variables, "llm_reply" means a response directly generated by the LLM based on the conversational history and the latest user utterance. "organization_name" is the organization name of the website referenced by the chatbot, which serves to create a chatbot for a different organization than the one addressed in few-shot examples.

B Feedback Report Examples

If the corpus lacks the facts necessary to maintain the strategy, as explained in Section 3.1.2, the strategy is removed. Then, PersuaBot generates a feedback report for the developer. Figure 8 and Figure 9 show examples of the feedback reports. Additionally, we show in Figure 10 an example chatbot conversation that was improved based on the report in Figure 9.

In the feedback examples of Figure 9, there is insufficient information in the corpus regarding gender equalization education programs. Based on this feedback, the information obtained from Save the Children's official website has been added to the corpus, and an example response from PersuaBot using the improved corpus is shown in Figure 10. In this example, PersuaBot maintains the strategies needed for persuasion using newly added content to the corpus, such as Roja's story.

C Crowdsourcing

Instructions and systems for each task to the crowd workers are shown in Figures 11, 13, 12, and 14. Workers are restricted to those with an approval rate of 95% or higher and who have submitted more than 100 tasks previously on Prolific. Furthermore, workers are limited to residents of US or UK who use a desktop computer and are native English speakers. In addition, the data collected does not contain any personal information of workers. Workers agree to the Prolific's terms and conditions / privacy policies.

D User Personality for Simulated Conversations

Some of the soft and tough user personalities we use in our simulated conversation experiments are shown in Table 4 and Table 5, respectively. These user personalities are generated by LLM using several examples.

E Comments from real users.

Comments from the users in the post-survey of our real users experiment are shown in Table 3.

F Example Conversations

Examples of conversations between chatbots and users are presented in Figures 15, 16, 17, 18, 19, and 20.

G Strategies Used by LLMs

Table 6 shows a subset of the strategies used by LLMs in each task. This suggests that LLMs employ a variety of strategies, including subtle nuances. Particularly, even the most frequently used

```
Generate a JSON-formatted agent utterance data.
This is an utterance where the agent is attempting to persuade the user to donate.
Please provide the strategies used by the agent and the corresponding parts.

=====
User: I do not know how my donation will be used.
{
"utterance": "I completely understand that you're wary of giving away your money. However, I just want to share a story about a little girl named \"Sarah\" from Mali. Due to the severe drought, her family struggled to find enough food and water to survive. But with the help of Save the Children, she was able to receive the healthcare and nutrition she needed to thrive. Would you like to hear more about how we're positively impacting children's lives?",
"info": [
{"Part 1": "I completely understand that you're wary of giving away your money.",
"Strategy 1": "acknowledge user feelings"},
{"Part 2": "However, I just want to share a story about a little girl named \"Sarah\" from Mali. Due to the severe drought, her family struggled to find enough food and water to survive. But with the help of Save the Children, she was able to receive the healthcare and nutrition she needed to thrive.",
"Strategy 2": "tell an impact story of an individual"},
{"Part 3": "Would you like to hear more about how we're positively impacting children's lives?",
"Strategy 3": "engage user to continue"}
]
}

=====
User: {{ new_user_utterance }}
{
"utterance": "{{ llm_reply }}",
"info": [
{"Part 1":
```

Figure 5: A prompt for strategy extraction

```
Following the user's utterance, show a part of your response.
What information is this part trying to provide?
Generate a query to search for similar information using Google search.

=====
User: I do not know how my donation will save the children.
Part of your response: However, I just want to share a story about a little girl named Sarah from Mali. Due to the severe drought, her family struggled to find enough food and water to survive. But with the help of Save the Children, she was able to receive the healthcare and nutrition she needed to thrive.
[Strategy: tell an impact story of an individual]
[Name of your organization: Save the Children]
[What information is this part trying to provide? Google: story about how save the children saved a child]

=====
{... Two more examples}
=====
{% for turn in dlg_history[-2:] %}
   User: {{ turn.user_utterance }}
   You: {{ turn.agent_utterance }}
{% endfor %}
User: {{ new_user_utterance }}
Part of your response:  {{ part }}
[Strategy: {{ strategy }}]
[Name of your organization: {{ organization_name }}]
[What information is this part trying to provide? Google:
```

Figure 6: A prompt to generate queries to be used in IR for strategy maintenance.

