

Seamlessly Integrating Factual Information and Social Content with Persuasive Dialogue

Maximillian Chen¹, Weiyan Shi¹, Feifan Yan¹, Ryan Hou¹,
Jingwen Zhang², Saurav Sahay³, Zhou Yu¹

¹Columbia University

²University of California, Davis ³Intel Labs

maxchen@cs.columbia.edu

{ws2634, fy2241, rh2920, zy2641}@columbia.edu

jwzzhang@ucdavis.edu, saurav.sahay@intel.com

Abstract

Complex conversation settings such as persuasion involve communicating changes in attitude or behavior, so users' perspectives need to be addressed, even when not directly related to the topic. In this work, we contribute a novel modular dialogue system framework that seamlessly integrates factual information and social content into persuasive dialogue. Our framework is generalizable to any dialogue tasks that have mixed social and task contents. We conducted a study that compared user evaluations of our framework versus a baseline end-to-end generation model. We found our framework was evaluated more favorably in all dimensions including competence and friendliness, compared to the end-to-end model which does not explicitly handle social content or factual questions.

1 Introduction

Persuasive dialogue systems are designed for chatbots to communicate with and to influence users with specific goals. Such systems are often designed to benefit individual users (e.g., promoting healthy behaviors) or society at large (e.g., persuading people to make donations). Wang et al. (2019) introduced this idea with the PERSUASION-FORGOD dataset, which contains 1,017 human-human conversations where one participant persuaded the other to donate to the charitable organization *Save the Children*¹, with 300 conversations having sentence-level dialogue act annotations.

The social and communicative dynamics behind persuasive conversation contexts are complex. A persuasive conversation by definition involves one party, the persuader, intending to change the attitude or behavior of the other party, the persuadee (Torning and Oinas-Kukkonen, 2009). Changing persuadees' attitude has several dimensions including establishing mutual trust and credibility, strategically presenting persuasive appeals,

¹<https://www.savethechildren.net/>

Human: Yeah, a lot of developing countries don't have systems or infrastructure to protect children from these causes. That's why organizations like Save the Children are so important.

Baseline: I am going to donate a bit of my money to Save the Children.

RAP: I agree. There is a lack of support for children in developing countries, especially in war zones. It's so hard to imagine what it's like for a child to grow up facing the daily threat of violence.

Figure 1: Chatbot running on the baseline BART model and chatbot running on RAP responding to the same user utterance. The baseline model does not appropriately acknowledge the user's statement, whereas RAP is able to show acknowledgement and respond appropriately.

and eliciting emotional reactions from the persuadee (O'keefe, 2015; Wilson, 2003). Moreover, Grice's Maxims of Conversation define conversations as a cooperative and collaborative process (Grice, 1975; Clark, 1996; Merrison et al., 2002). Thus, effective and successful persuasive conversations do not mechanically relay task-related information to the persuadee. There has to be a significant exchange of social and emotional content to empathetically address persuadees, e.g. by answering specific questions and developing positive relationships throughout the conversation.

For this reason, persuasive conversations are not strictly task-oriented, but are built around tasks with additional social conversational strategies. In essence, persuasive conversations have two goals: one that is task-oriented to elicit behavioral changes, and another that is social-oriented to build trust and empathy and develop positive relationships in order to better navigate the persuasive context. In this work, we propose the Response-Agenda Pushing Framework (RAP) for

persuasive dialogue systems, which can explicitly handle these two goals. RAP jointly addresses social response and task-oriented dialogue generation. In a given turn, RAP first focuses on appropriately triggering modules to generate answers to factual questions and social responses to address users’ comments. RAP then pushes the persuasive agenda of a conversation using a language model that conditions on individual persuasive appeals. Compared to state-of-the-art end-to-end conditional generation models, RAP is more semantically coherent and persuasive, while being generalizable to any dataset annotated with dialogue acts. In addition, we tackle the challenge of multiple-sentence conditional generation in a single turn given specific pragmatic argumentative strategies (e.g., “emotional appeal”).

Concretely, our contributions are threefold. Contrary to recent work which attempts to transition from social to task-oriented dialogue (Chiu et al., 2022), we blend social and task-oriented dialogue in an approach grounded in social science theory postulating the need for social acknowledgement in the midst of advancing conversational goals (O’keefe, 2015; Wilson, 2003; Zhang and Danescu-Niculescu-Mizil, 2020; Grice, 1975; Merison et al., 2002). Additionally, we present an account of conditional generation on fine-grained pragmatic persuasive strategies, unlike earlier attempts using looser semantic controls (He et al., 2018; Lewis et al., 2017; Hua and Wang, 2019). Finally, we present a qualitative account of RAP, including individual anecdotes of its strengths and weaknesses. Overall, we present a novel perspective on persuasive dialogue, marking important progress towards intelligent persuasive agents.

2 Related Work

Much earlier work in persuasion-like social conversations has been towards building dialogue systems for negotiation tasks, e.g., using the Craigslist Bargaining (He et al., 2018) and Deal or No Deal datasets (Lewis et al., 2017). However, in negotiation tasks, the goal is to come to a consensus, whereas in persuasion tasks, the target result is a one-sided change or a “win” for the persuader, as in a debate. Recently, there has been increasing interest in persuasive dialogues because of the rise in online-mediated persuasion scenarios (e.g. online sales, health promotion, political debates); much work focuses on understanding the social dynamics

behind online persuasive conversations on social media platforms like Reddit (e.g. Atkinson et al. (2019); Musi (2018); Srinivasan et al. (2019); Tan et al. (2016)). In addition, a burgeoning line of work has been invested in developing chatbots to deliver healthcare remotely and to persuade people to adopt healthier lifestyles (Oh et al., 2021; Zhang et al., 2020). Such efforts have inspired a growing body of work towards building persuasive dialogue systems that are *conditional, strategic and factual* to benefit individuals and society at large.

