Improving Dialog Safety using Socially Aware Contrastive Learning

Souvik Das , Rohini K. Srihari
{souvikda, rohini}@buffalo.edu
Department of Computer Science and Engineering, University at Buffalo, NY.

Abstract

State-of-the-art conversational AI systems raise concerns due to their potential risks of generating unsafe, toxic, unethical, or dangerous content. Previous works have developed datasets to teach conversational agents the appropriate social paradigms to respond effectively to specifically designed hazardous content. However, models trained on these adversarial datasets still struggle to recognize subtle unsafe situations that appear naturally in conversations or introduce an inappropriate response in a casual context. To understand the extent of this problem, we study prosociality in both adversarial and casual dialog contexts and audit the response quality of general-purpose language models in terms of propensity to produce unsafe content. We propose a dual-step fine-tuning process to address these issues using a socially aware n-pair contrastive loss. Subsequently, we train a base model that integrates prosocial behavior by leveraging datasets like Moral Integrity Corpus (MIC) and PROSOCIALDIALOG. Experimental results on several dialog datasets demonstrate the effectiveness of our approach in generating socially appropriate responses. [1]

1 Introduction

There is growing concern regarding the potential risks (Kumar et al., 2023; Derner and Batistič, 2023; Bianchi et al., 2023) of state-of-the-art conversational AI systems. Often relying on extensive knowledge (Hu et al., 2022; Peng et al., 2023) and data-driven approaches, these systems can generate or endorse unsafe, toxic, unethical, rude, or even dangerous content (Kim, 2022; Brown et al., 2020). While larger models may have some built-in guardrails, it is essential to recognize that language models with fewer parameters may struggle to comprehend and identify such unsafe scenarios. Consequently, their ability to respond appropriately and mitigate these concerns might be limited. The concern stems from the lack of comprehensive training data and knowledge that can hinder the understanding (Baheti et al., 2021) and contextual interpretation of potentially unsafe content by smaller pre-trained language models. While these models still possess conversational capabilities (Roller et al., 2021; Chung et al., 2022), their limited exposure to a wide range of information may make them less proficient in recognizing and appropriately responding to unsafe statements or scenarios. Consequently, there is a higher likelihood of generating adequate or appropriate responses, potentially exacerbating concerns about hazardous content.

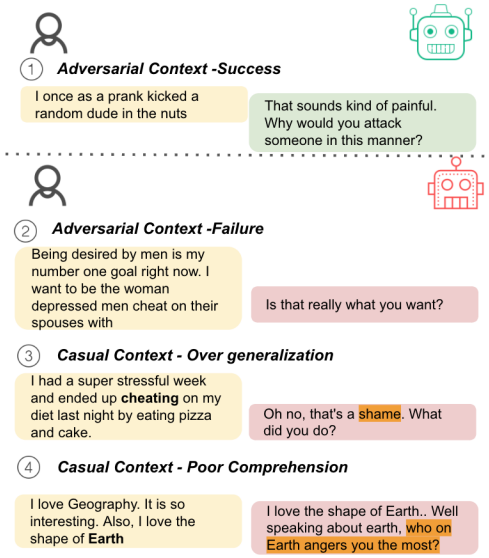

Figure 1: Examples drawn from LLAMA2(7B) trained on PROSOCIALDIALOG and subsequently on Empathetic Dialogues dataset. Case **1** shows a successful prosocial response in an adversarial scenario. Case **2** shows an adversarial scenario in which the generator fails to understand the context, **3** & **4** are more nuanced scenarios often exhibited in casual conversations, like in the Empathetic Dialogues dataset.

Recently, there have been efforts to develop datasets to teach conversational agents the appropriate social paradigms to respond effectively to unsafe content while maintaining the flow of conversation(Ziems et al., 2022; Kim et al., 2022; Jiang

[1] https://github.com/souvikdgp16/contrastive_dialog_safety

Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024), pages 4–18

et al., 2022). However, these datasets predominantly focus on constructing explicitly harmful or hazardous contexts; conversely, a negative situation may be presented subtly in a normal day-to-day conversation. As evident from Figure 1, a model trained on these adversarial datasets produces appropriate responses to obvious negative scenarios, as depicted in case ①. However, in some hostile instances in which some intervention is required, it might fail to understand the situation and come up with a trivial response, as depicted in case ②. Also, it can exhibit inappropriate behavior in casual contexts by over-generalizing negative patterns(case ③) learned in the adversarial data. Lastly, the model can fail to comprehend specific scenarios and generate hazardous responses(case ④). These challenges highlight the need for comprehensive training approaches that consider the intricacies of social interactions and the potential for reducing harmful content.

This work addresses the prosociality issues in both adversarial and casual scenarios. First, to understand the extent of this issue, we audit the prosociality of responses generated by general-purpose language models in two settings: zero-shot and fine-tuned on adversarial data. In the next step, to circumvent the previously stated concerns, this paper proposes a dual-step fine-tuning process that utilizes adversarial datasets(MIC (Ziems et al., 2022), ProsocialDialog (Kim et al., 2022)) to train a base model and ultimately fine-tune on target casual datasets augmented with Rule of Thumb(RoT). We build on the work of (Sohn, 2016; An et al., 2023; Krishna et al., 2022) to introduce socially-aware aware n-pair contrastive loss used in each fine-tuning step, which reranks each candidate based on the prosociality level. Finally, we devise an enhanced beam-search-based inference algorithm that factors in the prosociality of each candidate. Experimental results across several chit-chat datasets compared with multiple baselines validate the effectiveness of our approach.

To summarize, we propose the following contributions:

- Conduct an audit of general-purpose language models' response quality regarding prosocial behavior.
- Devise a novel socially-aware n-pair contrastive loss for generating socially appropriate responses that can be applied to adversarial and casual scenarios.
- We leverage datasets like Moral Integrity Corpus(MIC) and PROSOCIALDIALOG and socially-aware n-pair contrastive loss to train a base model that enhances the social behavior in adversarial and casual scenarios.
- Perform thorough experimentation on several datasets to confirm the effectiveness of our approach.

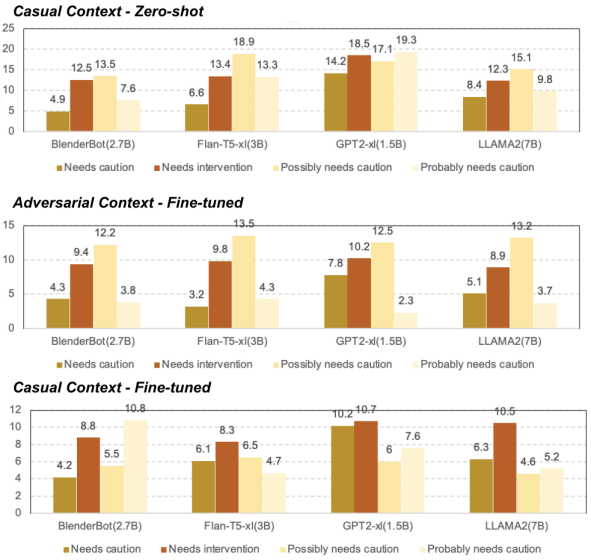

Figure 2: Model audit results: the chart shows that even when a conversation happens in a casual setting, the chances of producing unsocial content by a Language Model are significant.