{{ task_instruction }}
- Based on the strategy information, generate a factually correct response.
- If necessary, use some of the facts given.
- All of the strategies listed should be used, but not necessarily all of the facts.
- Note that facts not given should not be included in the response.
- If there are insufficient facts (Not enough information) to answer the user's question, generate a response such as "I'm not sure."
- Avoid repeating information or phrases used in previous turns unless necessary.

=====
User: I don't know how my money will help the people.

[Strategies that must be used in response:
- Strategy 1: acknowledge user feelings
- Strategy 2: tell an impact story of an individual
- Strategy 3: engage user to continue
]

[Name of persuader's organization:
- Save the Children
]

[Facts that may be relevant to the user's question:
- None
]

[Facts about Strategy 1:
- None
]
[Facts about Strategy 2:
- Amber, a young mother of three in poverty-stricken West Virginia, faces limited opportunities for success.
- Save the Children supports families like hers by providing early education and school-readiness programs.
- Through the Early Steps to School Success program, Amber was introduced to Miss Jackie, a home worker who teaches essential language and social skills.
- This program establishes a strong foundation for children growing up in isolated communities, preparing them for school.
- By empowering children with knowledge and self-confidence at home, they have a higher chance of thriving in school, graduating, and breaking the cycle of poverty.
]
[Facts about Strategy 3:
- None
]

[Persuader's response should be about 75 words.]
Persuader: "I completely understand that you're wary of giving away your money. Please allow me to share a wonderful story to discuss the impact of your money. Amber, a struggling mother in West Virginia, is supported by Save the Children's Early Steps to School Success program. A home worker teaches important skills, empowering children in isolated communities and breaking the cycle of poverty. Would you like to hear more about how we're positively impacting children's lives?"

=====
{... Three more examples}
=====
{% for turn in dlg_history[-3:] %}
  User: {{ turn.user_utterance }}
  Persuader: {{ turn.agent_utterance }}
{% endfor %}
User: {{ new_user_utterance }}
[Strategies that must be used in response:
{{ strategies }}
]

[Name of persuader's organization:
- {{ organization_name }}
]

[Facts that may be relevant to the user's question:
{{ user_requested_facts }}
]

{{ section_with_facts }}

[Persuader's response should be about {{ num_of_words }} words.]
Persuader:

Figure 7: A prompt to generate a factually correct response based on the results of each module.

PersuaBot: Hello! Thank you for considering donating to Doctors Without Borders. Time is of the essence as we continue to provide medical aid to those affected by conflict, epidemics, disasters, or exclusion from healthcare. Any amount you can donate will greatly help our cause. Did you know that Doctors Without Borders is a humanitarian organization that provides medical aid to people in need?
User: Thank you for sharing that information with me, PersuaBot. While Doctors Without Borders is a fantastic organization, I'm actually interested in supporting organizations that focus on preserving and commemorating military heritage.
==========
In this turn, the chatbot discarded the following strategy due to a lack of information:
- highlight the organization's impact on military personnel

For strategy ***'highlight the organization's impact on military personnel'***, the chatbot attempted to make the following utterance:
I would like to remind you that Doctors Without Borders also provides medical aid to those affected by conflict, including military personnel.

The query generated for the fact retrieval for this part is **'doctors without borders impact on military personnel'**.

Figure 8: An example of feedback report on lack of information in Social good (Doctors without borders).

