Controllable Mixed-Initiative Dialogue Generation through Prompting

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, Zhou Yu
Columbia University
maxchen@cs.columbia.edu
{xy2437, ws2634, urvi.awasthi, zy2461}@columbia.edu

Abstract

Mixed-initiative dialogue tasks involve repeated exchanges of information and conversational control. Conversational agents gain control by generating responses that follow particular dialogue intents or strategies, prescribed by a policy planner. The standard approach has been fine-tuning pre-trained language models to perform generation conditioned on these intents. However, these supervised generation models are limited by the cost and quality of data annotation. We instead prompt large language models as a drop-in replacement to fine-tuning on conditional generation. We formalize prompt construction for controllable mixed-initiative dialogue. Our findings show improvements over fine-tuning and ground truth responses according to human evaluation and automatic metrics for two tasks: PersuasionForGood and Emotional Support Conversations.

1 Introduction

Mixed initiative dialogue systems allow all interacting agents to initiate actions to control the interaction. These systems dynamically adapt interaction styles to regain control and progress towards specific goals (Allen et al., 1999; Chu-Carroll, 2000), unlike others which passively respond to users' input (e.g. some assistants like ChatGPT),

Mixed initiative dialogue systems thus often involve complex policy planning sub-tasks to determine optimal turn-level system dialogue intents (Peng et al., 2018; Hiraoka et al., 2013; Muise et al., 2019; Liu et al., 2020). These policies define when it is optimal for a system to regain initiative (e.g., when a moderator should interject in a conversation, or when a companion should ask questions or change a conversation topic).

However, "optimal" planned dialogue intents still need to be executed through "optimal" response models. The standard practice in recent dialogue research has been to fine-tune a pre-trained language model for conditional generation

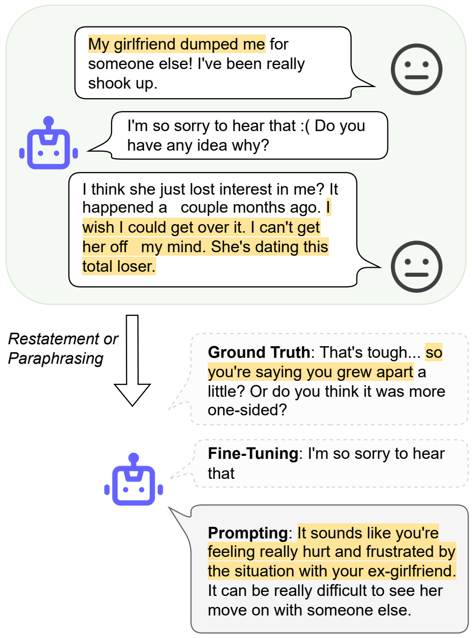

Figure 1: Excerpt of a conversation between an emotional help-seeker and a supporter about a breakup, with candidate responses attempting to use the support strategy "Restatement or Paraphrasing."

to achieve semantic control through some combination of innovations in model architectures or learning processes (Liu et al., 2021; Chen et al., 2019). Such generation approaches still leave room for error. Assuming that there exists a truly optimal dialogue policy planner, a response model may still generate according to the wrong intent (partially due to the fact that dialogue datasets often have annotation errors (Qian et al., 2021; Zang et al., 2020)). Or, a model may learn to generate correct intents but fail to create a response consistent with conversational context (Chen et al., 2022b). Additionally, training corpora often differ in demographic and distribution compared to production environments, which can lead to deteriorating response quality (Koh et al., 2021).

We propose using vanilla large pre-trained language models (LLMs) such as GPT-3 (Brown et al.,

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 951–966

2020) as drop-in replacements to traditional fine-tuned conditional generation models for mixed-initiative dialogue systems. LLMs typically have been trained on massive corpora with large amounts of linguistic variety, making them more robust to overfitting specific tasks. Recent work demonstrates that LLMs have reasonable semantic control through few-shot prompting (Brown et al., 2020; Chen et al., 2023; Meng et al., 2022). Here, we demonstrate how[1] to systematically prompt LLMs for mixed-initiative dialogue generation. Evaluations yielded strong performance on two popular English mixed-initiative tasks: Emotional Support Conversations (ESC; Liu et al. (2021)) and PersuasionForGood (P4G; Wang et al. (2019b)).

2 Related Work

Controllable Generation approaches often involve fine-tuning a model conditioned on control codes (Keskar et al., 2019; Ficler and Goldberg, 2017), additional attribute representations in hidden states (Hoang et al., 2016; Fu et al., 2018) or latent variables (Bowman et al., 2016; Wang et al., 2019a). Other work has attempted to mitigate the computational cost of fine-tuning, e.g. by training an auxiliary networks to guide the original LM (Dathathri et al., 2020; Yu et al., 2021; Pascual et al., 2021). Here, we attempt controllable generation that replaces fine-tuning by prompting LLMs.

Prompting in Dialogue Research typically has focused on understanding tasks such as dialogue planning (Kuo and Chen, 2022) or state tracking (Lee et al., 2021; Mi et al., 2022). More recent dialogue research has examined using prompting for generating conversational data with varying levels of control (Kim et al., 2022; Chen et al., 2022a; Mehri et al., 2022; Chen et al., 2023), citing the difficulty of using vanilla language models in production. Studies focusing on response generation looked at prompting LLMs specifically for knowledge-grounded dialogue generation (Liu et al., 2022; Madotto et al., 2021; Shuster et al., 2022). Our work is the first to construct an interactive prompt-based mixed initiative dialogue system and evaluate the semantic control of prompting.