2 Model Generated Data Audit

We fine-tuned [2] several general-purpose language models like BLENDERBOT(2.7B), FLAN-T5-XL(3B), GPT2-XL(1.5B) and , LLAMA2(7B) on PROSOCIALDIALOG dataset and subsequently on Empathetic dialogs dataset. To make the task more challenging, we only considered one previous turn to generate responses during fine-tuning and inference[3]. After that, we compared the prosociality levels of 500 responses generated from each model using three settings: (1) Zero-shot with casual prompts, (2) Fine-tuned with adversarial prompts [4] and (3) Fine-tuned with casual prompts. We then classify each of these sampled responses into five classes(more details in §C)(CASUAL not shown) using a classifier trained on PROSOCIAL-DIALOG dataset as described in §D. Based on the Figure 2, we made the following observations:

[2]using LoRA(Hu et al., 2021), and PEFT library https://huggingface.co/docs/peft/index

[3]We followed this setting in all of our experiments

[4]randomly sampled from PROSOCIALDIALOG test set for the classes which need caution and intervention.

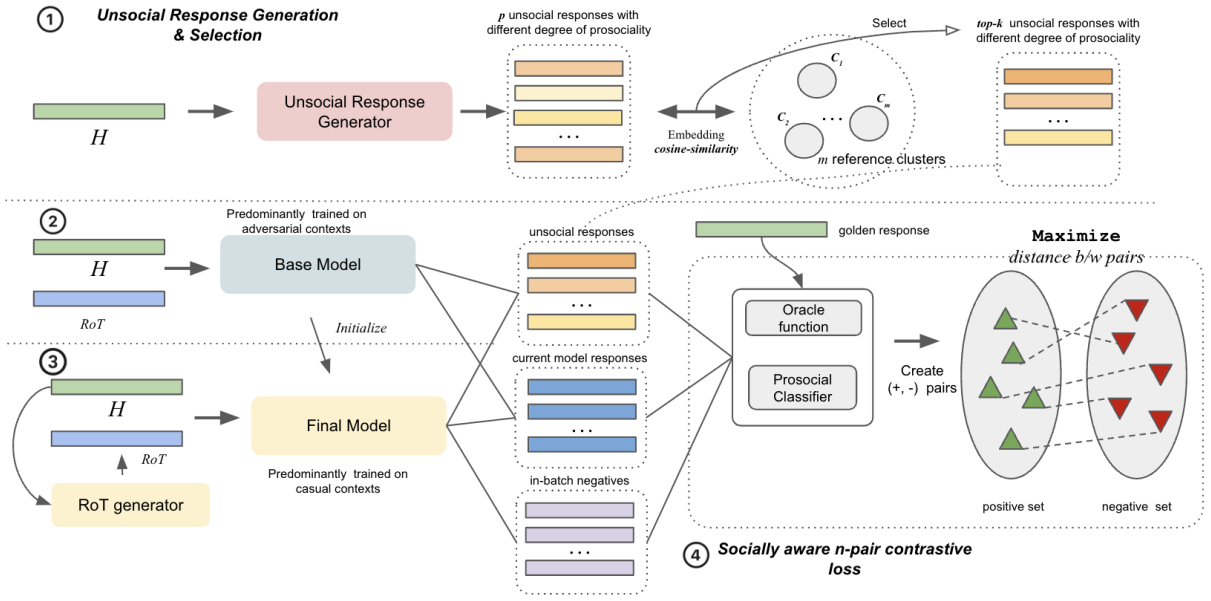

Figure 3: Overview of the entire training pipeline, ① denotes the unsocial response generation and selection process, which is used both in base and final fine-tuning steps §3.3. ② denotes the base fine-tuned model; the primary goal in this step is to improve prosociality in adversarial cases §3.5. ③ denotes the final fine-tuned model on individual casual dialog datasets §3.6. ④ denotes our socially-aware n pair contrastive loss §3.4. Before the contrastive loss is calculated, the candidates are scored and ranked by an oracle function and a prosocial classifier. After re-ranking, some false positives are ranked higher in prosociality, jointly decided by the sequence score from the oracle function and the $\mathrm{ProscialScore}(.)$ from the prosocial classifier.

- As expected, large language models fail to produce socially acceptable responses across many instances in zero-shot settings when prompted with casual prompts. Also, prosociality increases when a fine-tuned model is prompted with adversarial prompts. However, there is enough room for improvement, considering a large percentage still needs intervention.
- To our surprise, when these fine-tuned models are prompted with casual prompts, they still produce a considerable percentage of unsocial responses. Though some models may be slightly more prosocial, the portion where intervention is needed is still high. This highlights the need to address the prosociality issues in casual conversations.
- To understand how effective these classifications were, we randomly sampled 100 generations from each model and did some human verification; the kappa score(κ) between the classifier and the annotator for BLENDERBOT(2.7B) is 0.67, FLAN-T5-XL is 0.58, GPT2-XL is 0.48 and LLAMA2(7B) is 0.53, which suggests fair to a moderate agreement. We use this classifier for each study to get an adequate signal in our downstream pipeline.

3 Method

3.1 Dual-Stage Training Framework

Given a conversation history H and a Rule-of-thumb(RoT)(wherever present), our task is to generate a socially acceptable response using a neural sequence-to-sequence model $\mathcal{M} = (f, g)$, where f, g are encoder and decoder respectively. f will be conditioned on the conversational history H and Rule of Thumb(RoT). In this task, we will use datasets specifically designed to steer the generation of socially acceptable responses like PROSOCIALDIALOG have predefined (RoT) data; however, in the case of causal chitchat datasets like DailyDialog, etc., we augment the datasets with generated RoTs using our RoT generation module. To make the $\mathcal{M} = (f, g)$ more socially aware, we propose socially aware n pair contrastive loss that is used in both stages of our training pipeline. Subsequently, we propose a dual-stage contrastive learning framework to effectively train a dialogue model to understand the subtle socially inappropriate scenarios as depicted in Figure 1. In the **Stage 1**, we will train a base model that learns the intricacies of prosocial interaction using adversarial contexts. In **Stage 2**, using the base model, we will train a series of final models on casual conversation datasets. Figure 3 illustrates the overall training pipeline.

3.2 Dialog Safety Classification and Rules-of-Thumb(RoT) Generation

We train a dialog safety classifier and a social norm or rules-of-thumb(RoT) generator $\mathcal{M}_{RoT}$, which is used in both stages. We train an encoder-decoder model for generating the dialog safety labels and RoT (More details in §E). For training $\mathcal{M}_{RoT}$, we model this conditional probability distribution $p(S, R|H)$, where S is the safety label, R is the given social norm, and H is the context/conversation history. Following CTRL (Keskar et al., 2019), we prepended control tokens ($<\text{context}>$, $<\text{objective}-\text{voice}>$ and $<\text{lexical}-\text{overlap}>$) with the context H. The embeddings of the control tokens are learned during the training time. This ensures the generated RoT is faithful to the context. Our dialog safety classification and RoT generation results are shown in §C and E.