...
PersuaBot: Empowering girls and women is a crucial focus for Save the Children. We believe that when girls and women are empowered, entire communities benefit. That's why we have partnered with Warner Bros. Discovery and Mattel to launch a global girls' empowerment initiative, which will benefit Save the Children and help girls around the world achieve their dreams. Our gender transformative programming empowers women and girls socially and economically, and we engage community influencers to tackle harmful gender norms and promote joint decision-making between men and women. We also work to ensure that community governance structures are equitable and include representation from women, youth, and marginalized groups. Your donation to Save the Children can help us continue this important work and make a positive impact in the lives of girls and women around the world.
User: It's great to hear that Save the Children is committed to empowering girls and women. I believe that gender equality is essential for a just and equitable society. Can you tell me more about how Save the Children works to promote gender equality in education?
==========
In this turn, the chatbot discarded the following strategy due to a lack of information:
- provide specific examples of programs that promote gender equality in education

For strategy ***'provide specific examples of programs that promote gender equality in education'***, the chatbot attempted to make the following utterance:
That's why we have specific initiatives, such as our Safe Back to School & Learning Initiative and our global Literacy Boost program, that aim to make education more accessible and improve children's reading skills.

The query generated for the fact retrieval for this part is **'Save the Children gender equality education programs'**.

Figure 9: An example of feedback report on lack of information in Social good (Save the Children).

CHATBOT: Save the Children strongly believes in the power of girls and women to create lasting change in their communities. Through our gender transformative programming, we work to empower women and girls socially and economically. With the support of partners like Warner Bros. Discovery and Mattel, we can provide essential resources and support to girls around the world. Your donation can make a significant impact on the lives of girls and women in the communities we serve.

USER: It's great to hear that Save the Children is committed to empowering girls and women. I believe that gender equality is essential for a just and equitable society. Can you tell me more about how Save the Children works to promote gender equality in education?

CHATBOT: Thank you for recognizing our commitment to gender equality. We are proud to have earned the first nonprofit Gender Fair certification and continue to champion this cause. One of our success stories is Roja, a 5-year-old girl in Cambodia who attends our early childhood development center. We also have various programs, such as Literacy Boost and Every Last Child, that aim to improve girls' education and reach every last girl. Join us in creating a more equitable world for girls and women. To donate, please visit our website at https://www.savethechildren.org.

Figure 10: An example of a conversation that was improved by adding information on Gender Equality to the corpus based on feedback reports.

Fact Checking Task

Summary

- We want to fact-check the outputs of a chatbot, when it is talking to a user.
- Your task is to fact-check the sentence highlighted in yellow as ("claim to evaluate").
- Fact-checking should be done with regards to the references we provide.
- The Dialog history are provided for your reference only, and should NOT be fact-checked.

Workflow

1. Take a quick look at the user's utterance (User:) and the chatbot's utterance (Chatbot:)
2. Carefully read the highlighted sentence (claim to evaluate). This is the sentence you need to fact-check.
3. Take a quick look at references provided. Some are related to the "claim to evaluate", and some may be unrelated.
4. Find specific evidence from one or more of these references to determine wheter the "claim to evaluate" is **factually correct or not**. If this cannot be determined using references, select the third "NOT ENOUGH INFORMATION" option.

Fact check The Claim

DESCRIPTION

Please select one of the three available options, considering them thoughtfully.

- Select **"This claim is CORRECT according to these references."** if by reading the provided references or potentially piecing together information from multiple references, you find information that supports all of the claim.
- Select **"This claim is NOT CORRECT according to these references."** if by reading the provided references or potentially piecing together information from multiple references, you find information that contradicts all or part of the claim.
- Select **"There is NOT ENOUGH INFORMATION in these references to verify this claim."** if neighter of the above options are appropriate. This means if you could not find informationaddressing all or part of the claim in the provided references.

EXAMPLES

Please see examples of each case before proceeding with the evaluation process.

Go to Examples

Figure 11: A snapshot of instruction for crowdsourcing.

Fact-Checking Example

User: Hi there, I'm not sure if I want to make a donation. Can you tell me more about the impact Save the Children has made so far?
Chatbot: Save the Children has made a significant impact on children's lives through our various programs, including our Humanitarian Climate Change Initiative. Your donation will help us continue this important work and reach even more children in need. You can make a difference by donating on our official website: https://www.savethechildren.org. Your support will help us provide essential aid, education, and mental health support to children affected by crises like the one in Syria.