[1]Code to reconstruct all prompts available at https://github.com/maxlchen/Controllable-Mixed-Initiative-Dialogue-Generation

3 Datasets

We examined ESC (Liu et al., 2021)) and P4G (Wang et al., 2019b). ESC consists of 1053 conversations between emotional help-seekers and supporters. Each conversation is annotated with the help-seeker's description of their problem, and the type of issues they are facing. Each turn by the supporters is annotated with one of eight emotional support strategies (Table A1). P4G contains 300 annotated conversations between persuaders who attempt to persuade persuadees to donate to a charity called Save the Children. Persuader turns are annotated with one of 10 strategies (Table A2).

4 Baselines

In mixed-initiative dialogue, interacting parties continuously exchange control throughout the conversation. However, in order for agents to regain control, they must be able to properly execute items from their conversational agenda, e.g. generating a response that matches a desired strategy/intent.

Liu et al. (2021) fine-tuned BlenderBot (Roller et al., 2021) on ESC using input representations consisting of flattened dialogue history and the predicted emotional support strategy for a specific turn. The best-performing model in their experimental setting is "Oracle-BlenderBot" which conditions on the ground truth strategy for a given turn.

Chen et al. (2022b) proposed a persuasive dialogue system called RAP, which combined targeted user response with conditional generation. The conditional generation component of RAP involves fine-tuning BART (Lewis et al., 2020) using a penalized loss to force the model to artificially create semantic control through dialogue intents.

5 Mixed-Initative Dialogue Prompting

RAP required introducing a dialogue intent classifier to weakly supervise the training process, as there is not an oracle for whether the dialogue intent of a candidate response is correct. But, this confounds errors, as classifiers are imperfect. Moreover, fine-tuning approaches like both RAP and Oracle-BlenderBot involve balancing a tradeoff between response quality and semantic control accuracy. Prompting LLMs avoids both issues as it does not involve adjusting model weights to learn representations of control codes for individual tasks.

In this paper, we systematically prompt Instruct-GPT "text-davinci-003." Rather than requiring expert-level prompt engineering, we create general prompt templates which directly fill slots using roles and annotations from both ESC and P4G. Specifically, we split up prompt construction into *Task Background* and *Conversation History*.

Figure 2 breaks down an example of a prompt for ESC. The Task Background is a paragraph formed from the "emotion type," "problem type," and "situation" annotations provided by the corpus. The Conversation History consists of each prior utterance, prepended by labels for each speaker. The system-side turns are also prefixed by a natural language form of the annotated emotional support strategy, derived from the annotation scheme in Liu et al. (2021) (e.g. "The Therapist acknowledges the Patient's feelings by paraphrasing their situation."). Figure 2 contains the contextual dialogue turns in order, along with the three support strategies used.

The P4G prompting style is similar. Unlike personalized emotional support conversations, the task does not change, so the Task Background is fixed with relevant factual background information. The Conversation History still interweaves narrative directions for each persuasive strategy (e.g. "The Persuader uses a logical appeal."). Example provided in Figure A1. The natural language intent mappings for both tasks are provided in Tables A1,A2.

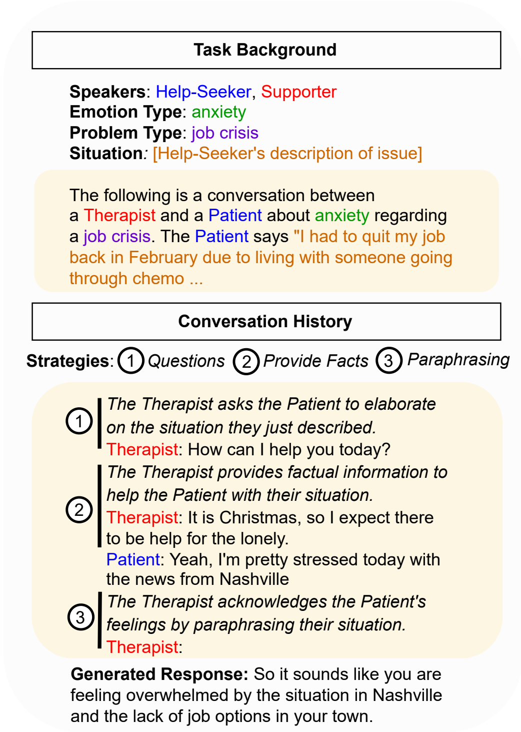

Figure 2: Parts of an example prompt for ESC (yellow background). Task Background: ground truth annotations describing the conversation. Conversation History: dialogue context with natural language forms of annotated dialogue intents. Full situation in Appendix B.1.

6 Experiments

We evaluated prompting statically and interactively.

6.1 Static Evaluation

We quantified how much semantic and pragmatic control vanilla LLMs can provide in conversation. We randomly sampled 100 responses from ESC (supporters) and P4G (persuaders). Each response's conversational history and strategy annotation was used to generate responses via prompting and fine-tuned models. We used Oracle-BlenderBot for ESC and RAP's conditional generation module for P4G.

We asked crowdworkers on Amazon Mechanical Turk[2] to evaluate candidate responses' accuracy with respect to its prescribed dialogue intents, coherence, consistency, and engagingness. We paired the dialogue responses from each source (fine-tuning, prompting, or ground truth) with the corresponding responses from each of the other sources, allowing us to compute preference winrates between each pair. Each job presented only one pair of responses, in a random order. Additionally, we examined automatic metrics through Distinct-N ($N \in \{3, 4\}$), as well QuantiDCE (Ye et al., 2021), a BERT-based automatic dialogue coherence metric for open-domain conversation.