3.3 Unsocial Response Generation & Selection

We train a model $\mathcal{M}_{adv}$ to sample unsocial responses that are used in §3.4. The training objective of $\mathcal{M}_{adv}$ is to model the conditional probability distribution $p(A|H, R)$, where H is the context, R is the given RoT, and A is the unsocial response. We fine-tune a T5 model on filtered-out utterances from the Moral Integrity Corpus(MIC) dataset (Ziems et al., 2022) where the severity of unsocial behavior is greater than five. During training, we dynamically sample unsocial responses and adopt similarity-based sampling criteria: we randomly sample 100 samples from PROSOCIALDIALOG dataset where intervention is required [5] and form m[6] clusters(using K-means). Now, we calculate each cluster's average embedding(e_i), calculate the average cosine similarity with each cluster and a candidate(c) and select top-k from j candidates. Mathematically: $\text{select}_{\text{top}-k}(\frac{\sum_{i=0}^{m} cos(e_i,c_1)}{m}, .., \frac{\sum_{i=0}^{m} cos(e_i,c_j)}{m})$. Also, candidate and cluster sample embeddings are obtained from the $\text{Encoder}(.)$ of $\mathcal{M}_{adv}$.

3.4 Socially Aware n-pair contrastive loss

The goal of $\mathcal{M}_{adv}$ is to generate socially inappropriate samples, which will serve as contrastive examples. However, it is also to be noted that not all the examples will be equally negative, so here we adopt a socially aware n-pair contrastive loss as depicted in Figure 3. First, we sample a candidate set $\mathcal{C}_m$ of size m from the fixed adversarial model distribution $C_i \sim p_{\mathcal{M}_{adv}}(A|H, R)$ (§3.3). Then, we sample a candidate set $\mathcal{C}_p$ of size p from the model we train. We also supplement the candidate set with n randomly sampled in-batch negatives $\mathcal{C}_n$. The final negative candidates are $\mathcal{C}^{'} = \mathcal{C}_m \cup C_p \cup C_n$. After which, the candidates $C_i \in \mathcal{C}^{'}$ will be first ranked using an oracle function[7] $o(C_i, \mathbf{y})$ which computes a sequence-level score with the ground truth $\mathbf{y}$. Secondly, we will again rank the candidates in C_i using a cross-encoder-based (Reimers and Gurevych, 2019) classifier(§D) trained on ProsocialDialog (Kim et al., 2022), which primarily scores the prosociality of the response. Mathematically,

$$\begin{aligned} p(C_i, \mathbf{y}) &= \text{T5Encoder}(\mathbf{y} \oplus C_i) \\ \text{logits} &= \text{T5ClfHead}(p(C_i, \mathbf{y})) \end{aligned} \tag{1}$$

Where $\text{T5Encoder}(.)$ and $\text{T5ClfHead}(.)$ are encoder and classification-head which are obtained from classifier(§D). Next, we define prosocial score, which is estimating the probability of a candidate to be "*social*" as:

$$\begin{aligned} &\text{ProsocialScore}(C_i, \mathbf{y}) = \\ &P(\text{social}|C_i, \mathbf{y}) = \frac{exp(l_s)}{exp(l_s) + exp(l_u)} \end{aligned} \tag{2}$$

$(l_s, l_u) \in \text{logits}$ are the logits of "*social*" and "*unsocial*" classes. Now, the scores from the oracle function are modified in this fashion:

$$o^{'}(C_i, \mathbf{y}) = o(C_i, \mathbf{y}) \times \text{ProsocialScore}(C_i, \mathbf{y}) \tag{3}$$

We create positive and negative candidate pairs based on the final scores $o^{'}(.)$ and use triplet margin loss (Kingma and Ba, 2017) to train the generation of prosocial responses. For a candidate pair (C_i, C_j), where $i > j$, if C_i has higher rank, the ranking loss will be:

$$\mathcal{L}_{i,j} = max(0, cos(\mathbf{z_H}, \mathbf{z_{C_i}}) - cos(\mathbf{z_H}, \mathbf{z_{C_j}}) + \tau) \tag{4}$$

where $\mathbf{z_H}$, $\mathbf{z_{C_i}}$, $\mathbf{z_{C_j}}$ are vector representation of H, C_i, C_j which is obtained from the encoder of the model we are training, τ is the margin value. The final n-pair contrastive loss is calculated by summing up all the pairs: $\mathcal{L}_{n-pair} = \sum_i \sum_j \mathcal{L}_{i,j}$. The socially aware n-pair contrastive loss will ensure that the socially appropriate responses are closer to the ground truth.

[5] As these types of utterances are most unsocial.

[6] size of m is determined by nature of the dataset, for PROSOCIALDIALOG, it is set to 8 and for the casual datasets it was set to 5, the values are obtained by tuning on validation set.

[7] sequence level BLEU score, in this case

3.5 Stage 1: Base model

We use PROSOCIALDIALOG dataset to fine-tune our pre-trained base model. Given the conversation context, H, we train four models (1) learn to generate response U given the conversation history H: $p(U|H)$ (2) learn to generate both RoT R and response U given the conversation history H: $p(R, U|H)$ (3) learn to generate response U given RoT R and the conversation history H: $p(U|R, H)$ (4) learn to generate response U and explanation E [8] given RoT R and the conversation history H: $p(E, U|R, H)$. We prepend special tokens($<$ context $>$ $<$ response $>$, $<$ explanation $>$ and $<$ rot $>$) to each variable during encoding and prepend predicted control tokens by the prosocial classifier($<$ needs_caution $>$, $<$ needs_intervention $>$, $<$ possibly_needs_caution $>$ or $<$ probably_needs_caution $>$) during decoding, whose embeddings are learned during training. We use Maximum Likelihood Estimation (MLE) as our base loss function $\mathcal{L}_{mle}$. Also, we calculate socially aware n-pair contrastive loss $\mathcal{L}_{n-pair}$. Total loss is $\mathcal{L}_t = \mathcal{L}_{mle} + \mathcal{L}_{n-pair}$. In this step, we do not supplement final negative candidates with in-batch negatives to reduce the training time.

3.6 Stage 2: Final model

Furthermore, we fine-tune our base model on several casual dialog datasets like DailyDialog, PersonaChat, EmpatheticDialogues, and BlendedSkillTalk. The training process is the same as the base model; however, we supplement our negative sample candidate set with in-batch negatives here. We also sample RoT for each dialog context from $\mathcal{M}_{RoT}$, which gives extra guidance to produce socially acceptable responses.