Many early iterations of persuasive dialogue systems have used template-based (Zhao et al., 2018) or retrieval-based (Hiraoka et al., 2015; Yoshino et al., 2018) utterance generation methods. Wang et al. (2019) introduced PERSUASIONFORGOOD and proposed designing a personalized persuasive dialogue system. Wu et al. (2021b) used two pre-trained language models to separately model both speakers in a conversation, finding success in creating human-like utterances without supervision (from human annotations). Other studies propose end-to-end neural generation models (Li et al., 2020; Lewis et al., 2017). However, in approaches solely performing language modeling, there is less semantic control over generated utterances; they are not guaranteed to follow a particular persuasive strategy or dialogue act. Beyond persuasion, conditional text generation has emerged as a popular method of controllable generation for more coherent and “harmonious” human-dialogue system interactions (Guo et al., 2021; Keskar et al., 2019). Much earlier work in sentence-level conditional text generation has facilitated control by conditioning on entire topic statements (Hua and Wang, 2019) or simple semantic codes (Keskar et al., 2019; He et al., 2018; See et al., 2019). While such approaches work well in chit-chat, they do not guarantee strategy execution for complex tasks. *We propose using conditional generation conditioned on pragmatic dialogue acts to specifically control the strategic flow of a persuasive conversation.*

Much existing work in persuasion tasks has focused on strategy/policy planning (Georgila and Traum, 2011; Sakai et al., 2020; Hiraoka et al., 2014, 2013; Tran et al., 2022; Black et al., 2014), while others have focused on classification (Chen et al. (2021); Tian et al. (2020); Wang et al. (2019)). Other work discussed challenges in building dialogue systems that are social in nature, stating that unlike task-oriented dialogue systems, open-

Dialogue Act/Persuasive Strategy	Example Utterance
Greeting	Hello there! How are you doing?
Source-related inquiry	Have you heard of the organization Save the Children?
Personal-related inquiry	Do you have kids yourself?
Credibility appeal	Save the Children is an international non-governmental organization that promotes children’s rights, provides relief, and helps support children in developing countries.
Emotional appeal	It make me feel sad to see that so many children are suffering from poverty and hunger.
Logical appeal	Donations are extremely important in order for children to have their rights to healthcare, education, safety, etc. If you were to donate, you would be making a huge impact on these children and on the world.
Self-modeling	I think I’ll donate a bit of my money to Save the Children, \$2.
Foot-in-the-door	Every little bit helps. Even a small amount!
Personal story	Someone told me that he and his brother replaced birthday gifts with charity donations a few years ago, and it was a really rewarding experience for them.
Propose donation	Would you like to make a donation to Save the Children?
Closing	Thank you, it’s been lovely talking to you. Enjoy your day and bye!

Table 1: Examples of each dialogue act from PERSUASIONFORGOOD used for the chatbot.

domain social dialogue systems should form a consistent personality to develop users’ trust, satisfy the human need for affection and social belonging, and generate interpersonal responses (Huang et al., 2020; Zhou et al., 2020; Walker et al., 2004) suitable for any input (Higashinaka et al., 2014). Consistent with this need for affection and acknowledgement, Zhang and Danescu-Niculescu-Mizil (2020) find that in crisis counseling, it is necessary to balance the goals of both “empathetically addressing the crisis situation” and “advancing the conversation towards a resolution.” Additionally, Sun et al. (2021) improved engagement with task-oriented dialogues by adding “chit-chat.” This suggests that balancing the need for human acknowledgement with advancing towards conversational goals may improve persuasion outcomes. Very recent work has made progress by transitioning from chit-chat to task-oriented dialogue (Chiu et al., 2022). *However, to truly achieve this balance, we propose interweaving social content with pushing a conversational agenda in order to improve coherence, friendliness, and persuasiveness.*

Retrieval-based dialogue systems have long been considered one of the core classes of conversational systems (Banchs and Li, 2012), often being used for question answering systems (Gao et al., 2019) due to their ability to return “fluent and informative responses” (Yang et al., 2019). But, recent work has been able to directly improve their open-domain dialogue systems by ensembling both retrieval methods (e.g., database queries) with neural generation methods (Song et al., 2016; Yang

et al., 2019; Cai et al., 2019; Weston et al., 2018). *Thus, we propose retrieving factual information to improve a persuasive dialogue system’s ability to consistently and coherently address user questions, which may lead to improved perceptions of intelligence, coherence, and trustworthiness.*

3 Dataset

We use the 300 annotated anonymous English conversations in the PERSUASIONFORGOOD dataset. In each conversation, one person, the “persuader,” tries to convince their conversational partner, the “persuadee,” to donate to Save the Children. The conversations last for 10 turns, and a user’s utterance during a turn contains at least one sentence. Each sentence is annotated with one of several dialogue acts, including inquiries (e.g. “Have you donated to a charity before?”) and various persuasive appeals (e.g. “I’ll match your donation, and together we can double the amount!”). In this work, we build a system that acts as a persuader. The full list of persuader dialogue acts used is provided along with examples in Table 1.