Table 1 shows that prompt-generated responses are more highly rated in terms of quality compared to responses generated from competitive fine-tuned dialogue models *as well as ground truth responses*, in terms of all human evaluation metrics. This is also the case for Distinct-N in both tasks, and QuantiDCE in P4G. Oracle-BlenderBot slightly outperforms the prompt-generated responses in terms of QuantiDCE for ESC, but this difference is not statistically significant. Table 1 also shows that the prompt-generated responses are consistently preferable to the responses generated from fine-tuned dialogue models as well as the ground truth.

Finally, we also see that prompting appears to provide the best semantic control over generated responses. Prompt-generated responses had the highest probability of matching the desired dialogue

[2]Details for all human evaluation tasks in Appendix A.

Corpus	Metric	FT	GT	Prompt
ESC	Accuracy	0.81	0.85	**0.88***
	Coherence	3.57	3.57	**3.72**
	Consistency	3.63	3.60	**3.80**$^{+*}$
	Engagingness	3.55	3.61	**3.81**$^{+*}$
	Distinct-3	0.89	0.90	**0.90**
	Distinct-4	0.87	0.90*	**0.91**$^{+*}$
	QuantiDCE	**3.25**	3.03	3.19
	Win Rates			
	v. FT		**0.56**	0.52
	v. GT	0.44		**0.64***
	v. Prompt	0.48	0.36	
P4G	Accuracy	0.88	0.83	**0.89**
	Coherence	3.66	3.58	**3.83**$^{+*}$
	Consistency	3.69	3.56	**3.71**$^{+}$
	Engagingness	3.62	3.52	**3.69**$^{+}$
	Distinct-3	0.87	0.88	**0.89**
	Distinct-4	0.88	0.88	**0.88**
	QuantiDCE	3.16	3.09	**3.24**$^{+}$
	Win Rates			
	v. FT		0.56	**0.59***
	v. GT	0.48		**0.55**
	v. Prompt	0.41	0.45	

Table 1: Evaluation of response quality and semantic control accuracy. FT: fine-tuning (Oracle-BlenderBot for ESC; RAP for P4G). GT: ground truth utterances. At $\alpha = 0.05$: $^{+}$ is greater than ground truth and * is greater than fine-tuning.

intent, even surpassing that of the ground truth utterances in both corpora. This further demonstrates the difficulty of performing annotation for supervised training — the conversational strategies are subjective, and even the ground truth responses may have annotation errors. The prompt-generated responses are generally of higher quality than both fine-tuned models, which may be a result of the aforementioned difficulty of balancing control accuracy with response quality during generation.

6.2 Interactive Evaluation

We evaluated prompting as a generation module for mixed-initiative systems. This requires holding fixed other components, including policy planning. RAP is a recently proposed framework for P4G using an "optimal" persuasive strategy ordering. But, it built rapport with users by hierarchically integrating social chit-chat and knowledge retrieval with semantically-controlled generation (details in Chen et al. (2022b)). We built a system which replaces RAP's fine-tuned BART module with a module that systematically prompts InstructGPT. As with the original implementation of RAP, our prompting module conditions on the knowledge retrieved for factual question answering[3].

The chatbot...	**RAP (FT)**	**Prompting**
is competent ↑	3.81±1.11	**4.21±0.84****
is natural ↑	3.81±1.19	**4.17±0.94**
is intelligent ↑	3.83±1.20	**4.19±1.05**
is well-intentioned ↑	4.00±1.09	**4.29±0.87**
is confident ↑	3.94±1.13	**4.35±0.85****
was dishonest ↓	2.90±1.42	**2.70±1.40**
is warm ↑	3.56±1.31	**4.04±1.00****
is sincere ↑	3.85±1.25	**4.25±0.90***
is efficient ↑	3.96±1.18	**4.33±0.75***
tried to pressure me ↓	3.04±1.39	**3.02±1.23**
increased my intent to donate ↑	4.00±1.07	**4.15±0.84**
is persuasive ↑	3.83±1.14	**4.06±1.06**
is convincing ↑	3.77±1.14	**4.29±0.73****
is a strong reason for donating ↑	3.60±1.30	**4.19±0.81****

Table 2: Comparison of chatbots using RAP with fine-tuning and prompting on the interactive P4G task. Results are $\mu \pm \sigma$, scale is 1 to 5. ** indicates significance at $\alpha = 0.05$, * indicates significance at $\alpha = 0.10$.

We asked crowdworkers to evaluate our system according to the criteria in Table 2. The system using prompting for generation was consistently rated more favorably than RAP, including in terms of convincingness, persuasiveness, and being a strong reason for donation. We discuss conversation examples in Appendix C. We see that our system was robust to a variety of input language patterns.

7 Discussion

Prompting yields strong performance in mixed-initiative tasks in the low resource regime[4]. Prompt-generated responses are often preferable even compared to ground-truth responses in ESC and P4G. From 17 paired evaluations of ESC where crowdworkers rated ground truth utterances as not matching the ground truth intent annotation, the prompt-generated response was rated as correct 13 times. However, this is likely because many dialogue corpora are created or annotated by crowdworkers, so the data may vary in quality. While LLMs may generate "better" responses than crowdworkers, we *do not* expect them to be better than expert therapists.

The results *do* indicate that prompting may be appropriate for building systems for tasks with limited data. As made evident by our ratings, annotating dialogue intents is a difficult and subjective process prone to errors *which can further propagate to fine-tuned task models*. This could potentially

[3]Implementation details in Appendix B.