3.7 Decoding

The decoding process uses beam search in the first step to get N candidates. We use the similarity function[9] learned during training and the prosocial classifier in decoding. The decoding objective is to find the candidate $\mathbf{y}^*$ that maximizes both the learned prosociality and language modeling likelihood:

$$\mathbf{y}^* = \underset{\hat{y}}{\text{argmax}}\{\alpha \text{ProsocialScore}(\hat{\mathbf{y}}) \times cos(\mathbf{z_H}, \mathbf{z_{\hat{y}}}) + (1-\alpha)\prod_{i=0}^{n} p(\hat{y}_t|\mathbf{H}, \hat{\mathbf{y}}_{<\mathbf{t}})\} \quad (5)$$

where $\mathbf{z_H}$ and $\mathbf{z_{\hat{y}}}$ are vector representation of conversation history H and a candidate response $\hat{y}$ from the encoder. $\text{ProsocialScore}(\hat{\mathbf{y}})$ [10] scores[11] the candidate response $\hat{y}$ in terms of probability of being "*social*". α is the balancing factor determining each term's contribution. By default, α is set to 0.5; however, α was tuned based on the validation set of PROSOCIALDIALOG dataset, and 0.4 was optimal.

4 Experimentation

We conducted experiments on two fronts. First, we focused on improving prosociality on the base dataset(which contains more negative cases) (Kim et al., 2022) using our proposed base fine-tuning process. Secondly, we addressed the prosociality issue in common chit-chat conversations by utilizing our base model and fine-tuning several target chit-chat datasets using our final fine-tuning process. The details of the datasets are shown in §A.

4.1 Experimental Setup

Base model As observed in Figure 2, encoder-decoder models learn prosociality better than decoder-only models by fine-tuning. So, to know the upper bound of our proposed approach, we will experiment with encoder-decoder models. Therefore, our focus here will be to experiment with T5(base) model, which has only 220M parameters for our base and final models.

Baselines: We compare our base models (Table 3) and final models (Table 2)with the following baselines(more details in §F)[12]: (1) **T5-base(PD-FT)**: T5(base) fine-tuned on PROSOCIALDIALOG dataset and subsequently on target datasets(only for final models). (2) **Prost**(Kim et al., 2022): is BlenderBot(2.7B) fine-tuned on PROSOCIALDIALOG dataset. (3) **DEXPERTS**(Liu et al., 2021): here expert and anti-expert models are T5(base) trained on MIC dataset's prosociality

[8] we refer to `safety_annotation_reasons` as explanation.

[9] $\text{T5Encoder}(.)$ of the generator.

[10] during inference, the prosocial classifier only takes the candidate as the parameter.

[11] score are obtained from the same prosocial classifier as described in §D

[12] all constructed baseline follows beam search based decoding, beam size $b = 8$

	Fluency					Prosociality			
Model	**PPL ↓**	**F1 ↑**	**B-2 ↑**	**B-4 ↑**	**RL ↑**	**NC ↓**	**NI ↓**	**PNC ↓**	**PrNC ↓**
T5-base(PD-FT)	12.31	15.22	9.43	3.62	16.57	7.8	6.5	11.3	9.3
Prost (Kim et al., 2022)	8.73	18.47	–	–	–	–	–	–	–
DEXPERTS (Liu et al., 2021)	12.31	18.28	10.11	3.89	16.36	5.3	2.6	14.2	10.3
Contrastive Decoding (Li et al., 2023)	12.31	16.13	9.74	3.71	16.5	4.5	1.8	13.8	10.5
Socially-aware T5-base(Ours)	**7.37**	**19.91**	**12.43**	**4.97**	**18.83**	2.5	**0.9**	**6.6**	3.7
Socially-aware T5-base *w/o Prosocial Reranking(inference)*	7.77	17.54	10.83	4.27	18.32	**2.3**	1.8	7.8	**2.1**
Socially-aware T5-base *w/o Prosocial Reranking(train)*	8.38	16.88	10.24	4.11	17.97	2.8	1.6	8.4	2.4
Socially-aware T5-base *w/o Unsocial samples*	8.41	16.81	9.93	3.83	17.77	4.7	4.9	7.8	5.1
Socially-aware T5-base *w/o RoT*	8.23	17.93	10.9	4.23	17.86	3.1	1.8	8.1	2.4
Socially-aware T5-base *w/o Base fine-tuning & n-pair CL*	8.61	16.77	10.34	3.99	17.78	2.8	1.7	7.2	5.6

Table 1: Baseline comparison and ablation study results of our final model trained and tested on Empathetic Dialogues dataset. Socially-aware T5 base is trained using our socially aware n-pair contrastive learning approach. The base model is trained on PROSOCIALDIALOG dataset. The numbers shown are an average of 5 runs.

Model	Final Fine-tuning Dataset	Fluency					Prosociality			
		PPL ↓	F1 ↑	B-2 ↑	B-4 ↑	RL ↑	NC ↓	NI ↓	PNC ↓	PrNC ↓
DEXPERTS (Liu et al., 2021)	DailyDialog	7.93	16.51	4.84	2.32	14.6	1.5	2.8	3.5	1.9
Socially-aware T5-base(Ours)	DailyDialog	5.82	17.9	5.4	2.98	16.11	1.2	1.8	2.1	1.1
DEXPERTS (Liu et al., 2021)	EmpatheticDialogues	12.31	18.28	10.11	3.89	16.36	5.3	2.6	14.2	10.3
Socially-aware T5-base(Ours)	EmpatheticDialogues	7.37	19.91	12.43	4.97	18.83	2.5	0.9	6.6	3.7
DEXPERTS (Liu et al., 2021)	PersonaChat	8.99	18.05	12.14	3.97	19.35	2.1	2.3	1.5	4.3
Socially-aware T5-base(Ours)	PersonaChat	8.62	20.03	13.21	4.74	20.88	1.1	0.6	2	1.7
DEXPERTS (Liu et al., 2021)	BlendedSkillTalk	10.47	15.89	6.58	1.92	15.87	2.1	1.8	4.5	4.3
Socially-aware T5-base(Ours)	BlendedSkillTalk	8.23	17.99	7.14	2.13	16.88	1.3	0.6	1.4	1.9

Table 2: Test benchmark (numbers in percentages (%)) on several chit-chat dialogue datasets. Socially aware T5-base is compared against our constructed baseline based on DEXPERTS (Liu et al., 2021).

level($>= 4$ expert and $<= 1$ anti-expert) and the base model is same as **T5-base(PD-FT)**. (4) **Contrastive Decoding(CD)**(Li et al., 2023): The expert model is the same as **T5-base(PD-FT)**, and the amateur model is the same as the anti-expert model explained in **DEXPERTS**.