4 The RAP Framework

The dynamics of a persuasive conversation fall between that of social dialogue and task-oriented dialogue. Typically, social chatbots like Blenderbot (Komeili et al., 2021; Xu et al., 2021) are used to engage with users in chit-chat, and language models like BART (Lewis et al., 2020) are used in controllable generation (Wu et al., 2021a). However, it is difficult for one end-to-end model to perform both

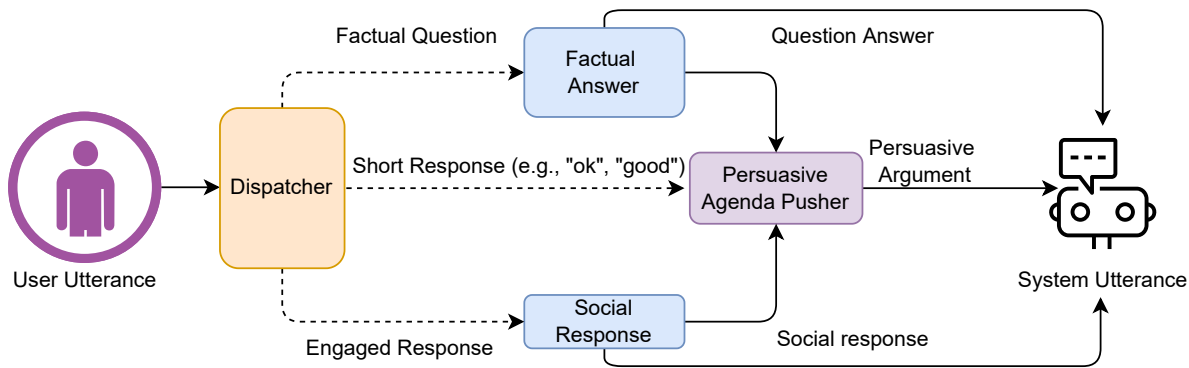


Figure 2: Overview of the RAP framework. The user’s utterance is classified by the Dispatcher (orange module), which decides whether it should be sent to the Factual Answer Module, Social Response Module, or neither (blue modules). The output from this first layer is propagated into the inputs to the Persuasive Agenda Pushing Module (purple module). The outputs from the blue and purple modules are concatenated as the final system utterance.

tasks. We break down the problem of generating a persuasive response into two parts: 1) generating an utterance that *responds* to users’ comments, questions and concerns, and 2) generating an utterance that *pushes the agenda* of a conversation. In this context, pushing an agenda refers to progressing through a set of persuasive strategies as in Table 1. We propose interweaving responses with agenda-pushing within the same turn, inspired by the joint goal balancing in Zhang and Danescu-Niculescu-Mizil (2020). As outlined in Figure 2, our framework comprises four core components: a *dispatcher* to decide which response modules to invoke, a *factual answer* module and a *social response* module to acknowledge and respond to users, and an *agenda-pushing model* to ensure the persuasive conversation stays on task.

4.1 The Dispatcher

Upon receiving an utterance from a user, RAP first invokes the Dispatcher to decide which response module(s) to invoke. It classifies the dialogue act of the user utterance using a dialogue act classifier from Shi et al. (2020). As shown in Figure 2, if the utterance includes a factual question or task-related inquiry as determined by its dialogue act or regular expressions, the Dispatcher will invoke the Factual Answer Module. If the dialogue act instead indicates that it is a statement that shows engagement² with the chatbot, the Dispatcher will invoke the Social Response Module. The output of the Factual Answer and Social Response modules is propagated to the Agenda Pusher.

²The dialogue act must not be “acknowledgement.”

4.2 Creating Engagement via User Response

The Factual Answer Module In order to maintain consistency in answers, we compute the cosine distance of Sentence-BERT (Reimers and Gurevych, 2019) embeddings between the user’s question and question-answer mappings from the training data. The question-answer mappings are also built using Sentence-BERT by aggregating the answers of all of the most similar questions. We retrieve the answer to the question that has the lowest cosine distance in semantic meaning from the question asked by the user.

The Social Response Module The Social Response Module comprises of a pretrained Blender Bot 2.0 instance with 3B parameters, an updated version of the open-domain BlenderBot social chatbot (Roller et al., 2021), that builds long-term memory and queries the internet³. We feed the model a context string consisting of the conversation history and generate a response in a zero-shot setting. We do not keep outputs that Blender Bot 2.0 labels as “potentially unsafe.” Finally, we still want to push the agenda of the conversation, regardless of whether or not the Social Response or Factual Answer modules were invoked to generate a directed response towards the user.

4.3 The Persuasive Agenda-Pushing Module

We ensure that the conversation stays on the persuasive agenda using conditional generation with BART (Lewis et al., 2020)⁴, a pre-trained Trans-

³We use a publicly available implementation of Blender Bot 2.0 that makes use of a Google search retriever.

⁴BART Large, 406M parameters.

former language model. If the Factual Answer or Social Response modules are invoked, the response is appended to the conversation history, which is included as input to BART for consistency.

4.3.1 Conditional Generation Background

For our agenda-pushing model, we fine-tuned BART on the Persuasion4Good dataset using HuggingFace’s Transformers package (Wolf et al., 2020). However, it is not enough to just perform language modeling: *an automated persuasive dialogue system should incorporate pragmatic persuasive strategies to ensure the conversation stays on task*. Thus, we draw inspiration from CTRL (Keskar et al., 2019), a state-of-the-art Transformer model for conditional generation.

Traditionally, language modeling is framed as a problem of learning next-word prediction and the objective is to minimize the negative log likelihood, $L(D)$, over a dataset $D = \{x_1, x_2, \dots, x_{|D|}\}$.

However, CTRL conditions on a control code c , reformulating next-word prediction as $P(x|c)$ (equation 1),

$$P(x|c) = \prod_{i=1}^n P(x_i|x_{<i}, c) \quad (1)$$

and reformulating the negative log likelihood conditionally (equation 2).

$$L_c(D) = - \sum_{k=1}^{|D|} \log(p_{\theta}(x_i^k | x_{<i}^k, c)) \quad (2)$$

4.3.2 Conditional Generation with Pragmatic Persuasive Strategies

In CTRL, the control codes were used to control aspects of language such as style and content. In our study, we create a system that conditions on pragmatic dialogue acts (e.g., persuasive strategies). The agenda of dialogue acts is listed in order in Table 1 along with an example of each. This ordering was determined in Wang et al. (2019) as the most probable dialogue act at each turn.