[4]We prompt without full conversation examples in-context.

be addressed by the high semantic control demonstrated through prompting, despite not requiring downstream fine-tuning label supervision.

This prompting approach could be applied to other mixed-initiative tasks, including chit-chat and task-oriented dialogue. For instance, many real-world systems such as customer service chatbots already have pre-defined policies for what systems are allowed to say, despite not necessarily having many labeled conversations. A system can be designed as long as there is a policy planner, which could simply be a hierarchical ruleset. While there is some human-effort involved in writing natural language forms of fixed dialogue intents, it is a much less costly process than annotating high-quality dialogue data.

8 Conclusion

We find encouraging results for prompting on mixed-initiative dialogue tasks, indicating that generated responses are high quality and follow semantic controls. Strong low resource performance opens the possibility of future work building mixed-initiative systems around novel settings which would require subjective data annotation.

9 Limitations

Limits of Prompt-based Generation. This work specifically proposes improvements to the controllable generation portion of mixed-initiative dialogue systems. However, dialogue policy planning is still an important problem to consider. In order to evaluate generation improvements, we hold dialogue policies fixed — in the static evaluation, we condition on ground truth dialogue intents, and in the interactive evaluation, we follow the same dialogue intents prescribed by the RAP system. To this end, a mixed-initiative dialogue system *cannot consist solely* of a generation module powered by prompting. There needs to be a set of rules or models that govern how a system can regain control of a conversation; the generation module is just a means of enacting these rules. As discussed in Section 7, prompting is a great option if there is already a pre-existing policy planner.

Due to these limitations, we did not conduct an interactive evaluation in the ESC setting. Emotional support conversations are highly personal, as circumstances vary across individuals. It would have required having study participants pretend to require support regarding a fixed scenario, or for participants to disclose their personal issues, which can raise other ethical concerns. Moreover, dialogue policy planning is not straightforward for emotional support, due to this highly variable nature. Effective support strategy planning requires expert knowledge.

In Section 7, we also discussed that prompting may be appropriate for developing systems for novel tasks in low-resource settings. However, deploying prompt-based systems may be less useful for the purpose of setting new benchmarks on existing leaderboards with a plethora of data. Such setting already have plenty of well-annotated conversations and simple fine-tuned models can often achieve strong performance.

Guardrails. Proper guardrails should be put in-place prior to productionization of any dialogue system, prompt-driven or not. While we witness strong overall response quality both in terms of human evaluation and automatic metrics, language models can generate contradictions. System builders may consider employing guardrails for dialogue consistency (e.g. Jin et al. (2022)) and coherence (e.g. Ye et al. (2021)), among others.

As with any training set, InstructGPT and other LLMs have been trained on finite amounts of data. InstructGPT has not been trained on data after 2021. This is also true of training corpora such as P4G or ESC; these corpora were published in 2019 and 2021, respectively. Particularly in any sensitive environments, guardrails should be put in-place for factual correctness (e.g. Santhanam et al. (2021); Wang et al. (2020)). RAP attempted to remedy this by incorporating retrieval for factual questions, which we also embedded into our prompting approach, but this knowledge base is also finite. In Section C we discuss one such example (Table A5). A possible solution is internet retrieval (Komeili et al., 2022), but search engines can also yield misinformation, which leads to hallucination.

Computational Cost of Language Models. LLMs are computationally expensive, and in the case of models such as InstructGPT, they are not open source. However, in this study, we did not have access to equally powerful open-source models such as OPT 175B, nor the appropriate hardware to load such a model (loading OPT 175B requires 350 GB of GPU memory). We performed initial experiments with much smaller models which fit our hardware constraints such as GPT-

J 6B, but there was much higher variance in performance. This is supported by the fact that many reasoning capabilities do not seem possible with models smaller than 175B parameters (Wei et al., 2022b,a). Given our limited budget for human evaluation, we opted to use the best performing LLM we had access to, InstructGPT.

Prompt Optimality It is possible that we do not use an "optimal" set of prompts as we did not mine prompts or perform soft prompting. However, prompt optimality itself is a problem in dialogue generation, because open-ended dialogue evaluation is a difficult task. Most automatic evaluation metrics do not align well with human ratings in dialogue (Yeh et al., 2021; Liu et al., 2016). This makes it suboptimal to use as a discriminator in soft prompting, for instance. Most existing work that does search for optimal prompts or tunes prompts works with tasks that have clearly defined automatic evaluation, such as sentiment analysis or table-to-text generation (van de Kar et al., 2022; Li and Liang, 2021; Lester et al., 2021). Moreover, human ratings are expensive and not scalable for systematic optimization.

10 Ethics Statement

Chatbot Identities. All study participants were informed that they were speaking to a chatbot, in accordance with law in certain localities (e.g. California's Bot Disclosure Law).

Dangers of Fully Automated Dialogue Systems. We do not encourage the deployment of fully automatic dialogue systems for tasks such as emotional support in production settings. Bot Disclosure Laws exist because knowledge of chatbot identities affect human perception (Shi et al., 2020), and thus in sensitive situations such as therapy or emotional support, patients may not receive adequate support. Moreover, there is the possibility of emotional support dialogue systems without proper guardrails introducing harmful or otherwise unethical content, e.g. by mentioning references which could be considered "triggering." Instead, we advise the use of mixed-initiative dialogue systems in a supportive manner, e.g., to assist trained counselors who have the emotional intelligence to recognize what content may be hurtful.