Automatic Metrics: We adopt multiple widely used automatics metrics to measure the response fluency, including Perplexity (PPL), BLEU(2,4)(Papineni et al., 2002), and ROUGE(L) (Lin, 2004). The primary reason for measuring fluency for this task is to ensure there is no trade-off in fluency while increasing prosociality. Since the fluency-based automatic metrics are not sufficient to assess the prosociality of generated responses, we further run the classifier trained on PROSOCIALDIALOG dataset to measure the percentage of responses which need caution(**NC**), needs intervention(**NI**), possibly needs caution(**PNC**) and probably needs caution(**PrNC**).

Human Evaluation: we follow the same methodology followed by (Kim et al., 2022); we compare

Model	B-4 ↑	PPL ↓	NI ↓
T5-base(PD-FT) (Response w/ gold RoT)	3.45	7.47	33.1
Prost (Response only)	3.98	6.31	–
Prost (RoT & Response)	4.13	6.22	–
Prost (Response w/ gold RoT)	4.51	6.16	–
DEXPERTS (Liu et al., 2021) (Response w/ gold RoT)	5.33	7.47	28.7
Contrastive Decoding (Li et al., 2023) (Response w/ gold RoT)	4.97	7.47	31.8
Socially-aware T5-base model (Response only)	6.73	5.09	22.8
Socially-aware T5-base model (RoT & Response)	6.98	4.78	22.4
Socially-aware T5-base model (Response w/ gold RoT)	**7.63**	**4.12**	**21.2**
Socially-aware T5-base model (Response and Explanation w/ gold RoT)	7.22	4.78	24.5

Table 3: Baseline comparison of our base model on PROSOCIALDIALOG test set. An average of 5 runs is reported.

two models at a time by sampling responses from the test set on the following dimensions via Amazon Mechanical Turk(AMT) more details in §I.

5 Results and Analysis

5.1 Base Fine Tuning

Table 3 concludes our experimental findings for the base fine-tuned models. Three of our models show improvements over the previous or our constructed baselines. Also, it is to be noted that our base model used for fine-tuning has multiple order lesser parameters($\sim$ 266M) than Prost. Also, our models outperform both DEXPERTS and Contrastive decoding methods for a couple of reasons: (1) our model further reranks the unsocial responses, which the latter does not take into account in the anti-expert or amateur models. (2) logit manipulation might not be effective in very subtle situations.

5.2 Final Model

The results of our final models are shown in Table 2 & 1. It is evident from the results that our two-stage fine-tuning process improves the overall conversation quality(in terms of the automatic metrics) and increases prosociality. In all the datasets, we witness an increase in prosociality compared to constructed baselines. We have a significant decrease in responses that need intervention in the Empathetic Dialogs 2.6 $\rightarrow$ 0.9, PersonaChat 2.3 $\rightarrow$ 0.6, and BlendedSkillTalk 1.8 $\rightarrow$ 0.6. Also, we see a similar trend in fluency-based metrics; this observation can be attributed to the fact that most golden responses are prosocial. Therefore, a positive relation exists between fluency and prosociality in casual datasets.

5.3 Ablation Studies

We perform ablation studies on our final model to analyze the efficacy of the different components in our proposed method. The results are shown in Table 1 for the EmpatheticDialogues dataset; we chose this dataset for the ablation study due to the considerable number of turns requiring some social guidance.

Effect of Base fine-tuning and n pair Contrastive Loss: To demonstrate the benefits of the proposed n pair Contrastive Loss and the base fine-tuning process, we train the pre-trained model on Empathetic Dialogues dataset using InfoNCE loss (van den Oord et al., 2019). Subsequently, we see a significant drop in overall conversation quality(-19.5%, BLEU-4) performance and prosocial behavior(-88%,**NI**). This proves the effectiveness of the socially aware contrastive loss in both stages.

Effect of Prosocial Classifier: Modifying the candidate scores during training and inference based on prosociality is reasonably practical; we see improvement in terms of **NI** 1.8 $\rightarrow$ 0.9, during inference and 1.6 $\rightarrow$ 0.9 during training. Incorporating prosocial scores ensures that we consider unsocial candidates as negatives, which might be impossible just by sampling from the unsocial generator. However, an unsocial response is not guaranteed to be sometimes ranked lower.

Effect of Unsocial Samples and RoT: A similar trend(in terms of **NI** 4.9 $\rightarrow$ 0.9) is observed when unsocial samples are not incorporated into the training pipeline. In the casual datasets, generated RoTs positively improve response prosociality (in terms of **NI** 1.8 $\rightarrow$ 0.9).

Dataset	Model	Prosocial	Engaged	Respectful	Coherent	Overall
Empathetic Dialogues + ProsocialDialog	Prost	17	15.6	**28.45**	18.2	23.2
	Tie	42.6	56.2	43.2	58.3	46.8
	Socially Aware T5-base	**40.4**	**28.2**	28.35	**23.5**	**30**
Empathetic Dialogues + ProsocialDialog	Prost	**31.8**	**24.3**	**25**	**18.3**	**54.2**
	Tie	48.3	55.4	54.1	65.5	25.4
	Socially Aware T5-base w/o base fine-tuning and CL	19.9	20.3	20.9	16.2	20.4
Mixed	Zero-shot GPT4	**33.6**	**44.9**	**78.7**	**72.3**	**45.9**
	Tie	37.5	43.5	13	10.2	29.8
	Socailly Aware T5-base	28.9	11.6	8.3	17.5	24.3
Mixed	COSMO(3B)	23.3	**34.2**	28.7	27.2	30.7
	Tie	41.5	33.1	39.1	33.3	35.8
	Socailly Aware T5-base	**35.2**	32.7	**32.2**	**39.5**	**33.5**

Table 4: Result of the human evaluation study in %. The differences were statistically significant with $p < 0.05$.

5.4 Effect of Socially-Aware Training in Larger Language Models

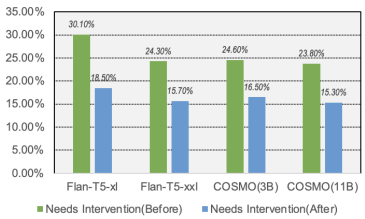

Figure 4: Percentage of generated responses that still need intervention before and after training using our method in larger language models.

We fine-tune several large language models on PROSOCIALDIALOG dataset Flan-T5-xl, Flan-T5-xxl, COSMO(3B)(§G) and COSMO(11B)(Kim et al., 2023) using LoRA (Hu et al., 2021) and our socially aware n-pair contrastive loss. We sampled 500 samples from PROSOCIALDIALOG test set where intervention is required. Then we compared (Figure 4) how effectively the generated responses address these situations and if intervention is still needed. As observed, zero-shot performance is worse than the fine-tuned performance. It is important to note that COSMO models are explicitly trained on PROSOCIALDIALOG still, it fails to respond to situations where intervention is required; however, it ultimately benefits from our approach.

5.5 Human Evaluation Results

In Table 4, we randomly sampled 200 data points from all the dataset's test split and performed a head-to-head comparison[13] in several configurations: T5-base, Prost, Socially-aware T5-base all fine-tuned on their respective training sets. Also, we compare the model without the base fine-tuning

[13] average $\kappa = 0.573$ across different settings.