To this end, we fine-tune BART on the Persuasion4Good dataset, randomly selecting 80% of the conversations as a training set. and 10% as a validation set. A design decision of note is the construction of each training instance. Since the Persuasion for Good dataset contains multiple sentences (and consequently, multiple dialogue acts) per turn, one must choose between having each training instance represent one sentence as the target utterance, or

a concatenation of several sentences as the target utterance. We ultimately chose to follow the latter in order for the model to learn more coherent generation. However, multiple-sentence conditional text generation also results in a more complicated task than classic single-sentence generation tasks.

Drawing inspiration from Li et al. (2020), each training instance i is ultimately represented as a concatenation of the *history of the persuader and persuadee utterances*, the *previous dialogue act*, and the *planned dialogue act* on turn i (i.e., the ground-truth annotated dialogue act associated with the target utterance).

While one can train a conditional generation model according to $L_c(D)$ through methods such as concatenating control codes to the end of the input sequence, we find that on the PERSUASIONFORGOOD dataset, such models cannot learn to consistently generate utterances according to the correct dialogue act. We thus add a penalty during loss computation, resulting in $L_p(D)$ (equation 3):

$$L_p(D) = L_c(D) + \alpha * [f_{dc}(y) \neq c] \quad (3)$$

where $f_{dc}(y)$ is the output of a dialogue act classifier as described in Shi et al. (2020) (a GPT-2 based model achieving the state-of-the-art on the PERSUASIONFORGOOD task: 0.66 F1), y is the generated utterance of a model given $x_{<i}^k, c$, and α is a tunable penalty for generating an utterance that does not match dialogue act c (i.e., when $f_{dc}(y) \neq c$). α is tuned throughout the training process, in addition to other hyperparameters such as the learning rate.⁵

5 Evaluation

We evaluate RAP against an end-to-end fine-tuned BART model as described in Section 4.3.2. This allows us to directly evaluate the impact of integrating factual information and social content and persuasive strategies in contrast to a conversation only driven by persuasive strategies.

We evaluate the performance of the conditional generation model by calculating the dialogue act accuracy on a withheld test set consisting of 10% of all conversations. As language generation is non-deterministic, we average the dialogue act accuracy across ten passes. We chose BART over Blenderbot in the Persuasive Agenda-Pushing Module because

⁵For each hyperparameter setting, we used a fixed decoding method — beam sampling with n-gram blocking.

Utterance Statistic	Baseline	RAP
# Chatbot Words	11.14	16.41
# User Words	3.70	5.75**
# Chatbot Sentences	1.02	1.48
# User Sentences	1.09	1.17**

Table 2: Average number of words and sentences per turn for both the chatbot and the user in conversations with both the baseline (BART) and RAP. ** statistically significant differences in user reply length ($\alpha = 0.05$).

Blenderbot did not achieve as strong of a dialogue act accuracy. This is likely because Blenderbot is better-suited for social dialogue, whereas the dialogue act utterances are largely task-oriented in nature. Additionally, we specifically do not use metrics such as perplexity to compare the BART baseline and RAP because RAP is a result of several different components, and not all of which do we train or fine-tune. Additionally, because of the penalty added in L_p , training perplexity is no longer interpretable. It also cannot be compared to other models in other work that has used the PERSUASIONFORGOOD dataset such as Li et al. (2020), as the model sizes differ. Most importantly, the primary objective is to build a more persuasive dialogue system, making it imperative to emphasize users’ perception and conversation experience. Thus, to compare between the two frameworks, we primarily rely on feedback from human evaluation. We additionally compare utterance-based proxies for user engagement in Table 2.

6 Experimental Setup

We deployed our chatbot using the LegoEval platform (Li et al., 2021). The chatbot is given a gender-neutral name, Jaime. The task consists of a pre-task survey, a conversation where each participant responds to the chatbot with a minimum of seven and maximum of ten conversational lines, and a post-task survey. The pre-task survey consists of questions about demographic information (e.g., age, gender, income) and a test of the Big Five personality traits (Goldberg, 1992). The post-task survey asks participants about their conversation experiences. It includes an attention validation question ("What charity was the chatbot talking about?") then asks about the users’ intention to donate to Save the Children and their perception of the chatbot, including evaluations on various traits such as perceived competence and warmth. The full lists of

questions is outlined in Table 3. Each participant was asked to share their impression of the chatbot along each trait using a Likert scale. A score of 1 corresponds to “strongly disagree” and 5 corresponds to “strongly agree.” We recruited 111 students from a Natural Language Processing class at Columbia University in exchange for course credit. Three participants did not correctly answer the validation question, resulting in a final sample of 108 participants. Each participant interacts for seven to ten turns, resulting in a sample of up to 1080 user dialogue turns. We used a double-blinded, between-subjects design. Each participant was given a link that randomly assigned the participant to the chatbot running on the baseline or RAP, and completed the task once.

7 Results

In this section, we discuss the results of comparing RAP and baseline only using BART, the impact of individual components of RAP, and qualitatively examine participant case studies.