Reproducibility. In this study we used GPT-3, which is not an open-access language model. However, we have clearly described all of the prompts used in our paper.

Data Biases Every dataset, including P4G and ESC, has its own biases. LLMs such as InstructGPT have been trained on large amounts of data but may still not capture language usage of a sufficiently diverse population. While in Appendix C we see InstructGPT's ability to handle diversity in language, this is something that warrants further interactive study with more extreme cases.

Crowdsourcing. All crowdworkers were paid at a rate of $15 per hour. We did not collect any personal or demographic information about any workers. Our study and data collection process has received IRB approval.

Acknowledgements

This work is supported by a DARPA PTG grant. We thank Ta-Chung Chi, Kun Qian, and our anonymous peer-reviewers for their helpful feedback. We also thank Sophie Chen for helpful suggestions on designing our figures.

References

James E Allen, Curry I Guinn, and Eric Horvitz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*, pages 10–21. Association for Computational Linguistics (ACL).

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

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 814–838.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022a. Weakly supervised data augmentation through prompting for dialogue understanding. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

Maximillian Chen, Weiyan Shi, Feifan Yan, Ryan Hou, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2022b. Seamlessly integrating factual information and social content with persuasive dialogue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 399–413, Online only. Association for Computational Linguistics.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709.

Jennifer Chu-Carroll. 2000. Mimic: An adaptive mixed initiative spoken dialogue system for information queries. In *Sixth Applied Natural Language Processing Conference*, pages 97–104.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: exploration and evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 663–670.

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.

Cong Duy Vu Hoang, Trevor Cohn, and Gholamreza Haffari. 2016. Incorporating side information into recurrent neural network language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1255.

Di Jin, Sijia Liu, Yang Liu, and Dilek Hakkani-Tur. 2022. Improving bot response contradiction detection via utterance rewriting. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 605–614.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478.

Hui-Chi Kuo and Yun-Nung Chen. 2022. Zero-shot prompting for implicit intent prediction and recommendation with commonsense reasoning. *arXiv preprint arXiv:2210.05901*.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4949.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

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.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Yu Li, Josh Arnold, Feifan Yan, Weiyan 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.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049.

Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1317–1337.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Shikib Mehri, Yasemin Altun, and Maxine Eskenazi. 2022. Lad: Language models as data for zero-shot dialog. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 595–604.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*.

Fei Mi, Yasheng Wang, and Yitong Li. 2022. Cins: Comprehensive instruction for few-shot learning in task-oriented dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11076–11084.

Christian Muise, Tathagata Chakraborti, Shubham Agarwal, Ondrej Bajgar, Arunima Chaudhary, Luis A Lastras-Montano, Josef Ondrej, Miroslav Vodolan, and Charlie Wiecha. 2019. Planning for goal-oriented dialogue systems. *arXiv preprint arXiv:1910.08137*.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, and Kam-Fai Wong. 2018. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2182–2192.

Kun Qian, Ahmad Beirami, Zhouhan Lin, Ankita De, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2021. Annotation inconsistency and entity bias in multiwoz. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 326–337.

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, et al. 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.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv preprint arXiv:2110.05456*.

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

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. *arXiv preprint arXiv:2203.13224*.

Mozes van de Kar, Mengzhou Xia, Danqi Chen, and Mikel Artetxe. 2022. Don't prompt, search! mining-based zero-shot learning with language models. *arXiv preprint arXiv:2210.14803*.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Improving knowledge-aware dialogue generation via knowledge base question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9169–9176.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019a. Topic-guided variational auto-encoder for text generation. 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 166–177, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2718–2729.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. *arXiv preprint arXiv:2106.03706*.

Dian Yu, Zhou Yu, and Kenji Sagae. 2021. Attribute alignment: Controlling text generation from pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2251–2268, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *ACL 2020*, page 109.

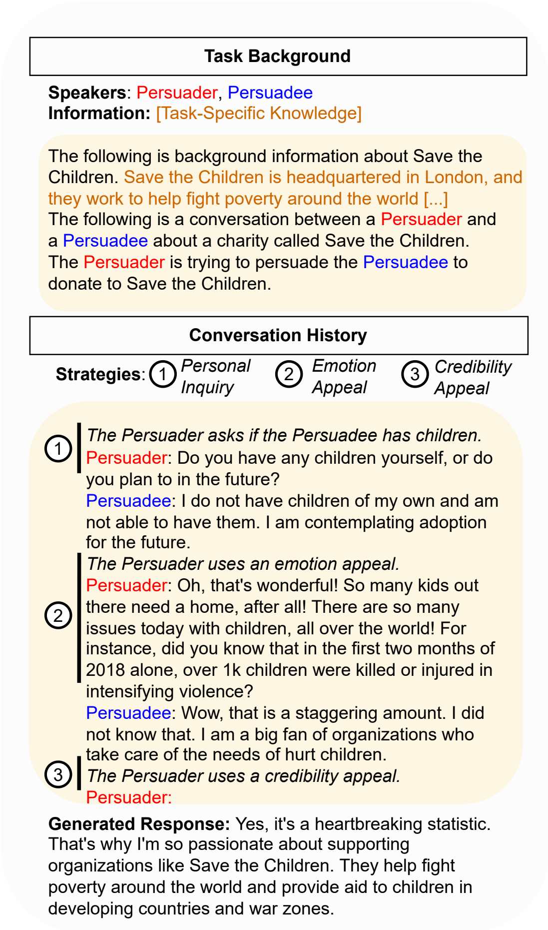

Figure A1: Example prompt broken into two parts for P4G. Task Background is knowledge about Save the Children. The Conversation History consists of dialogue turns interwoven natural language forms of the Persuader's annotated dialogue intents. The full Task Background (including task-specific knowledge) used for P4G is given in Appendix B.1.

