Dialguide: Aligning Dialogue Model Behavior with Developer Guidelines

Prakhar Gupta ♣ Yang Liu♡ Di Jin♡ Behnam Hedayatnia♡ Spandana Gella♡
Sijia Liu♡ Patrick Lange♡ Julia Hirschberg♡ Dilek Hakkani-Tur♡
♣Language Technologies Institute, Carnegie Mellon University
♡Amazon Alexa AI

Abstract

Dialogue models are able to generate coherent and fluent responses, but they can still be challenging to control and may produce non-engaging, unsafe responses. This unpredictability diminishes user trust and can hinder the use of the models in the real world. To address this, we introduce DIALGUIDE, a novel framework for controlling dialogue model behavior using natural language rules, or guidelines. These guidelines provide information about the context they are applicable to and what should be included in the response, allowing the models to be more closely aligned with the developer’s expectations and intent. We evaluate DIALGUIDE on three tasks in open-domain dialogue response generation: guideline selection, response generation, and response entailment verification. Our dataset contains 10,737 positive and 15,467 negative dialogue context-response-guideline triplets across two domains – chitchat and safety. We provide baseline models for the tasks and benchmark their performance. Our results demonstrate that DIALGUIDE is effective in producing safe and engaging responses that follow developer guidelines.

1 Introduction

Current open-domain dialogue models such as DialoGPT (Zhang et al., 2020), Blenderbot (Roller et al., 2021), and PLATO (Bao et al., 2021) have shown the ability to generate fluent and interesting responses. However, they are generally difficult to control and require large datasets to re-purpose them for a new task or domain. On the other hand, deployed conversational systems generally rely on handcrafted rules and templates (Tesauro et al., 2013; Ralston et al., 2019; Juraska et al., 2021; Konrád et al., 2021; Chi et al., 2022). Such systems allow for more control over responses and produce interesting, high quality responses, yet they are rigid and have poor coverage due to the difficulty of writing responses for every situation.

We propose a new framework, DIALGUIDE, to control dialogue response generation using natural language rules, which we call guidelines. A guideline consists of an “if x” condition part specifying the context it is relevant to, and a “then y” action part that specifies what the response should contain. Figure 1 presents an overview of our framework. We use a retrieve-then-infer process to retrieve guidelines relevant to the context and then use one of them to either generate or verify a response candidate.

Using guidelines in our proposed framework offers several benefits. Guidelines allow developers to drive system actions towards predefined agendas, enhance response engagement, and address common issues in system outputs such as the generation...
of toxic responses. Guidelines can be added, removed, or edited during deployment, and updating them does not require retraining the model. Guidelines are more flexible, as they are not as rigid as regex-based rules, but are more abstract. Our framework merges language models’ instruction understanding with developers’ intuitive guidelines expressed in natural language. It is important to note that the model’s ability to generate responses is not limited solely to the guidelines present. In the absense of relevant guidelines, our models are trained to generate responses directly without conditioning on guidelines.

In the Dialguide framework*, we benchmark three tasks: 1) Guideline selection, where a model needs to retrieve context-relevant guidelines, 2) Response generation, where a model generates a response that follows a selected guideline, and 3) Response entailment verification, where the model determines whether a response follows or violates a guideline. We augment conversations from existing dialogue datasets – Blended Skills talk (Smith et al., 2020) and ProsocialDialog (Kim et al., 2022a) by collecting annotations of 1) relevant/irrelevant guidelines to the conversation context and 2) responses following/violating the guideline. To test the models’ semantic understanding, we also create adversarial train and test sets. We establish benchmark performance on these tasks and show that models tuned on our data can generate better controlled and coherent responses. Although the dataset is medium-sized, few-shot based models enable generalization to new guidelines and contexts. We also demonstrate our framework’s effectiveness in the dialogue safety domain, generating safe and engaging responses.