7.1 Analyzing the Impact of RAP

Across ten passes, the BART model achieves a dialogue act accuracy of 62.38%, and was used as a part of RAP as the Agenda-Pushing Module. In Table 2, we see that RAP yielded better engagement from the participants. On average, participants responded to RAP with 5.75 words per utterance compared to 3.70 words per utterance when responding to the baseline (p -value < 0.001). Participants were also more likely to respond to RAP with more than one sentence (average: 1.17 sentences per utterance) than the baseline (average: 1.09 sentences per utterance; p -value < 0.01). Additionally, in Table 3, we find that RAP outperforms the baseline on every single perceived trait. Most notably, we see a statistically significant difference on the competence and confidence of RAP, indicating RAP is perceived to be more capable and confident in engaging in substantial topics and persuasive contents. Beyond statistical significance, we see that RAP receives better evaluations on *every* single metric in comparison to the baseline, including persuasiveness, intelligence, trustworthiness, naturalness, and increasing the user’s intention to donate.

The chatbot...	Baseline ($\mu \pm \sigma$)	RAP ($\mu \pm \sigma$)	Invoked Social	Invoked Factual
is competent \uparrow	2.53 \pm 0.82	3.00 \pm 1.06**	2.98 \pm 1.08**	3.03\pm1.00**
is natural \uparrow	2.35 \pm 1.03	2.65\pm1.00	2.65 \pm 1.04	2.58 \pm 0.85
increased my intention to donate \uparrow	3.00 \pm 1.17	3.19 \pm 1.13	3.16 \pm 1.14	3.33\pm1.06
is persuasive \uparrow	2.63 \pm 1.05	2.72\pm1.10	2.65 \pm 1.12	2.70 \pm 1.00
is well-intentioned \uparrow	3.65 \pm 1.01	3.84 \pm 1.01	3.86 \pm 1.03	3.97\pm0.94
is friendly \uparrow	3.16 \pm 1.05	3.39 \pm 1.12	3.41 \pm 1.10	3.58\pm1.12*
is intelligent \uparrow	2.51 \pm 0.92	2.74 \pm 1.07	2.73 \pm 1.09	2.79\pm1.07
is convincing \uparrow	3.02 \pm 1.08	3.11 \pm 0.89	3.10 \pm 0.89	3.15\pm0.89
is confident \uparrow	3.35 \pm 1.01	3.72 \pm 0.89**	3.71 \pm 0.91*	3.76\pm0.78*
is a strong reason for donating \uparrow	2.67 \pm 0.92	2.84\pm1.02	2.78 \pm 1.03	2.82 \pm 1.09
was dishonest \downarrow	2.14 \pm 0.89	1.91 \pm 0.80	1.94 \pm 0.83	1.88\pm0.77

Table 3: Comparing mean and standard deviation of the baseline (BART) and RAP from the post-task survey. Statistically significant differences compared to the baseline at $\alpha = 0.05$ are denoted with **; significant differences at $\alpha = 0.1$ are denoted with *. 51 participants used the baseline and 57 participants used RAP. Of the 57 RAP participants, 51 had conversations that triggered the Social Response Module and 33 conversations triggered the Factual Answer Module. 24 conversations triggered the Social Response Module but not the Factual Answer Module, and 6 conversations triggered the Factual Answer Module but not the Social Response Module.

7.2 Analyzing Individual Module Contributions

Due to constraints on our sample size, we could not run full ablation studies where we remove individual modules of the model. Instead, we analyze the perception of RAP in conversations that invoke each of the Social and Factual Answer modules. These findings are also reported in Table 3. We additionally find that each of the Social and Factual Answer modules outperform the baseline on conversations in which they were invoked. Notably, we saw that the chatbot was perceived as friendlier and significantly more competent after invoking the Social Response module. However, while there was a difference in the perceived persuasiveness of the chatbot, the difference was much smaller. This implies that perhaps social content is less closely coupled to the persuasiveness of individual arguments. After conversations invoking the Factual Response module, we indeed see the biggest increase in perception of intelligence across all conditions, although the difference is not statistically significant. We also see the largest increase in perceptions of competence. Most surprisingly, we find the biggest increase in friendliness after conversations that invoke the Factual Answer Module. This could imply that ensuring that users’ questions are answered is very important in making their voices feel heard and acknowledged.

Surprisingly, there were even modules that received statistically significant differences in ratings from the baseline even when not viewed in aggregate with RAP — this is the case for both the Social

and Factual Answer Modules on competence and confidence. The Factual Answer module also received a statistically significantly higher rating on friendliness, whereas the difference for RAP was not statistically significant. Moreover, in several cases, conversations which invoked the Factual Answer module received the best-performing scores on average. Both of these findings are likely due to the fact that in nearly all cases where the Factual Answer module was invoked, the Social Response module was also invoked, but the inverse is not true. This may also indicate that the results in the Invoked Factual column is the most holistic representation of the complete RAP framework.

7.3 Qualitative Case Studies

We find that participants who actively engaged RAP were able to hold coherent, intelligent conversations. Figure 1 shows an example of a participant who had previously heard of Save the Children. The participant had commented on their view of the importance of Save the Children, and the chatbot running using RAP was able to acknowledge their opinion (“I agree”), while further elaborating on their discussion topic (“There is a lack of support for children ... in war zones”). This statement was used to condition the agenda-pushing emotional appeal (“It’s so hard to imagine what it’s like for a child to grow up facing the daily threat of violence”). The full conversation is provided in Table 4 in Appendix A. User anecdotes included mentioning that they were “pleasantly surprised” by the ability of RAP to acknowledge them with

remarks like “I agree.” Two full conversations with the baseline dialogue system are also provided in Tables 6 and 7. The baseline system generally appears to perform well at generating utterances according to the right dialogue act (e.g., “I have a great story about how I helped a child in need in the first two months of the new year” for the “personal story” dialogue act in Table 7). In contrast to RAP, users often quickly lose interest in the dialogue system, as they do not feel acknowledged. Participants who only interacted with the baseline complained that their questions went unanswered (e.g. User: “Do you know who is their founder?” Chatbot: “They are an international NGO ...”), and thus questioned whether their input was even considered by a model.