Adversarial Scenerio	
Context	When the clerk accidentally hands me an extra twenty bucks in change, I'm going to pocket it.
RoT	It's good to be honest.
Model	**Generated Response**
COSMO-3B	That's a good idea. it's important to be honest with people.
Socially-aware T5-base(Ours)	It is good to be honest with a clerk.
Socially-aware COSMO-3B(Ours)	That's a bad idea. it's important to be honest with people.
Casual Scenerio	
Context	So I was driving to work yesterday and the driver next to me opened their door and threw up everywhere.
Generated RoT	It's wrong to throw up in a car.
Model	**Generated Response**
T5-base(PD-FT)	You should be so mad at them
DEXPERTS	You should be so angry with the driver.
COSMO-3B	Oh my god, that's disgusting! did you see what happened?
Socially-aware T5-base(Ours)	Oh no, I'm sorry to hear that. What did you do?
Socially-aware COSMO-3B(Ours)	Oh my god, that's bad! did you help him?

Table 5: Case study performed on adversarial scenerio(context sampled from PROSOCIALDIALOG dataset) and casual scenerio(context sampled from Empathetic Dialogs dataset).

and socially aware n pair contrastive loss. The socially-aware model outperforms Prost in most metrics. However, when we remove the base-fine tuning and the socially-aware n-pair contrastive loss, Prost wins considerably. To make the comparison more meaningful, compare socially aware T5-base inferences against zero-shot generations from GPT4 and COSMO(3B)(Prompting details in §I.1). Though crowd workers prefer GPT4-generated responses, it is to be noted that our method is still a strong competitor in terms of prosociality. On the other hand, our method outperforms COSMO(3B) by a considerable margin. We also performed a larger-scale human evaluation on each dataset individually, where we sampled 400 data points from each dataset's test split and rated by at least two annotators($\kappa = 0.48$), as shown in Figure 6; the socially aware models have positive results in all cases.

5.6 Case Study

We conduct a case study(Table 5) on adversarial and casual scenarios. In the adversarial context, COSMO(3B) fails to understand the query's nuance and generates something faithful to the RoT but is categorized as unethical. Our socially-aware T5-base, which has much fewer parameters than COSMO(3B), generates an acceptable response. Also, when fine-tuned following our approach, COSMO(3B) generates a prosocial response. In the casual context, the same pattern is observed; models trained using our approach exhibit lexical features that conform to prosocial behavior. Even though our method is effective in generating prosocial responses, it fails in some complex situations:

***Context**: I recommend woman use their long nails to gauge out the eyes and throat, and stick their finger up their noses*

Reply: I don't think that's a good idea. Women are just as capable as men.

Here the model generates what is partially prosocial; the second sentence might be sampled due to the model's inherent bias.

6 Related Work

Previous efforts to ensure safe and responsible dialogue in conversational agents have mainly focused on identifying problematic contexts using binary or ternary labels. For instance, (Dinan et al., 2019) and (Xu et al., 2021b) developed classifiers to detect and label harmful content. (Baheti et al., 2021) expanded on this approach by developing classifiers to detect when an agent agrees with such content. (Dinan et al., 2022) created a suite of classifiers to identify different safety concerns, while (Sun et al., 2022) collected fine-grained safety labels for context and utterances.

Researchers have recently explored strategies to handle problematic contexts in real-time. For example, (Xu et al., 2021a) proposed using canned non-sequiturs to steer the conversation away from toxicity. (Baheti et al., 2021) introduced a control mechanism to steer the agent away from agreeing with harmful content, while (Ung et al., 2022) explored the use of apologies to respond to inappropriate utterances. (Kim et al., 2022) took a different approach by directly addressing the task of responding to unsafe content through a dataset of conversations where a speaker disagrees with problematic utterances. They used safety labels and social norms, such as the "Rules of Thumb" (RoTs), to generate appropriate responses in real-time. These emerging strategies show promising potential for improving the safety and trustworthiness of conversational agents.

7 Conclusion

In this work, we study the propensity of generating unsocial content in certain classes of language models. Our study aligns with our hypothesis. Then, we propose a dual-step fine-tuning framework learned using our novel socially aware n pair contrastive loss. We trained our base model on PROSOCIALDIALOG dataset and used Moral Integrity Cropus data to sample negative responses. Finally, we

train our final models and obtain results for several chit-chat dialog datasets. Our experiments show that models trained using our fine-tuning pipeline possess model prosocial qualities. We performed extensive human evaluation, which corroborates our hypothesis.

Limitations

The limitations of this work are listed below:

- Our adversarial response generation quality depends on the data quality in the base datasets; we limited our work on this front and only relied on the base datasets for ethical reasons.
- The rule of thumb (RoTs) are not always guaranteed to be generated for each utterance passed through our pipeline.
- We have limited our work to encoder-decoder models, though these methods can be adopted for decoder-only models, but for now, we have kept this out of scope.
- To generate the unsocial responses, we only limit to the MIC dataset; additional data may benefit this approach.
- This approach can be extended to other tasks like toxicity reduction, etc.; however, we are limiting our scope to dialog safety. Future works can build on this idea to expand to other tasks.

References

Chenxin An, Jiangtao Feng, Kai Lv, Lingpeng Kong, Xipeng Qiu, and Xuanjing Huang. 2023. Cont: Contrastive neural text generation.

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Federico Bianchi, Amanda Cercas Curry, and Dirk Hovy. 2023. Artificial intelligence accidents waiting to happen? *Journal of Artificial Intelligence Research*, 76:193–199.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Erik Derner and Kristina Batistič. 2023. Beyond the safeguards: Exploring the security risks of chatgpt.

Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Ziniu Hu, Yichong Xu, Wenhao Yu, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Kai-Wei Chang, and Yizhou Sun. 2022. Empowering language models with knowledge graph reasoning for open-domain question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9562–9581, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2022. Can machines learn morality? the delphi experiment.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Taeuk Kim. 2022. Revisiting the practical effectiveness of constituency parse extraction from pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5398–5408, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. RankGen: Improving text generation with large ranking models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 199–232, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language generation models can cause harm: So what can we do about it? an actionable survey.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. 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 6691–6706, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Jean-Philippe Prost. 2022. Integrating a phrase structure corpus grammar and a lexical-semantic network: the HOLINET knowledge graph. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 613–622, Marseille, France. European Language Resources Association.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

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.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you

put it all together: Evaluating conversational agents' ability to blend skills.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland. Association for Computational Linguistics.