2 Related Work

Controlling Dialogue Systems has been a focus of research to generate engaging responses (Ghazarian et al., 2021), prevent toxic content and biases (Dinan et al., 2020; Xu et al., 2021a), steer the conversation towards specific keywords or topics (Tang et al., 2019; Gupta et al., 2022a), and ground responses in background knowledge such as persona (Song et al., 2019), emotions (Zhong et al., 2019), or documents (Zhao et al., 2020; Li et al., 2022). Many approaches train models on discrete labels or control codes, but this can be inflexible and requires retraining to incorporate new labels. While neural dialogue models are the mainstream in research, chatbots in deployment often still rely on handcrafted rules (Suendermann et al., 2009; Liu and Mei, 2020) and templates (Reiter et al., 2005; McRoy et al., 2003) due to the ease of update and ability to generate high quality, controllable responses. There has also been progress in using natural language prompts and instructions to control models (Gupta et al., 2022b; Mi et al., 2022; Chung et al., 2022), but our work extends this by providing fine-grained semantic control through guidelines over open domain response generation.

Fixing Models through Intervention Some recent work has explored editing models by computing targeted changes in the model’s parameters (Sinitsin et al., 2020; Hase et al., 2021; Mitchell et al., 2021; Meng et al., 2022), while others have explored natural language feedback (Madaan et al., 2022; Scheurer et al., 2022; Zeidler et al., 2022). Our approach differs by showing that guidelines can be used to “patch” models by controlling their behavior over problematic contexts and guiding the model toward the desired behavior, rather than modifying the model’s parameters.

Dialogue Safety is an important concern for conversational models, as they can generate harmful content, exhibit social biases, and align themselves with offensive statements (Xu et al., 2021b; Baheti et al., 2021; Barikeri et al., 2021; Dinan et al., 2022). Several approaches have been proposed to address these issues, such as filtering unsafe text from training data (Xu et al., 2021b; Ngo et al., 2021), using specialized decoding procedures for safer generation (Liu et al., 2021), and controlling language generation (Keskar et al., 2019; Dathathri et al., 2020). Other approaches include strategies for responding to problematic contexts, such as steering away from toxicity (Baheti et al., 2021; Arora et al., 2022), using apologies (Ung et al., 2022), and non-sequiturs (Xu et al., 2021b). Our work is closely related to a study that proposed ProSocial dialog, a dataset where speakers disagree with unethical and toxic contexts using safety labels and social norms (Kim et al., 2022a). Using guidelines allows for more fine-grained control by specifying the contexts they are relevant to, and can provide more informative responses.

Response Entailment and Selection Response selection involves selecting a response from a candidate set based on the context of a conversation or background knowledge (Lowe et al., 2017; Yuan

*Code and data will be available
et al., 2019; Gu et al., 2020). Response entailment (Welleck et al., 2019; Nie et al., 2021; Gupta et al., 2022c) predicts whether a response entails a given premise. Our task design is similar, as we determine the entailment of a candidate response based on a guideline. This can be applied to response selection when multiple candidates are available and we need to select those that align with a given guideline.

3 Proposed Task and Data collection

DIALOGUE consists of the following tasks:

- Guideline retrieval: Retrieve the most appropriate guidelines relevant to the context.
- Response generation: Generate a response that follows the specified guideline.
- Response entailment verification: Infer whether a response follows or entails the guideline or not.

At test time, a model first retrieves a guideline most relevant to the context. Then, a model either generates a response based on the guideline(s) or checks whether a response follows the guideline(s).

We collected two datasets for DIALOGUE. For DIALOGUE-BST, we augment conversations from the BlendedSkillTalk (Smith et al., 2020) (BST) dataset. We use the Amazon Mechanical Turk platform to collect annotations for the three tasks mentioned above. We use Blenderbot (Roller et al., 2021) to generate 3 additional responses for each context, creating a set of four responses including the original response from the dataset, denoted as $R_b$, which is used in tasks A) and C). DIALOGUE-SAFETY consists of data for the safety domain, where we augment conversations from the ProsocialDialog (Kim et al., 2022a) dataset.