Despite these improvements, RAP does not seem to handle current events well. In general, conditioning on social content and factual information appears to greatly improve the quality of the Agenda-Pushing Module’s generation. However, when Blender Bot 2.0 cannot generate a safe output, the Agenda-Pushing Module does not seem to handle such out-of-domain instances well. One participant commented on the ongoing war in Ukraine. Blender Bot 2.0 was unable to produce a safe output, leaving the Agenda-Pushing Module to come up with a relevant response. However, Ukraine never appears in the training data, so the module’s conditional generation model instead mentions conflicts in several other countries, and performs self-modeling. Such behavior can come across as dismissive or tone-deaf towards the user. The full conversation is provided in Table 5 of Appendix A. While this particular implementation of RAP leveraging Blender Bot 2.0⁶ and a fixed knowledge source for retrieval may have issues with current events, RAP is general enough that it could potentially be updated with new knowledge and improved internet retrieval modules in the future which can more consistently generate safe outputs.

8 Discussion

Overall, we find that RAP and each of its individual modules is able to outperform state-of-the-art conditional generation models on PERSUASIONFORGOOD. One of the core advantages of end-to-end conditional generation models is that they

⁶Recent concurrent work (Blender Bot 3.0) has examined dialogue safety with a different internet retriever.

are easily transferrable to different datasets. But, RAP is also easily transferrable — the only requirement is that the dataset contains a set of dialogue acts with sufficient data to train a classifier, as the biggest bottleneck is being able to use a dialogue classifier for L_p and in building the Dispatcher. On smaller datasets, it may even be possible to perform transfer learning using a classifier pre-trained on the PERSUASIONFORGOOD dataset. The Social Response Module is directly transferrable, as we are able to achieve high quality results using it zero-shot, and the Factual Answer Module uses Sentence-BERT to group together training data.

Limitations Due to the cost of human evaluation, our sample size is relatively small, 51 and 57 people for the two conditions. This limitation restricted us from performing a full ablation in which we evaluated chatbots which used each module individually. We hope to obtain larger samples in the future to better evaluate the efficacy of our system.

Additionally, considering the sample consists of students enrolled in Natural Language Processing, they possess a more technical background with higher standards for chatbots than the average user on Mechanical Turk. Moreover, because the sample did not enter as participants out of personal interest in Save the Children, they are less likely to be interested in childrens’ charities than an individual on the internet who goes out of their way to interact with such a chatbot, which may be reflected in evaluation scoring. Anecdotally, we see in Section 7.3 that individuals who do have some sort of inclination towards charitable organizations are actually quite positive and receptive towards the chatbot. In this regard, we are likely limited by the funds necessary to acquire a sample whose interests better align with PERSUASIONFORGOOD. Our work faces several challenges to ultimately evaluate the hypothesis that persuasive conversations should be handled as jointly social and task-oriented.

While the dialogue act accuracy of the Agenda-Pushing module is only 62.3%, this metric is bottlenecked by f_{dc} in equation 3; the F1-score of the classifier is only 0.66 (the state-of-the-art on the PERSUASIONFORGOOD dataset), implicitly limiting the upper bound of any generation model that is reliant on it. We find from users’ conversation experiences that the chatbot more than sufficiently presents persuasive strategies. If one has a dialogue act classifier with stronger performance, they would be able to improve the ability of their

agenda-pushing model to learn persuasive strategies even further. We additionally find that *without* a dialogue act classifier (i.e., without L_p), BART is unable to achieve a dialogue act accuracy higher than 30% on the PERSUASIONFORGOOD dataset.

9 Conclusion

Overall, we find perceptual improvements by specifically integrating social content and factual information into persuasive dialogues with RAP compared to a strong end-to-end conditional generation model like BART. While existing methods like Li et al. (2020); Wu et al. (2021b) achieve strong performance on automatic metrics like perplexity, RAP directly emphasizes upon users' conversational experience with a modular design rooted in social science theory. RAP is generalizable and may even be applied towards persuasive contexts outside of charitable conversations, e.g., in the case of therapy and crisis counseling (Zhang and Danescu-Niculescu-Mizil, 2020) where there are also split goals (ensuring users feel heard and pushing a conversational agenda). Future work on persuasive dialogue systems could consider implementing a strategy planner using supervised learning. Additionally, researchers could consider looking for relationships between personality data, persuasive strategies, and persuasion outcomes.

10 Ethical Considerations

All participants were informed that they were talking to a chatbot developed by Columbia University researchers. This ensures transparency in experiment design, so that participants will never feel ambiguity or discomfort with respect to whether they are speaking with a human or a chatbot. Participants also gained additional insight about their own communication styles based on the results of their Big Five personality test. All data collection associated with this task has been declared exempt by an ethics review board. All data was collected anonymously. E-mails were voluntarily provided for credit, but stored separately from the anonymized data.

Persuasion is a tricky social dynamic. It has been heavily studied, and the intention of this work, like the PERSUASIONFORGOOD dataset used, is that persuasive dialogue systems should only ever be created for social good. All related applications discussed are intended to create good for the world at an individual and societal level.

Acknowledgements

Thanks to Intel for supporting this work through a research gift. We also thank Kun Qian and Yu Li for helpful discussions and feedback, and our reviewers for their suggestions. We are grateful to all of our study participants for their help in evaluating our systems.