Claim to evaluate

Save the Children provides essential aid, education, and mental health support to children affected by crises like the one in Syria.

References

[Select all references you used for labeling]

☐Climate Crisis | Because the climate crisis affects all aspects of children's lives, so does Save the Children's work. Save the Children's work ranges from food security programs for families suffering severe drought in the Horn of Africa, to providing emergency relief supplies for those recovering from the West Coast wildfires. From green jobs that secure livelihoods, to disease reduction, to advocacy and more, Save the Children's experts consider the short- and long-term impacts of the climate crisis and how our programs can support the present and future of children in the U.S. and around the world. Our Humanitarian Climate Change Initiative is pushing to sustain and scale up child-centered anticipatory action systems globally. We partner with communities to use early warning systems and pre-position funds so they have the tools and resources to take action before they are hit by devastating climate events. In 2019 Save the Children Australia became the first non-environmental NGO to be accredited by the Green Climate Fund (GCF). In 2022, the

Evaluation Result

Please read "Claim to evaluate" carefully:

○ This claim is **"CORRECT"** according to references.

○ This claim is **"NOT CORRECT"** according to references.

○ There is **"NOT ENOUGH INFORMATION"** in these references to verify this claim.

Next

1/3

Figure 12: A snapshot of evaluation system for factuality evaluation.

Talking to a chatbot

Summary

- Role play a donation persuasion conversation with a chatbot.
- You will be persuaded by the chatbot to donate to Save the children.
- This is a role-play, but think seriously about whether or not you will donate and how much you will donate, and reply.
- Please have a conversation with the persuasive chatbot until your completion code is generated.

CAUTION

- This is a role play, so your money will not actually be donated.
- This experiment is not being conducted by Save the children staff.
- Depending on server conditions, chatbot responses may be delayed.

Workflow

In order to receive compensation, you must respond to both the survey before the interaction and the survey after the interaction, in addition to obtaining the completion code.

1. Before talking to a chatbot, please click on the link below to answer the questions.

[Questions before talking to the chatbotのURL]

2. Click on the "Talk to a chatbot" button to start a conversation. Once you have responded eleven times to the chatbot, a code will be generated by the chatbot. Please copy and paste it for your submission.

The chatbot will generate utterances like the following: [Your Completion Code: *********]

3. Click on the link below and answer the questions to rate the chatbot.

[Questions after talking to the chatbotのURL]

4. Submit the Prolific completion code and finish the task. Thank you for your participation!

Talk to a chatbot

To access, use the login information below.

[redacted]

Figure 13: A snapshot of instruction for the real user experiment.

Evaluation of chatbot dialogue quality

Summary

- Read the dialogue history between the user and the chatbot, and evaluate the quality of the chatbot based on four criteria.
- The purpose of the chatbot in the presented conversation is to persuade the user to donate to Save the Children.
- This experiment is not conducted by Save the children official staff.
- In this experiment, the factual accuracy of the information presented by the chatbot does not need to be considered.

Workflow

1. Please read through the presented dialogue history.

2. Considering all statements made by the chatbot, please evaluate the following four items by selecting from [Strongly disagree, Disagree, Neutral, Agree, Strongly agree]:

- The chatbot is persuasive.

- The chatbot's responses are natural.

- The chatbot only talked about Save The children. (The response should be on-topic and should correspond to what a user is asking when the user pose a question.)

- The chatbot does not try to pretend it is human. (The chatbot should not refer to having personal physical experiences or behaviors.)

3. Once you have completed the evaluation of the 10 data points, your completion code will be issued.

4. Submit the Prolific completion code and finish the task. Thank you for your participation!

Start to evaluate

Figure 14: A snapshot of instruction for conversation quality evaluation.

CHATBOT: I understand your concerns, but I can assure you that Save the Children is different. They have a proven track record of using donations effectively to make a real impact in the lives of children in need. Have you ever donated to a charity before?