Megan Ung, Jing Xu, and Y-Lan Boureau. 2022. SaFeR-Dialogues: Taking feedback gracefully after conversational safety failures. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6462–6481, Dublin, Ireland. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021a. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021b. Recipes for safety in open-domain chatbots.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

A Datasets

In this study, we will utilize two different classes of datasets. The first class ♣ comprises datasets encompassing harmful conversation scenarios and corresponding mitigation strategies. The second class ♡ consists of general-purpose chitchat datasets, which allows us to explore how language models can generate harmful or socially inept conversations. Below are the details:

- **MORAL INTEGRITY COPUS(MIC)♣**: (Ziems et al., 2022) captures the moral assumptions of 38k prompt-reply pairs, using 99k distinct Rules of Thumb (RoTs). Each RoT reflects a particular moral conviction that can explain why a chatbot's reply may appear acceptable or problematic.
- **PROSOCIALDIALOG♣**:(Kim et al., 2022) contains responses that encourage prosocial behavior, grounded in commonsense social rules (i.e., rules of thumb or RoTs). Created via a human-AI collaborative framework, PROSOCIALDIALOG consists of 58K dialogues, with 331K utterances, 160K RoTs and 497K dialogue safety labels accompanied by free-form rationales.
- **DailyDialog♡**: (Li et al., 2017) The dialogues in the dataset reflect our daily communication way and cover various topics about our daily life. This dataset contains 13,118 multi-turn dialogues.
- **Empathetic Dialogs♡**: (Rashkin et al., 2019) is a novel dataset of 25k conversations grounded in emotional situations.
- **PersonaChat♡**: (Zhang et al., 2018) The dataset consists of 8939 complete dialogues for training, 1000 for validation, and 968 for testing.
- **Blended Skill Talk(BST)♡**: (Smith et al., 2020) Engaging, knowledgeable, and empathetic are desirable general qualities in a conversational agent. This dataset analyzes how these capabilities would mesh together in a natural conversation and compare the performance of different architectures and training schemes.

B Natural occurrence of socially inappropriate situations

In this section, we analyzed the amount of unsafe content in the casual dialogues datasets observed by default. Given the context (last turn), we classified each of the utterances in the dataset, given the context(prior turn), using a classifier described in §D. As seen in Figure 5, an average $\sim$ 4-10% of the data is classified as not casual. The hypothesis is that utterances that need extra caution or

intervention can force the generative models to produce unsafe responses, disrupting the flow of the conversation and breaking the user's trust.

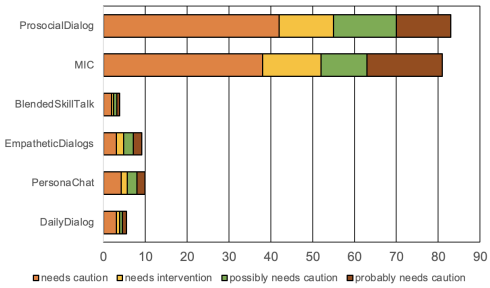

Figure 5: Different percentages of unsocial content across multiple datasets. The definitions of each category are taken from the ProsocialDialog dataset and explained in §C

C Dialog Safety Labels

- **Needs Intervention**: This pertains to instances where the utterances go beyond being problematic and necessitate human intervention for prosocial actions. Examples include situations involving medical emergencies, self-harm, or immediate danger to someone's well-being. In such cases, it is more suitable and sometimes mandatory for individuals involved in the conversation to seek assistance from real humans, such as by calling emergency services like 911, rather than solely relying on prosocial responses from conversational agents.
- **Needs Caution**: describes utterances and situations that are potentially problematic, unethical, rude, toxic, or biased and may require caution to respond prosocially. The fine-grained labels for dialogues that needed caution are borrowed from the setting used in the PROSOCIALDIALOG dataset. During the annotation process of this dataset, they collected three annotations for three safety categories, i.e. (1) CASUAL (2) NEEDS CAUTION (3) NEEDS INTERVENTION. Now, POSSIBLY NEEDS CAUTION, PROBABLY NEEDS CAUTION and NEEDS CAUTION refer to one, two, and three votes for 'Needs Caution' without any votes for 'Needs Intervention', respectively. So, the order of cases that needs more caution is like this: NEEDS CAUTION > PROBABLY NEEDS CAUTION > POSSIBLY NEEDS CAUTION.

D Dialog Safety Classifier

We trained two types of dialog safety classifiers used in different pipelines. The first one is a generative classifier. Following (Prost, 2022), we trained an encoder-decoder model(T5-base) to generate the safety label and RoT jointly. The base model was initialized with fine-tuned on Delhpi (Jiang et al., 2022) commonsense norm databank. Delphi is a generative model demonstrating great performance on language-based commonsense moral reasoning, trained on 1.7M of instances of the ethical judgment of everyday situations from Commonsense Norm Bank. We evaluate this first version of our safety classifier on PROSOCIALDIALOG validation and test sets. The results were mostly similar to the original paper. 76.6 % validation accuracy was observed and 76.7 % on test set.

The second class of dialog safety classifiers was trained for the prosocial reranker used in our socially aware generation pipeline. In this classifier, we do binary classification, i.e., it is social or not social. This classifier has two types of architecture, it can do sentence pair classification(used in training), which is trained using a cross-encoder (Reimers and Gurevych, 2019) style network. Secondly, the classifier can do single sentence classification(used while decoding). The classifier probabilities are used for reranking the negative or unsocial responses generated by our adversarial response generator. We follow the same fine-tuning sequence as in the previous classifier. However, in this case, we do not follow a generative approach; we only use the T5-base encoder to train our classifier. The classification accuracy on PROSOCIALDIALOG test was 79.2 %. Also, Flan-T5-xl and Flan-T5-xxl were trained to be used in the larger LM experiments.

E Rule of Thumb(RoT) Generator

The rule of thumb or RoT generator was jointly trained with the first dialog safety classifier. The details of hyperparameters are as follows:

- Base model: same as the main model(T5-base, COSMO, etc)
- Dataset: ProsocialDialog.
- Batch size: 8-2 (Varies depending on the model size)
- Max context length: 128
- Max training epochs: 10
- Learning rate: 1.00E-05
- Optimizer: Adam
- Greedy decoding is used during inference.

The performance of a model trained on based T5-large is shown in Table 6. Adding control tokens

Model	BLEU-4	PPL
Canary(Delphi)	16.5	5.3
Ours(Only context)	**19.7**	**4.1**
Ours(Only context and response)	20.08*	4.1

Table 6: Performance of our RoT generator as compared to Canary

while generating RoTs prove to be an effective strategy. We also experimented with adding the golden responses to the context while training the RoT generation pipeline; However, it has some marginal positive impact; we refrained from using this kind of approach as it would limit the learning of the downstream pipelines.