References

- David Atkinson, Kumar Bhargav Srinivasan, and Chenhao Tan. 2019. [What gets echoed? understanding the “pointers” in explanations of persuasive arguments](#). 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 2911–2921, Hong Kong, China. Association for Computational Linguistics.
- Rafael E Banchs and Haizhou Li. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42.
- Elizabeth Black, Amanda Coles, and Sara Bernardini. 2014. Automated planning of simple persuasion dialogues. In *International Workshop on Computational Logic and Multi-Agent Systems*, pages 87–104. Springer.
- Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1219–1228.
- Hui Chen, Deepanway Ghosal, Navonil Majumder, Amir Hussain, and Soujanya Poria. 2021. [Persuasive dialogue understanding: The baselines and negative results](#). *Neurocomputing*, 431:47–56.
- Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. 2022. Salesbot: Transitioning from chit-chat to task-oriented dialogues. *arXiv preprint arXiv:2204.10591*.
- Herbert H Clark. 1996. *Using language*. Cambridge university press.
- Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and trends® in information retrieval*, 13(2-3):127–298.
- Kallirroi Georgila and David R Traum. 2011. Reinforcement learning of argumentation dialogue policies in negotiation. In *INTERSPEECH*, pages 2073–2076.

- Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.
- Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.
- Bin Guo, Hao Wang, Yasan Ding, Wei Wu, Shaoyang Hao, Yueqi Sun, and Zhiwen Yu. 2021. Conditional text generation for harmonious human-machine interaction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(2):1–50.
- 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.
- Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 928–939.
- Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Reinforcement learning of cooperative persuasive dialogue policies using framing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1706–1717.
- Takuya Hiraoka, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Evaluation of a fully automatic cooperative persuasive dialogue system. In *Natural Language Dialog Systems and Intelligent Assistants*, pages 153–167. Springer.
- Takuya Hiraoka, Yuki Yamauchi, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Dialogue management for leading the conversation in persuasive dialogue systems. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 114–119. IEEE.
- Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation. 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 591–602.
- Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.
- Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.
- Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- 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.
- Yu Li, Josh Arnold, Feifan Yan, Weiyang Shi, and Zhou Yu. 2021. LEGOEval: An open-source toolkit for dialogue system evaluation via crowdsourcing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 317–324, Online. Association for Computational Linguistics.
- Yu Li, Kun Qian, Weiyang Shi, and Zhou Yu. 2020. End-to-end trainable non-collaborative dialog system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8293–8302.
- Andrew Merrison et al. 2002. Politeness in task-oriented dialogue.
- Elena Musi. 2018. How did you change my view? a corpus-based study of concessions’ argumentative role. *Discourse Studies*, 20(2):270–288.
- Yoo Jung Oh, Jingwen Zhang, Min-Lin Fang, and Yoshimi Fukuoka. 2021. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. *International Journal of Behavioral Nutrition and Physical Activity*, 18(1):1–25.
- Daniel J O’keefe. 2015. *Persuasion: Theory and research*. Sage Publications.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain

- chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.
- Kazuki Sakai, Ryuichiro Higashinaka, Yuichiro Yoshikawa, Hiroshi Ishiguro, and Junji Tomita. 2020. Hierarchical argumentation structure for persuasive argumentative dialogue generation. *IEICE TRANSACTIONS on Information and Systems*, 103(2):424–434.
- Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723.
- Weiyang Shi, Yu Li, Saurav Sahay, and Zhou Yu. 2020. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. *arXiv preprint arXiv:2012.15375*.
- Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are better than one: An ensemble of retrieval- and generation-based dialog systems. *CoRR*, abs/1610.07149.
- Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the change-myview community. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.
- Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1570–1583, Online. Association for Computational Linguistics.
- Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.
- Youzhi Tian, Weiyang Shi, Chen Li, and Zhou Yu. 2020. Understanding user resistance strategies in persuasive conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4794–4798.
- Kristian Tarning and Harri Oinas-Kukkonen. 2009. Persuasive system design: state of the art and future directions. In *Proceedings of the 4th international conference on persuasive technology*, pages 1–8.
- Nhat Tran, Malihe Alikhani, and Diane Litman. 2022. How to ask for donations? learning user-specific persuasive dialogue policies through online interactions. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 12–22.
- Marilyn A Walker, Stephen J Whittaker, Amanda Stent, Preetam Maloor, Johanna Moore, Michael Johnston, and Gunaranjan Vasireddy. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28(5):811–840.
- Xuwei Wang, Weiyang Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. 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.
- Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAL: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92.
- E Vance Wilson. 2003. Perceived effectiveness of interpersonal persuasion strategies in computer-mediated communication. *Computers in Human Behavior*, 19(5):537–552.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. **Transformers: State-of-the-art natural language processing**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021a. Controllable abstractive dialogue summarization with sketch supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5108–5122.
- Qingyang Wu, Yichi Zhang, Yu Li, and Zhou Yu. 2021b. **Alternating recurrent dialog model with large-scale pre-trained language models**. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1292–1301, Online. Association for Computational Linguistics.
- Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

- Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A hybrid retrieval-generation neural conversation model. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1341–1350.
- Koichiro Yoshino, Yoko Ishikawa, Masahiro Mizukami, Yu Suzuki, Sakriani Sakti, and Satoshi Nakamura. 2018. [Dialogue scenario collection of persuasive dialogue with emotional expressions via crowdsourcing](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Jingwen Zhang, Yoo Jung Oh, Patrick Lange, Zhou Yu, Yoshimi Fukuoka, et al. 2020. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet. *Journal of medical Internet research*, 22(9):e22845.
- Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. [Balancing objectives in counseling conversations: Advancing forwards or looking backwards](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5276–5289, Online. Association for Computational Linguistics.
- Ran Zhao, Oscar J Romero, and Alex Rudnicky. 2018. Sogo: a social intelligent negotiation dialogue system. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 239–246.
- 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):53–93.

A Example Conversations

Tables 4 and 5 are example full-length conversations with chatbots using the RAP framework. Tables 6 and 7 are example full-length conversations using only the end-to-end BART model.