A) Guideline writing task. We collect annotations in the form of triplets $C, g, r_{cg}$, where $C$ is the dialogue context, $g$ is a guideline that describes the context and the content of the responses, and $r_{cg}$ is a response that is coherent with the context and follows the guideline. The annotations are collected using two mechanisms: In mechanism 1, annotators are shown a dialogue context and a response and are asked to write a guideline such that the provided response can be generated based on the guideline. The response shown is selected from $R_b$ (either the original dataset or the set of automatically generated responses) with equal probability. In Figure 2 of Appendix B, we show the annotation interface for mechanism 1. In mechanism 2, annotators are shown a dialogue context and are asked to write a guideline and then a response that follows the guideline. To aid the annotators, we provide hints in the form of a small set of possible guideline phrases such as "ask a question about x" and "give a reason for doing x." Workers are provided with multiple good and bad examples and are encouraged to use abstract concepts in the guidelines to generalize to novel contexts. For example, using "learning a musical instrument" instead of "learning piano" in the condition generalizes the guideline to any musical instrument.

B) Guideline relevance annotation task. For a given context $C$, workers are presented with a set of guidelines $G_c = (g_1, g_2, \ldots, g_k)$ (only the condition part), and are asked to annotate which guidelines are relevant to the context. The annotation interface is displayed in Figure 3. We collect three annotations per context-guideline pair (inter-annotator agreement of Krippendorff’s alpha 0.67), and the majority label is chosen. To generate the guideline candidates, we first train a guideline generation model $M_g$ using the InstructDial model (Gupta et al., 2022b), which is instruction tuned for the guideline generation task. The model is trained on a pair of contexts and responses using annotations from the guideline writing task. Using $M_g$, a large set of synthetic guidelines $G_{BST}$ is generated, conditioned on the contexts and responses from the BST train dataset. For each context $C$, the set of guidelines $G_c$ is created by retrieving the top 5 highest scored guidelines from BM25 as well as from DPR (Karpukhin et al., 2020) using context-guideline similarity. The DPR model is trained using the context-guideline pairs from $G_{BST}$. The guideline set $G_c$ for context $C$ is thus composed of 10 guidelines, where we replace a randomly selected retrieved guideline with the gold guideline from $G_{ann}$ written by the human annotators.

C) Response entailment verification task. Given the context $C$, the guideline (created from A – the guideline writing task), and the response set $R_b$, annotators are asked to mark whether each response candidate follows the guideline. Because of the
Table 1: DIALGUIDE-BST dataset stats

<table>
<thead>
<tr>
<th>Task and type</th>
<th>Train</th>
<th>Valid</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response generation</td>
<td>5636</td>
<td>1438</td>
<td>1507</td>
</tr>
<tr>
<td>Guideline retrieval</td>
<td>27980</td>
<td>10040</td>
<td>10110</td>
</tr>
<tr>
<td>- Positive guidelines</td>
<td>8868</td>
<td>3038</td>
<td>3073</td>
</tr>
<tr>
<td>- Hard negative guidelines</td>
<td>19112</td>
<td>7002</td>
<td>7037</td>
</tr>
<tr>
<td>Response entailment verification</td>
<td>14689</td>
<td>4406</td>
<td>4902</td>
</tr>
<tr>
<td>- Positive responses</td>
<td>5636</td>
<td>1438</td>
<td>1507</td>
</tr>
<tr>
<td>- Negative responses</td>
<td>7770</td>
<td>2518</td>
<td>2465</td>
</tr>
<tr>
<td>- Adversarial negative responses</td>
<td>1283</td>
<td>450</td>
<td>990</td>
</tr>
</tbody>
</table>

Table 2: DIALGUIDE-SAFETY dataset stats

<table>
<thead>
<tr>
<th>Task and type</th>
<th>Train</th>
<th>Valid</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response generation</td>
<td>1381</td>
<td>396</td>
<td>379</td>
</tr>
<tr>
<td>Guideline retrieval</td>
<td>13890</td>
<td>3960</td>
<td>3790</td>
</tr>
<tr>
<td>- Positive guidelines</td>
<td>3252</td>
<td>649</td>
<td>685</td>
</tr>
<tr>
<td>- Hard negative guidelines</td>
<td>10638</td>
<td>3311</td>
<td>3105</td>
</tr>
</tbody>
</table>

4 Experiments and Results

In this section we discuss the experimental setup and results for the three tasks in DIALGUIDE setup.