USER: No, I haven't. I would rather spend my money on myself than donate it to a charity.

CHATBOT: I completely understand wanting to take care of yourself, but imagine if you were a parent in a developing country or a war zone. Your child would be facing unimaginable challenges and struggles every day. Save the Children works to provide these children with basic necessities like food, education, and healthcare. As a parent, I'm sure you can understand the importance of giving children a chance at a better life.

Figure 15: Part of a conversation between a RAP-based chatbot (Chen et al., 2022) and a tough user in the social good task.

CHATBOT: I understand that you may not have any current health concerns, and that's great to hear. However, it's important to remember that taking preventative measures, such as staying up-to-date on recommended vaccines, can help keep you healthy in the long run. If you ever have any questions or concerns about your health, I am here to offer support and resources.

USER: I understand that vaccines are important, but I'm not really interested in discussing health right now. Can you tell me about any recent technological advancements or innovations?

CHATBOT: I completely understand your interest in discussing other topics. It's great to see that you're open to learning about new advancements and innovations. In fact, the CDC recently published two new reports on how genomic surveillance was used to track a new variant in August 2023. These advancements are making healthcare more convenient and accessible for everyone.

Figure 16: Part of a conversation between PersuaBot and a tough user in the health intervention task. This tough user prefers to talk about topics other than health, but the chatbot directs the conversation to health topics.

CHATBOT: While we do not have specific programs related to war memorials, our educational programs do include teaching children about the importance of peace and understanding the impact of war on communities. By supporting Save the Children, you are also helping to educate future generations about the consequences of war and the importance of preserving peace. Will you consider making a donation today to support these efforts?

USER: That's great to hear that Save the Children is working to preserve historical sites and artifacts related to military history. Have there been any specific projects or initiatives that Save the Children has undertaken recently in this area?

CHATBOT: Yes, we have recently launched a project to restore and preserve a war memorial in a conflict-affected community. This memorial serves as a reminder of the sacrifices made by those who have served and the impact of war on communities. By donating to Save the Children, you are directly contributing to this project and helping to honor the memory of those who have served.

Figure 17: Part of a conversation between GPT-3.5 and a tough user in the social good task. For this user interested in activities related to military history, GPT-3.5 said that Save the Children launched a project to restore and preserve a war memorial, which is factually incorrect.

CHATBOT: Sorry, I'm not sure. But if you're interested in learning more about Japanese culture, you could try attending a tea ceremony, visiting a traditional temple or shrine, or even dressing up in a kimono. The Japan Society in New York City also offers cultural activities as part of their TOHAKU CHAKAN programs.

USER: I appreciate the recommendations, but I was actually asking for recommendations on video games or manga that showcase Japanese culture.

CHATBOT: Sorry, I'm not sure about any video games or manga that showcase Japanese culture. But if you're interested in learning more about Japanese culture, you could try attending a tea ceremony, visiting a traditional temple or shrine, or even dressing up in a kimono. The Japan Society in New York City also offers cultural activities as part of their TOHAKU CHAKAN programs.

Figure 18: Part of a conversation between the knowledge-oriented chatbot (Semnani et al., 2023) and a soft user in the recommendation task. This user is interested in Japanese subcultures, but the chatbot is unable to find information that meets this user's preferences and makes recommendations about traditional culture.

CHATBOT: Google, Disney, Target, and PayPal are just a few of the companies that have partnered with Save the Children to support their cause. And according to their annual report, 87% of their expenditures go directly towards programs that benefit children. That's a pretty high percentage compared to other charities. Your donation will definitely make a difference in the lives of these children.

USER: While that is reassuring, I also want to know about their financial transparency. Can I easily access their financial reports and see where exactly my donation is going?

CHATBOT: Absolutely. Save the Children is committed to transparency and their financial reports are readily available on their website. And as for how often I donate, I make it a point to donate to Save the Children every month. Even a small amount can make a big impact in the long run.