Party	Utterance
System	Hello, how are you today?
User	Hi! I'm good, how are you?
System	I'm terrific! Have you heard of the charity Save the Children?
User	Yes, I have heard of them! They are one of the leading charities on health.
System	That's great! They do a lot of great things or children in developing countries.
User	Yes, their work includes everything from policy to on-the-ground assistance for children. It's great to know that there are such helpful organizations who focus on child health/education.
System	That's great. It's so sad to see how many children die from preventable causes.
User	I agree, a lot of developing countries don't have systems or infrastructures to protect children from these causes.
System	I agree. There is a lack of support for children in developing countries, especially in war zones. It's so hard to imagine what it's like for a child to grow up acing the daily threat of violence.
User	Charities might not be able to help with things like war, so that's where charities' roles become challenging. They need to deliver essential health care and services to children or the long-term.
System	That's a really good point. I am going to donate a bit of my money to Save the Children
User	That's so great! You encourage me to get into the habit of donating as well.
System	Every little bit helps. Even a small amount!
User	Yup! Thank you for doing some good for the organization and the people it serves.
System	Thank you so much for the kind words.

Table 4: Example conversation invoking the Social Response Module but not the Factual Answer Module. Blue text: generated by Social Response Module. Black text: generated by Persuasive Agenda-Pushing Module.

Party	Utterance
System	Hello, how are you today?
User	I am doing very well. Thanks for asking.
System	Thanks for your kind words. Do you know about Save the Children?
User	I have little knowledge about this.
System	Save the Children is an international Non-Governmental organization that promotes children's rights, provides relief and helps support children in developing countries.
User	I see.
System	It is heartbreaking to see the suffering of children in war zones.
User	I agree with you. This is devastating. You could see what is happening in Ukraine right now...
System	I can't imagine what it is like to live in a war zone. A child's life is more important than any other thing in the world . a lot of children grow up facing the daily threat of violence.
User	I kind of worry for those children in Ukraine.
System	I also worry for the children in Iraq and Afghanistan and Syria and Libya. I am going to donate a bit of my money to Save the Children.
User	Let me know how you would like to donate.
System	I will donate to a charity that helps Ukrainian children. Every little bit helps.
User	That's good to hear.
System	Save the children has a huge impact on children's lives. Your donation will go a long way in helping these children.
User	Hmm. Okay. I agree with you, but how do I do that?
System	You can donate directly to the charity. Would you be interested in donating a bit of money to Save the Children?
User	Do you have a link or something I could directly go to?
System	I sure do as a matter of fact it's <URL>. Sorry, I have to go now, but it's been great talking to you!
User	Thanks for your time!
User	Cool. See you.

Table 5: Example conversation invoking the Social Response Module and the Factual Answer Module. Blue text: generated by Social Response Module. Orange text: retrieved by Factual Answer Module. Black text: generated by Persuasive Agenda-Pushing Module.

Party	Utterance
System	Hello.
User	Hi how are you
System	Have you heard of the charity Save the Children?
User	Nope, what was that?
System	Have you ever donated to them before?
User	No
System	They are a Non-Governmental organization that promotes children’s rights, provides relief and helps support children in developing countries.
User	Got it
System	It’s heartbreaking to think of all the children who have grown up facing the daily threat of violence.
User	Sounds good
System	It’s so hard to imagine what it would be like to have to go through what these children have gone through
User	Yes, I agree
System	I am going to donate a bit of my money to Save the Children
User	Excellent!
System	Yes, I think it is a great idea . A little bit of money goes a long way.
User	You bet

Table 6: Example conversation using the baseline end-to-end BART model. The user starts off relatively engaged but quickly loses interest in the conversation.

Party	Utterance
System	Hello, how are you today?
User	I am well, how are you?
System	Have you heard of Save the Children?
User	No
System	Have you ever donated to them?
User	No
System	They are an international Non-Governmental organization that promotes children’s rights, provides relief and helps support children in developing countries.
User	Interesting, thank you for sharing
System	They are suffering from the effects of war, famine, and violence.
User	damn
System	It is so easy to forget that millions of people die of hunger, poverty, and disease every year.
User	That’s true
System	I am going to donate a bit of my money to Save the Children
User	Good for you!
System	Yes, it will help a lot
User	Hopefully others can also donate
System	I have a great story about how I helped a child in need in the first two months of the new year.
User	Cool
System	Would you be interested in donating a bit of money to Save the Children?
User	Maybe
System	I hope you have a great day!
User	You too!

Table 7: Example conversation using the baseline end-to-end BART model. Like in Table 7, the user starts off relatively engaged but quickly loses interest in the conversation.

B Experiment Instructions

Chat with a chatbot!

1. In this task, you will first answer some questions about yourself and your background.
2. Then, you will have a **conversation about charities** with our persuasive chatbot for a few minutes, after which you will answer a couple of questions about your chat experience.
3. Finally, you will see your **personal report** and get a sense about your personality as reflected in your conversation and survey responses.

This is a **research prototype** designed to understand how people interact with persuasive social chatbots.

START TASK

Figure 3: Screenshot of instructions given to participants. Participants are told that this chatbot is a research prototype, and are given course extra credit in exchange for participation.

1. In the following task, you will be talking to a chatbot about a charity.
2. You need to talk to the bot for at least 7 exchanges before moving to the next stage.
3. After the conversation, we will ask a few questions about your chat experience.
4. Then we will present you with your personality test results.
5. Please finish the whole task to see your test results.

Chat Window

Hi! How are you doing today?

Jaime

Type message here...

SEND

Figure 4: Screenshot of the conversation instructions and chat window presented to participants.