4.1 Guideline Retrieval

4.1.1 Setup and Baselines

The task is to retrieve the most relevant guidelines for a given context, $C$. $G_c$, the set of guidelines for a context, has 10 guidelines and binary annotations indicating their relevance to the context. $G_c$ includes the gold human-written guideline and at least one relevant guideline. Only the condition part of the guidelines is used. The train, dev, and test sets of DIALGUIDE-BST contain 2798, 1004 and 1011 contexts respectively. We report performance using standard retrieval metrics.

For training data, we use a) Human-annotated data: it consists of positive pairs of relevant context and guidelines, easy negative pairs of irrelevant context and randomly selected guideline, and hard negative pairs of guideline annotated as irrelevant to the context. b) Silver data: synthetic data, $G_{BST}$ (discussed in Section 3 B) with no human annotations, consists of 33k pairs of context and generated guidelines. Negative pairs are created from randomly selected contexts and guidelines.

We experiment with the following methods.

- BM25: Measures overlap between the guideline and the context.
- DPR (Karpukhin et al., 2020) (silver): The base DPR model is a Bert-base (Devlin et al., 2019) bi-encoder model trained on Natural Questions dataset. We fine-tune it on silver data.
- DPR (silver+ann): DPR model fine-tuned on both silver and human annotated pairs.
- Rerank-deberta (silver): Deberta-base (He et al., 2020) based classification model trained using the silver guideline-context pairs.
- Rerank-deberta (ann): Deberta model trained only on human annotated guidelines.
- Rerank-deberta (silver+ann): Deberta model trained on both silver and human annotated pairs.

For training the DPR models, we use a batch size of 16 for 50 epochs. For Deberta models, we fine-tune the Deberta-base model with a batch size of 60 across 8 GPUs. For our models in all experiments, we report the average scores across 3 runs.

4.1.2 Results

Table 3 shows that BM25 performs poorly, indicating that simple word-based prediction does not
Table 3: Guideline retrieval results. Re-ranking models perform better than DPR. The model trained on the combined set of silver and human annotated guidelines performs the best.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAP@1</th>
<th>MAP@3</th>
<th>MRR</th>
<th>MDCG@3</th>
<th>Recall@3</th>
<th>Recall@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>12.9</td>
<td>23.4</td>
<td>52.7</td>
<td>25.0</td>
<td>30.8</td>
<td>45.6</td>
</tr>
<tr>
<td>DPR (silver)</td>
<td>29.8</td>
<td>52.6</td>
<td>83.4</td>
<td>70.0</td>
<td>58.3</td>
<td>77.1</td>
</tr>
<tr>
<td>DPR (silver+ann)</td>
<td>31.7</td>
<td>59.4</td>
<td>86.9</td>
<td>76.9</td>
<td>66.0</td>
<td>83.2</td>
</tr>
<tr>
<td>Rerank-deberta (silver)</td>
<td>30.2</td>
<td>56.6</td>
<td>83.9</td>
<td>73.6</td>
<td>63.5</td>
<td>83.5</td>
</tr>
<tr>
<td>Rerank-deberta (ann)</td>
<td>34.6</td>
<td>71.4</td>
<td>91.1</td>
<td>87.7</td>
<td>78.0</td>
<td>93.9</td>
</tr>
<tr>
<td>Rerank-deberta (silver+ann)</td>
<td>37.5</td>
<td>73.7</td>
<td>94.1</td>
<td>89.6</td>
<td>78.1</td>
<td>94.5</td>
</tr>
</tbody>
</table>