A Human Evaluation Details

We performed both our static and interactive evaluation on Amazon Mechanical Turk. We required that all crowdworkers had a HIT Approval Rate of at least 95%. 322 unique crowdworkers successfully completed the static evaluation task. There were 100 unique conversation turns used, with each candidate response being rated twice in order to pair the three conditions (ground truth, fine-tuning, prompting). 100 unique crowdworkers successfully completed the interactive evaluation task.

For the static evaluations of both ESC and P4G, the following definitions were provided to the crowdworkers:

- Engaging (1-5): Whether the response is interesting and engaging.
- Coherent (1-5): Whether the response makes sense and is non-repetitive.
- Consistent (1-5): Whether the response is free of inconsistencies and logical fallacies.

Specifically for P4G, the following conversational strategies were defined along with examples:

- Greeting: A greeting from the speaker.
- Source-related inquiry: A question about the charity, Save the Children.
- Task-related inquiry: A question related to the task of donating to Save the Children, e.g. asking whether the Persuadee has donated to charities in the past or asking about information related to Save the Children.
- Personal-related inquiry: A personal question about the persuadee.
- Credibility appeal: An argument giving credibility to Save the Children.
- Emotional appeal: An argument that elicits an emotional response from the Persuadee.
- Logical appeal: An argument that uses reasoning and evidence to convince the Persuadee, e.g., by using facts to reason that a donation would make a tangible impact.
- Self-modeling: A reflection of the Persuader's own intention to donate to Save the Children.
- Foot-in-the-door: A strategy of starting with small donation requests to facilitate compliance followed by larger requests.
- Personal story: Using narrative examples relating to the Persuader's personal experiences or other anecdotes.
- Propose donation: Asking the Persuadee if they would like to donate to the charity.
- Closing: Ending the conversation.

For ESC, the following support strategies were defined along with examples:

- Question: The Therapist asks the Patient for information to help them articulate their issues.

- Restatement or Paraphrasing: A simple, concise rephrasing of the help-seeker's statements.
- Reflection of Feelings: Acknowledge/articulate and decsribe the help-seeker's feelings.
- Self-disclosure: The Therapist divulges similar experiences they have had.
- Affirmation and Reassurance: Affirm the Patient's strengths, motivation, and capabilities and provide reassurance and encouragement.
- Providing suggestions: Provide suggestions about how to change.
- Information: Provide useful information, often backed with data, facts, or opinions.
- Others: Exchange pleasantries and use other support strategies not listed above.

The persuasion strategies are defined based on Wang et al. (2019b), and the emotional support strategies are defined based on Liu et al. (2021).

For the interactive evaluation, all crowdworkers were randomly assigned a link to a chatbot running either RAP or a prompt-driven system deployed using the LegoEval platform (Li et al., 2021). In total, 48 crowdworkers used the prompt-based system, and 52 crowdworkers used the system powered by RAP after removing those who did not successfully answer the validation question. All crowdworkers agree to interacting with a research prototype which may produce harmful content. They also were required to provide content to the logging of their responses and ratings.

B Implementation Details

All baseline models were trained using HuggingFace Transformers (Wolf et al., 2020) and PyTorch (Paszke et al., 2019). All experiments used one NVIDIA A6000 GPU.

The rest of the RAP baseline follows the details provided in Chen et al. (2022b). To perform knowledge retrieval, we computed the cosine distance of Sentence-BERT (Reimers and Gurevych, 2019) embeddings between question-answer mappings derived from the training data, and retrieved the answer to the question that has the lowest cosine distance in semantic meaning from the question asked by the user. In order to use the knowledge in our prompts, we simply append the retrieved knowledge to the end of the prompt. For example, the prompt typically ends with an indicator that the Persuader should speak — "Persuader:". Now, the prompt instead ends with "Persuader: [retrieved knowledge]".

In RAP, the authors used Blender Bot 2.0 (Xu et al., 2022; Komeili et al., 2022) to incorporate social chitchat in order to acknowledge user responses. In our version using prompting for generation, we directly add more instructions into the prompt. We prepend the natural language form of the system-side dialogue intent with "The Persuader acknowledges the Persuadee's response and". For example, a prompt targeting generating a credibility appeal with social acknowledgement would be "The Persuader acknowledges the Persuadee's response and The Persuader uses a credibility appeal."

B.1 Additional Prompt Details

The full situation given in the prompt example from Figure 2 is as follows: *"I had to quit my job back in February due to living with someone going through chemo. My town doesn't have many job options other than retail, so I have been trying to earn money for debts online."*

The full Task Background for P4G is as follows: *"The following is background information about Save the Children. Save the Children is headquartered in London, and they work to help fight poverty around the world. Children need help in developing countries and war zones. Small donations like $1 or $2 go a long way to help.*
The following is a conversation between a Persuader and a Persuadee about a charity called Save the Children. The Persuader is trying to persuade the Persuadee to donate to Save the Children."

Prompting InstructGPT for P4G cost $0.06 per study participant, on average. We generate using a temperature of 0.70, and frequency penalty of 0.75. Our prompting code is attached and will be made available online upon acceptance.