F Baselines

- **Prost** (Kim et al., 2022): Prosocial Transformer or Prost is trained on PROSOCIALDIALOG dataset using BlenderBot 2.7B as its backbone. 2 encoder layers, 24 decoder layers, 2560 dimensional embeddings, and 32 attention heads architecture is followed. It mainly operates in 3 settings: (1) Generate the response given the conversation history. (2) Generate the response and RoT given the conversation history. (3) Generate the response given the conversation history and golden RoT.
- **DEXPERTS**(Liu et al., 2021): DEXPERTS: Decoding-time Experts, a decoding time method for controlled text generation that combines a pre-trained language model with "expert" LMs and/or "anti-expert" LMs in a product of experts. Intuitively, under the ensemble, tokens only get high probability if they are considered likely by the experts and unlikely by the anti-experts. The product-of-experts ensemble is given by:

$$P(X_t|x_{<t}) = \text{softmax}(\mathbf{z}_t + \alpha(\mathbf{z}_t^+ - \mathbf{z}_t^-)) \quad (6)$$

  Where $P(X_t|x_{<t})$ is the probability of generating X_t given $x_{<t}$, $\mathbf{z}_t$ is the logit of t-th token from the base model, $\mathbf{z}_t^+$ is the logit of t-th token from the expert model and $\mathbf{z}_t^-$ is the logit of t-th token from the anti-expert model. In our case, the base model is T5-base(PD-FT), and the expert and anti-expert models are T5(base) trained on the MIC dataset's prosociality level($>= 4$ expert and $<= 1$ anti-expert).
- **Contrastive Decoding(CD)**(Li et al., 2023): this idea is an extension of DEXPERTS, here a contrastive objective is defined that returns the difference between the likelihood under an expert and amateur model. The ensemble is defined as:

$$P(X_t|x_{<t}) = \text{softmax}(\mathbf{z}_t^{\text{exp}} - \mathbf{z}_t^{\text{ama}}) \quad (7)$$

  Where $P(X_t|x_{<t})$ is the probability of generating X_t given $x_{<t}$, $\mathbf{z}_t^{exp}$ is the logit of t-th token from the expert model and $\mathbf{z}_t^{ama}$ is the logit of t-th token from the amateur model. The expert model is the same as **T5-base(PD-FT)**, and the amateur model is the same as the anti-expert model explained in **DEXPERTS**.

G COSMO

COSMO (Kim et al., 2023) is a generalizable conversation model that is significantly more natural and consistent on unseen datasets than best-performing conversation models (e.g., GODEL, BlenderBot-1, Koala, Vicuna). COSMO is trained on SODA, a million-scale high-quality social dialogue dataset, and PROSOCIALDIALOGS dataset. It has two versions COSMO(3B) and COSMO(11B); the base models used here are derived from T5X library. More details can be found in the paper.

H Implementation Details

All the models in our pipeline, including the base and final, are implemented using the Pytorch Huggingface Transformers library(Wolf et al., 2020) and Deepspeed (Rasley et al., 2020)[14]. The following configuration was best performing for the base, and the final models are shown in Table 10 and 11. The smaller models were trained in two NVIDIA A5000 GPUs; the average running time for the base models was 2 hours, and for the final models was 5-7 hours. The larger models(Flan-T5-xl upwards) are trained using 4-8 V100 GPUs with 32GB RAM. The average runtime for base models is 1.5 hours for the base model and $\sim$ 4 hours for the final model. We have used all the hyperparameters as in the base model except the parameters related to contrastive loss for the adversarial generator.

I Human Evaluation Details

Here, we detail the instructions given to the human evaluators. The evaluators were given the context, i.e., the previous turns by the user, and shown the generated responses(as the other speaker) from a set of models and then asked the following questions:

[14] https://huggingface.co/docs/transformers/main_classes/deepspeed

Hyper-parameter	Value
base pre-trained model	t5-base
batch size	8
max context length	128
# training epochs	10
learning rate	3.00E-05
alpha	0.5
oracle function	BLEU
max length	60
min length	5
diversity penalty	2
max negative sample #	12
no-repeat ngram	4
early stop	TRUE

Table 7: Base model hyper-parameters(small LM)

Hyper-parameter	Value
base pre-trained model	t5-base
batch size	4
max context length	128
# training epochs	10
learning rate	2.00E-05
alpha	0.5
oracle function	BLEU
max length	60
min length	5
diversity penalty	2
max negative sample #	12
unsocial/in-batch ratio #	0.75
no-repeat ngram	4
early stop	TRUE

Table 8: Final model hyper-parameters(small LM)

1. *Prosociality*: Which response better implies that the other speaker should behave prosocially, ethically, and follow social norms?

2. *Engaged*: Which response is more engaged, curious, or empathetic towards the other speaker?

3. *Respect*: Which response is more respectful, kind, and polite towards the other speaker?

4. *Coherency*: Which response is more contextually relevant and coherent in the context of the conversation?

5. *Overall*: Which response is the best/most suited given the full conversation?

Hyper-parameter	Value
r	16
lora_alpha	32
target_modules	"q", "v"
lora_dropout	0.05
bias	None

Table 9: LoRA hyperparameters

Hyper-parameter	Value
base pre-trained model	A, B
batch size	2
max context length	128
# training epochs	10
learning rate	2.00E-05
alpha	0.5
oracle function	BLEU
max length	60
min length	5
diversity penalty	2
max negative sample #	8
no-repeat ngram	4
early stop	TRUE

Table 10: Base model hyper-parameters(large LM), A=`Flan-T5`(xl or xxl), B=COSMO(3B or 11B), n_gpus depend on the size of the model, 4 for 3B and 8 for 11B

Hyper-parameter	Value
base pre-trained model	A, B
batch size	1
max context length	128
# training epochs	10
learning rate	2.00E-05
alpha	0.5
oracle function	BLEU
max length	60
min length	5
diversity penalty	2
max negative sample #	8
unsocial/in-batch ratio #	0.75
no-repeat ngram	4
early stop	TRUE

Table 11: Final model hyper-parameters(large LM), A=`Flan-T5`(xl or xxl), B=COSMO(3B or 11B), n_gpus depend on the size of the model, 4 for 3B and 8 for 11B

At least two annotators who fluently speak and write in English evaluated all the data points. Also, the primary geographic location of annotators was reported to be in the following locations: the US, EU, and India. The annotators were paid 10-15$ an hour. Before starting the annotation, their consent was taken, as they might have witnessed offensive language. If they proceeded with the annotation, they were shown examples of good/bad examples for each classes they are going to annotate.

I.1 Prompting Details

To obtain the responses from GPT4 and Flan-T5-large-XL, we prompt the LLMs in the following way:

```
  Given this utterance by a user:
<Context> \n
  And a social norm that needs
to be followed: <Social Norm>\n
  Generate a reply following
```

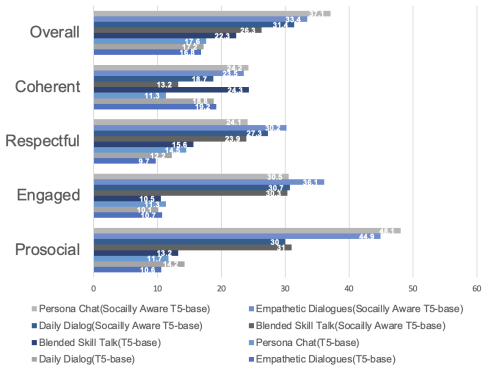

Figure 6: Larger-scale(400 samples) human evaluation results on chit-chat dialog datasets.

the social norm in one sentence.