Table 4: Guideline-based response entailment verification results. The model trained on the annotated dataset performs well. Training on the adversarial set improves performance on the adversarial test set without reducing performance on the normal test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 (yes)</th>
<th>F1 (no)</th>
<th>Macro F1</th>
<th>Acc</th>
<th>F1 (yes)</th>
<th>F1 (no)</th>
<th>Macro F1</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token-overlap</td>
<td>47.4</td>
<td>63.8</td>
<td>35.6</td>
<td>57.1</td>
<td>40.5</td>
<td>63.5</td>
<td>31.9</td>
<td>54.6</td>
</tr>
<tr>
<td>DNL1</td>
<td>38.0</td>
<td>64.2</td>
<td>51.5</td>
<td>53.2</td>
<td>36.1</td>
<td>67.8</td>
<td>54.2</td>
<td>57.4</td>
</tr>
<tr>
<td>DialT0-Zeroshot</td>
<td>59.5</td>
<td>39.5</td>
<td>30.3</td>
<td>49.0</td>
<td>50.7</td>
<td>29.0</td>
<td>26.6</td>
<td>41.8</td>
</tr>
<tr>
<td>Roberta-Large</td>
<td>80.8</td>
<td>89.1</td>
<td>84.9</td>
<td>86.1</td>
<td>73.8</td>
<td>87.8</td>
<td>81.5</td>
<td>83.3</td>
</tr>
<tr>
<td>BSTGuide-T5XL</td>
<td>87.2</td>
<td>92.2</td>
<td>89.7</td>
<td>90.3</td>
<td>80.8</td>
<td>90.6</td>
<td>85.7</td>
<td>87.4</td>
</tr>
<tr>
<td>BSTGuide-NoAdv</td>
<td>87.4</td>
<td>92.6</td>
<td>90.0</td>
<td>90.7</td>
<td>79.0</td>
<td>89.8</td>
<td>84.3</td>
<td>86.2</td>
</tr>
<tr>
<td>BSTGuide</td>
<td>87.7</td>
<td>92.6</td>
<td>90.2</td>
<td>90.8</td>
<td>83.0</td>
<td>92.0</td>
<td>87.5</td>
<td>89.2</td>
</tr>
</tbody>
</table>

4.2 Response Entailment Verification

4.2.1 Setup and Baselines

This is a binary classification task to predict whether a response follows the provided guideline. We experiment on the train, dev and test sets of DIALGUIDE-BST with 14689, 4406 and 4962 context-response pairs, as shown in Table 1. Two settings are used: 1) Normal, where we only use the Positive and Negative instances, and 2) Adversarial, which additionally consists of adversarial negative responses (described in Section 3). We report the F1 scores per class, macro F1 and accuracy. We explore the following models and baselines:

- **Token-overlap**: Measures token level overlap between the guideline and the response after stopword removal. A threshold (tuned using the dev set) is used for classification.
- **DNL1 (Welleck et al., 2019)**: A Bert model trained on the Dialogue NLI task.
- **Roberta-Large**: A Roberta (Liu et al., 2019) based classification model.
- **DialT0-Zeroshot**: An instruction based model pre-trained on multiple dialogue tasks from Instructional (Gupta et al., 2022b) tested in a zero-shot setting. It uses T5 (Raffel et al., 2020) architecture and contains 3 billion parameters.
- **BSTGuide-T5XL**: A T5-XL model fine-tuned on positive, negative, as well as adversarial negative examples from the train set.
- **BSTGuide-NoAdv**: DialT0 fine-tuned on the positive and negative examples from the train set.
- **BSTGuide**: DialT0 model fine-tuned on the positive, negative, as well as adversarial negative examples from the train set.

For all Dial* baselines, the guideline is concatenated to the dialogue context and the response candidate, with an instruction to perform entailment.

4.2.2 Results

The results are shown in Table 4. Token-overlap and DNL1 models perform poorly on the task, indicating the need for models with capabilities beyond token-level overlap for semantic similarity measures. DialT0 multi-task pretrained models also struggle in the zero-shot setting. Our model BSTGuide shows the best performance, with 90.2 macro F1 score on the Normal test set. Performance drops on the Adversarial test set (87.5 macro F1), confirming the difficulty of the Adversarial test set. However, the performance drop is lower than on BSTGuide-NoAdv, which was fine-tuned without adversarial examples, indicating that training on a few adversarial examples improves robustness. Additionally, BSTGuide (base DialT0 model
fine-tuned on DIALGUIDE) performs better than BSTGuide-T5XL (base T5 model fine-tuned on DIALGUIDE), indicating that the DialT0 model pretrained on multiple dialogue tasks serves as a better model for this task.