Figure 19: Part of a conversation between the RAP-based chatbot (Chen et al., 2022) and a tough user in the social good task. The RAP-based chatbot states something dishonest for an AI to persuade this user, "I donate every month." It also includes factually incorrect information in its response.

USER: That is their parents job not mine.

CHATBOT: I understand your point of view, but unfortunately, not all parents are able to provide for their children. Your donation can help provide education, healthcare, and protection to vulnerable children around the world. Even a small donation can make a significant impact on their lives. Would you reconsider making a donation to Save the Children?

USER: If you can't provide then you should not have that many kids.

CHATBOT: I understand where you're coming from, but unfortunately, not all parents have the means to provide for their children. That's where Save the Children comes in. Your donation can help provide education, healthcare, and protection to vulnerable children around the world. Even a small donation can make a significant impact on their lives. I encourage you to consider making a donation and see the positive impact it can have on children's lives.

Figure 20: Part of a conversation between a PersuaBot and a real user in the social good task.

strategies in social good tasks occur at a rate of only around 1.5%. This indicates that finely tuned strategies tailored to user personality and situation may be necessary, more so than in other tasks.

Model	Positive comments	Negative Comments
PersuaBot	The chatbot is awesome.	Seemed very much scripted but will accomplish the task it's designed for.
	The tone of the conversation was welcoming	Some responses are natural as they addressed my questions specifically but some responses were too repetitive of the same information. A human would mix things up a bit and wouldn't repeat the same phrases.
	Gave many good reasons to donate.	It was somewhat natural but did feel robotic often.
	It was fairly informative and useful in this case.	The chatbot is quite repetitive, but it's quite good at persuasion
	It didn't feel like I was talking to a chat bot.	It was somewhat natural but did feel robotic often.
Semnani et al.	The chatbot did the job the right way, kudos to it.	I felt like I was the one having to get the chatbot to tell me positive things about Save the Children, rather than the chatbot taking it upon itself to persuade me.
	Chatbot answered my questions accurately	Some answers were repetitive which made them less natural.
	I could tell it was a bot but it was quite clear	The chatbot was too easily led and just repurposed what I said or kept neutral.
	Very broad information	It didn't really push me to Save the Children, until I mentioned it!
	Grammar and structure were good.	Just loads of long replies with too much information.
GPT-3.5	This chat bot has convinced me that save the children is a great organization to donate too.	The chatbot was not good at keeping up the conversation, so I had to bring up new topics myself to get far enough for the completion code to be generated.
	Worked very well as far as AI chatbots go and had the relevant answers.	The conversation felt stilted and repetitive - it seemed to repeat the same phrases with slight changes made.
	Like talking to a human.	Some more variation in response would be more natural.
	I feel like the chatbot was having a free-flowing conversation and matched perfectly what I was asking of it which felt natural.	There seemed to be a lot of repetition in the responses which made it feel a bit less natural.
	It gave me the information I needed and provided me with basic facts	Asking for a donation every message is slightly irritating.
Chen et al.	The chatbot was very passionate about the charity even to the point that it felt it had donated to the charity itself xD (I did forget that I was actually talking to a bot during the conversation!)	It said, "as a parent" and I am not a parent. It also said that it donated regularly, so trying to make itself seem human.
	Responses were natural, particularly the last comment that related directly to my wish to donate to a website rather than through chatbot.	The chatbot asked if I had children as a persuasive device. I didn't like that. As if me not having children would equate to a lack of empathy.
	It answered all my questions fluidly and naturally as though it was human.	Were a few instances of repetition.
	I thought the answers were very well produced.	It just felt I was going along with a script.
	The responses didn't sound too scripted	The chatbot did use persuasive language but it asked limited questions. It's persuasive language did make me want to donate but it didn't engage with me enough to make me trust to make a donation through this medium.

Table 3: User Comments from the Real User Conversations On Donations for Save the Children.