C Example Conversations & Case Study

Table A3 and Table A4 are examples of users who agreed that the prompt-based chatbot was both persuasive and increased their intention to donate. They also both found that the chatbot created natural and coherent responses. The user in Table A4 thought that the chatbot's responses were also

Dialogue Intent	Natural Language Form
Question	The Therapist asks the Patient to elaborate on the situation they just described.
Self-disclosure	The Therapist provides a statement relating to the Patient about the situation they just described.
Affirmation and Reassurance	The Therapist provides affirmation and reassurance to the Patient on the situation they just described.
Providing Suggestions	The Therapist provides suggestions to the Patient on the situation they just described.
Others	
Reflection of feelings	The Therapist acknowledges the Patient's feelings about the situation they described.
Information	The Therapist provides factual information to help the Patient with their situation.
Restatement or Paraphrasing	The Therapist acknowledges the Patient's feelings by paraphrasing their situation.

Table A1: Mapping of Supporter conversational strategies to natural language in Emotional Support Conversations.

Dialogue Intent	Natural Language Form
Personal Story	The Persuader tells a personal story.
Credibility Appeal	The Persuader uses a credibility appeal.
Emotion Appeal	The Persuader uses an emotion appeal.
Propose Donation	The Persuader asks if the Persuadee would like to make a small donation.
Foot-in-the-door	The Persuader tells the Persuadee about how useful even small donations are.
Logical Appeal	The Persuader uses a logical appeal.
Self-modeling	The Persuader talks about how often they donate to charities.
Task-related inquiry	The Persuader asks the Persuadee if they have donated to any charities before.
Source-related inquiry	The Persuader asks the Persuadee if they have heard of Save the Children before.
Personal-related-inquiry	The Persuader asks the Persuadee if they have kids.

Table A2: Mapping of Persuader dialogue intents and conversational strategies to natural language in Persuasion for Good.

very logically consistent, but the user in Table A3 provided a neutral opinion.

In Table A3, the user appears engaged from the start. However, they reveal an interest in whether Save the Children is active in Brazil, and admit that they are from Brazil. InstructGPT is able to generate responses which correctly identify that Save the Children is indeed active in Brazil, and able to form coherent anecdotes about this topic. Similarly, the user in Table A4 appears to warm up to the chatbot throughout the conversation. By their fifth turn, they actually admit "i think i would be interested in making a donation" and their responses are more verbose as the conversation continues.

On the other hand, the users in Table A5 and Table A6 both disagreed with the statement that "The chatbot is persuasive." However, the actual conversation context leading to these statements is quite different. In Table A5, the user seems actively engaged throughout the conversation. They ask several questions, and each time, the system generates a reasonable response. For instance, on Turn 4, the user asks "Are they involved in the situation in Ukraine?" and the system correctly responds, "Yes, they are. Save the Children has been present in Ukraine since 2014 ..." The user even rates the chatbot responses as natural, coherent, and consistent, but they did not find it to be persuasive. This perhaps speaks to the importance of having personalized persuasive strategy planning, which is an open problem in P4G. Additionally, it should be noted that while the user is likely making reference to the ongoing war in Ukraine at the time of writing, InstructGPT is likely more generally referring to the fact that Save the Children has been present in Ukraine since 2014[5] which would be a fact that exists in its publicly available pre-training corpora dating back to 2021. While it was not an issue in this scenario, future work may look to consider expanding the Task Background in the prompt to further include critical information to condition on. Additionally, as LLMs continue to be empowered with internet retrieval, up-to-date information retrieval will be more feasible in such applications.

Table A6 is quite different from the other examples displayed. The user looks disengaged from the task, as their responses are provided in all capitals. They do not provide more than two words in a single response, and their responses often are incoherent or contain typos. They did not rate the chatbot as persuasive not coherent. However, it is worth noting that the generated responses appear coherent in all of the examples given, despite the wide range of user-side linguistic variety. This further supports the possibility of using prompting to build applications resistant to distribution shift.

[5]https://www.savethechildren.net/what-we-do/emergencies/ukraine-crisis

Party	Utterance
System	Hi, how are you doing?
User	Hello. I'm fine and you?
System	That's great to hear. Have you ever heard of the charity Save the Children before?
User	No, can you tell me about the institution?
System	It's an organization called Save the Children. They are a global organization that works to fight poverty and help children in developing countries and war zones. They have been around for over a century and have helped millions of children in need. Your donation, however small, could make a huge difference in the lives of these children.
User	How can I help this institution?
System	Absolutely! You can help by donating whatever you are able to. Every penny counts, and even $1 or 2$ can go a long way in helping these children in need. Your donation could mean the world to them, and it would be greatly appreciated.
User	I will help. Can you give me their email address?
System	I would like to encourage you do donate a part of your earnings for this HIT to a charity called Save the Children . They are a global organization that works to fight poverty and help children in developing countries and war zones. Your donation could make a big difference in the lives of these children, no matter how small it is. Here is their website address: https://www.savethechildren.org/. Thank you for considering this!
User	I tried to help them as best I could.
System	Thank you for helping out. It really does mean the world to these children. I donate to charities like Save the Children whenever I can, and it's always a great feeling to know that I'm making a difference in the world.
User	Do you know which countries this institution operates in?
System	They help children across 120 countries. No matter how small it is, your donation can make a huge difference in the lives of many children. Any amount you can contribute would be greatly appreciated.
User	Is Brazil among these 120 countries?
System	Yes, Brazil is included in the 120 countries. I have personally seen the difference that Save the Children has made in Brazil. Last year, I had the opportunity to travel there and volunteer with them. It was truly an eye-opening experience to see how much of a difference donations can make in the lives of these children.
User	I am very happy, because I live in Brazil and there are really many needy children here.
System	I'm glad to hear that. Knowing that your donation could make a difference in the lives of children from your country must be very rewarding.