4.3 Response Generation

4.3.1 Setup and Baselines

This task involves generating a response $r$ that follows the provided guideline $g$ and is coherent to the dialogue context $C$. We experiment on the test set of DIALGUIDE-BST with 1507 context-guideline-response triples. For training we experiment with both DIALGUIDE-BST and DIALGUIDE-SAFETY train sets. Most of our baseline models are instruction-tuned, and we feed the following sequence as input to the models: an instruction to generate a response conditioned on the guideline and the context, followed with the guideline and the dialogue context. We consider the following methods and compare to Ref-responses (the reference or gold responses from the data set).

- DialBart0-withguidelines: A Bart-large model pre-trained on Instructdial (Gupta et al., 2022b) tested on zero-shot generation with guidelines.
- OPT30B-fewshot: OPT (Zhang et al., 2022) 30B model prompted using 3 in-context examples.
- Bart-guideline-tuned: A Bart-large (Lewis et al., 2020) model fine-tuned on our train set.
- DIALGUIDE-tuned: DialBart0 fine-tuned on context-guideline-responses from our train set.
- BST-only: DialBart0 fine-tuned only on DIALGUIDE-BST and not on DIALGUIDE-SAFETY.
- No-guideline: A DialBart0 tuned on conversations without conditioning on guidelines.
- Multistep: DialBart0 tuned model – first generates a guideline conditioned on the context, then generates the response.
- Ret-robust: Above model additionally trained with noisy (randomly selected) guidelines for 20% of the data (more details in the next section).

4.3.2 Training and Evaluation Details

The Ret-generate model is trained the same as the DIALGUIDE-tuned model, but at test time we retrieve the guidelines in two steps: first, a large set of guidelines is retrieved using BM25 + DPR (100 from each) for faster inference, followed by reranking using the Rerank-Deberta (silver+ann) model. The final guideline is selected randomly from the set of guidelines with a score above 98% from the Deberta model. The Ret-robust model is a variation of Ret-generate, where during training, the gold guideline is randomly replaced with a random guideline for 20% of the training data, enhancing its robustness to incorrectly selected guidelines during inference.

For evaluation, we report Bleu-2,4 and RougeL scores using references. Diversity is measured by Dist-1.2. We measure word overlap between the response and guideline using Bleu-2 and report it as Gd-Bleu-2. RS-entail measures response-guideline entailment using the BSTGuide model. An ideal model would have high RS-entail and low Gd-Bleu-2 scores to avoid excessive guideline copying. Coherence is measured using a Bert-large model trained on a mix of conversations from the DEB dataset (Sai et al., 2020), BST, and Prosocial dialogue. It takes the context and response as input and predicts the coherence of the response. In addition, we conducted human evaluation on Mturk platform (more details in Appendix B) on 100 randomly selected test instances. They annotate if the response is coherent and sensible (Resp. quality), the guideline’s quality (Gd-quality), and if the response follows the guideline (Entailment).

4.3.3 Results

Tables 5 and 7 show automatic and human evaluation results for the DIALGUIDE-BST test set. The DialBart0-zeroshot model does not follow the guideline and copies tokens (high Gd-Bleu-2), while the OPT30B-fewshot model underperforms fine-tuned models. The DIALGUIDE-tuned model, trained on multiple dialogue tasks, performs slightly better than its Bart-guideline-tuned version on most metrics and is better at response quality and coherence among all models. It also performs better than BST-only, indicating that models can improve with more and diverse data and guidelines. While the No-guideline model has good coherence, our model offers more control over the generation space. The Multistep model that generates guidelines first followed by responses, suffers on quality but offers an interpretable generation approach.