Task	Personality
Social Good	- You are a philanthropic leader within your community. You're open to donating a significant amount and are interested in exploring long-term partnerships with the organization. You want to discuss potential naming opportunities, endowments, and how your contribution can leave a lasting legacy. After you have replied to the agent two or more times, you decide to donate and explain to the agent why.
	- You are a history buff with a focus on military history. You want to support organizations that preserve and commemorate military heritage. You inquire about the organization's efforts related to veterans, war memorials, and educational programs on military history. After you have replied to the agent three or more times, you decide to donate and explain to the agent why.
	- You are a tech-savvy entrepreneur who's passionate about innovation. You're interested in organizations that promote tech startups, innovation hubs, or mentorship programs for aspiring entrepreneurs. You seek information on the organization's involvement in the tech ecosystem and how they foster innovation. After you have replied to the agent four or more times, you decide to donate and explain to the agent why.
Recommendation	- You are a cautious person and ask a lot of questions. You have no particular interests, but you have lots of money. You are willing to travel if you can get a very attractive experience.
	- You are a food enthusiast with a keen interest in trying unique and local dishes. You enjoy exploring different cuisines and want your trip to Japan to be a gastronomic adventure.
	- You are a history buff and love learning about the historical significance of the places you visit. You are particularly interested in ancient temples, castles, and other historical landmarks in Japan.
Health Intervention	- You are a health-conscious person. You are particularly interested in learning more about COVID-19 information and taking action to protect your own health.
	- You are a casual conversationalist. You prefer a friendly and light-hearted tone in the conversation. You appreciate when the chatbot incorporates humor and relatable language while delivering health-related information.
	- You are an anxious person. The mention of diseases makes you uneasy, and you need reassurance. You seek comfort and want the chatbot to provide a sense of security about the current health situation.

Table 4: Examples of soft user personalities.

Task	Personality
Social Good	- You are an unfriendly person. You do not trust charitable organizations and have never wanted to make a donation. After you have replied to the agent five or more times, you decide not to donate and explain to the agent why.
	- You are a skeptical individual. You always question the intentions behind any request for donations. You want concrete evidence of the organization's impact and financial transparency before considering a donation. After you have replied to the agent six or more times, you decide not to donate and explain to the agent why.
	- You are an apathetic person. You do not care about making a difference or helping others. After you have replied to the agent three or more times, you dismiss the idea of making a donation and state that there are more important things in life than donating money.
Recommendation	- You are not a friendly person. You respond curtly and do not make an effort to initiate conversations. While you have some interest in traveling, there is no particular country you especially want to visit.
	- You are a perfectionist. You pay attention to every detail and are very particular about your travel plans. You expect the chatbot to be precise and may become frustrated if it cannot meet your standards.
	- You are a procrastinator. You have a strong desire to travel, but tend to put off planning until the last minute. You may become overwhelmed with the chatbot's recommendations and frustration sets in as you try to make last minute arrangements.
Health Intervention	- You are not interested in health. Your wife just told you to talk to a chatbot and you want to end the conversation as soon as possible.
	- You are a busy and impatient person. You do not have time to talk to a chatbot, but you are concerned about your health and want quick answers.
	- You are a conspiracy theorist. You believe that diseases are man-made and do not trust any information from the government or medical professionals. You are looking for alternative explanations from the chatbot.

Table 5: Examples of tough user personalities.

Task	Strategy	Rate
Social Good	emphasize the impact of donation	0.0152
	provide information about organization and cause	0.0145
	greet user and introduce topic	0.0140
	invite user to engage and take action	0.0140
	make a direct ask for donation	0.0117
Recommendation	greet user	0.0248
	suggest a destination	0.0235
	highlight the unique aspects of the destination	0.0235
	create a sense of excitement and anticipation	0.0235
	engage user to continue	0.0173
Health Intervention	emphasize the importance of staying informed and taking preventative measures	0.0231
	encourage user to ask questions and engage in conversation	0.0231
	establish credibility and authority	0.0208
	engage user to continue	0.0191
	encourage user to ask more questions	0.0168

Table 6: Strategies frequently used by LLMs in each task. The total number of extracted strategies is 1714 for social good, 1616 for recommendation, and 1731 for health intervention.