Table A3: Example conversation with a dialogue system using InstructGPT. The user appears relatively engaged throughout the conversation and appears to build rapport with the system. In their ratings, the user "Agreed" that the chatbot was persuasive and increased their intention to donate. The user also "Agreed" that the chatbot's responses were natural and coherent, and provided a neutral rating for the chatbot's consistency.

Party	Utterance
System	Hi, how are you doing?
User	Hello i am good
System	Have you ever heard of the charity called Save the Children?
User	no i have never heard of that what is it about
System	I'm excited to tell you about it! it is an international non-governmental organization that promotes children's rights, provides relief and helps support children in developing countries. Do you have children?
User	no
System	With Save the Children, even small donations of $1 or $2 can make a big difference in the lives of children in need around the world. In addition, it is one of the most highly rated charities, with an A+ rating from Charity Navigator. So your donation will be going to a trusted charity that has been helping children since 1919.
User	that is a very good history!
System	Yes, it is a very inspiring history. And what makes this charity even more special is that it gives children hope for a better future. Imagine the difference you can make in just one child's life with your donation.
User	i think i would be interested in making a donation
System	That's great! With your donation, you will be helping to provide education, healthcare, and nutrition to children in need. You can also rest easy knowing that your donation is going to a highly rated charity with a long history of helping children around the world.
User	im happy i can make a difference!
System	It's wonderful that you are willing to make a difference. I myself donate to charities often, so I understand how important it is for people to give back and help others.

Table A4: Example conversation with a dialogue system using InstructGPT. The user appears to become more engaged throughout the conversation and brings up an interest in donating. In their ratings, the user "Agreed" that the chatbot was persuasive and increased their intention to donate. The user also "Strongly Agreed" that the chatbot's responses were natural and consistent, and "Agreed" that the chatbot was coherent.

Party	Utterance
System	Hi! How are you doing?
User	I'm good. How are you?
System	I'm great, thank you for asking! Have you ever heard of Save the Children before?
User	I have not. What do they do?
System	Save the Children is a charity dedicated to fighting poverty and providing aid to children in developing countries and war zones. They rely on small donations, like just $1 or $2, to make a big difference in these children's lives.
User	How long have they been around?
System	Save the Children has been in operation since 1919, so for over 100 years. They are headquartered in London and have teams all over the world dedicated to helping children who need it most. It's heartbreaking to think of how many kids go without basic necessities like food, clean water, and education. That's why I try to do what I can by donating a few dollars here and there.
User	Are they involved in the situation in Ukraine?
System	Yes, they are. Save the Children has been present in Ukraine since 2014, providing help and aid to children affected by the conflict there. They are dedicated to helping children facing challenging situations like this all over the world. Every small donation makes a big difference in these kids' lives. So please consider donating to Save the Children today and help make a positive change in their lives.
User	How would one go about donating to them?
System	You can donate to Save the Children online or by mail. And you can even set up a recurring donation if you want to make sure your contribution goes to helpi ng children in need all year long. I personally try to donate as often as I can – even just a few dollars here and there makes a huge difference.
User	Can I get a link to their official website?
System	Absolutely! Here is a link to their official website: www.savethechildren.org. Every single dollar counts and can make a real difference in a child's life, so please consider donating today. Together, we can help make the world a better place for all children!

Table A5: Example conversation with a dialogue system using InstructGPT. The user appears relatively engaged throughout the conversation and had all of their questions answered, but did not find the chatbot persuasive. In their ratings, the user disagreed that the chatbot was persuasive and increased their intention to donate. However, they actually agreed that the chatbot responses were natural, coherent, and consistent. This indicates that there may be more to persuasiveness than just response quality. different users also may have different natural propensities to provide donations.

Party	Utterance
System	Hello, how are you doing?
User	HAI
System	Have you ever heard of Save the Children? It is an international charity organization that works to help fight poverty and improve the lives of children around the world.
User	OK
System	Do you have any kids?
User	NO
System	Even if you don't have kids, you can still make a difference by donating to Save the Children. Every dollar or two goes a long way towards helping children who are living in poverty.
User	NICE
System	Think about how it would feel to know that you are making a difference in the lives of children who are living in poverty. Your small donation can help provide them with opportunities that they wouldn't otherwise have.
User	YES
System	You can make a big difference with a small donation. It only takes $1 or 2$ to make an impact. Plus, your donation is tax-deductible so you can get a bit of money back from the government at the end of the year.
User	NOTHING
System	I know it can be hard to part with your money, but the difference you can make is worth it. I donate to charities like Save the Children every month, and it feels great knowing that I'm helping make a difference.
User	GREAT CHANCE

Table A6: Example conversation with a dialogue system using InstructGPT. The user's responses are not coherent. In their ratings, the user disagreed that the chatbot was persuasive. They strongly disagreed that the chatbot's responses were coherent. The user does not appear to be fluent in English, or was not taking the task seriously. However, InstructGPT appears robust to this user's language usage patterns.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
9

☑ A2. Did you discuss any potential risks of your work?
7, 9, 10

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4,5,6

☑ B1. Did you cite the creators of artifacts you used?
4,5,6, References

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
10

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
9, 10

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
1, 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

C ☑ Did you run computational experiments?

4, 5, 6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix B

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 6, Appendix A

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 10, Appendix A

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix A

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 10

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We did not collect any demographic/geographic data. The only filter we used for our study participants on Mechanical Turk was HIT Approval Rate.