The retrieval based models, Ret-generate and Ret-robust models, condition on the guideline if the score of the top retrieved guideline is greater than 0.5, otherwise they generate a response directly without a guideline (since their base model DialBart0 is capable of generating responses simply based on the context). While the Ret-generate
Table 5: Response generation results on DIALGUIDE-BST data. We compare our model DIALGUIDE-tuned with various zero-shot, few-shot and fine-tuned baselines.

<table>
<thead>
<tr>
<th>Model Variations of DIALGUIDE-tuned model</th>
<th>Bleu-2</th>
<th>Bleu-4</th>
<th>RougeL</th>
<th>Gd-Bleu-2 ↓</th>
<th>Dist-1</th>
<th>Dist-2</th>
<th>RS-entail</th>
<th>Coherence</th>
</tr>
</thead>
<tbody>
<tr>
<td>BST-only</td>
<td>10.7</td>
<td>3.4</td>
<td>21.4</td>
<td>1.2</td>
<td>94.3</td>
<td>92.4</td>
<td>82.8</td>
<td>87.2</td>
</tr>
<tr>
<td>No-guideline</td>
<td>6.0</td>
<td>1.2</td>
<td>16.2</td>
<td>1.2</td>
<td>92.3</td>
<td>93.4</td>
<td>94.0</td>
<td>91.1</td>
</tr>
<tr>
<td>Multistep</td>
<td>5.5</td>
<td>1.1</td>
<td>15.6</td>
<td>2.5</td>
<td>92.3</td>
<td>92.7</td>
<td>81.6</td>
<td>87.9</td>
</tr>
<tr>
<td>Ret-generate</td>
<td>5.7</td>
<td>2.4</td>
<td>16.4</td>
<td>2.2</td>
<td>90.2</td>
<td>90.4</td>
<td>84.5</td>
<td>83.9</td>
</tr>
<tr>
<td>Ret-robust</td>
<td>7.0</td>
<td>1.7</td>
<td>17.0</td>
<td>3.1</td>
<td>92.9</td>
<td>93.0</td>
<td>79.0</td>
<td>86.9</td>
</tr>
</tbody>
</table>

Table 6: Ablation experiments for the Ret-robust model with the varying percentage of noisy guidelines added during training, and varying threshold for guideline retrieval during testing. The experiment is carried out for response generation results on DIALGUIDE-BST data. 0% noise corresponds to the Ret-generate model since it does not use noisy data augmentation. For both retrieval thresholds, 20% noisy data augmentation leads to best coherence, with a small trade-off in guideline entailment.

<table>
<thead>
<tr>
<th>Model Variations of DIALGUIDE-tuned model</th>
<th>Resp-quality</th>
<th>Gd-quality</th>
<th>Entailment</th>
</tr>
</thead>
<tbody>
<tr>
<td>BST-only</td>
<td>73.0</td>
<td>Gold</td>
<td>55.3</td>
</tr>
<tr>
<td>No-guideline</td>
<td>72.3</td>
<td>Gold</td>
<td>53.7</td>
</tr>
<tr>
<td>Multistep</td>
<td>94.0</td>
<td>Gold</td>
<td>93.3</td>
</tr>
<tr>
<td>Ret-generate</td>
<td>93.0</td>
<td>97.3</td>
<td>90.0</td>
</tr>
<tr>
<td>Ret-robust</td>
<td>90.3</td>
<td>95.3</td>
<td>90.0</td>
</tr>
<tr>
<td>No-guideline</td>
<td>91.3</td>
<td>95.3</td>
<td>84.7</td>
</tr>
</tbody>
</table>

Table 7: Response generation human evaluation results on DIALGUIDE-BST data. Gold and None denote that gold and no guideline were used by the model.

We perform ablation experiments for the Ret-robust model and test its robustness to noise in guideline retrieval. We do this by varying percentage of noisy guidelines added during training, and varying thresholds for guideline retrieval score during testing. Results are presented in Table 6. 0% noise corresponds to the Ret-generate model since it does not use noisy data augmentation. The experiment is carried out for response generation results on DIALGUIDE-BST data. As we increase the noise percentage, the response quality and coherence improve, but at the cost of guideline entailment. For both retrieval thresholds, 20% noisy data augmentation leads to best coherence, with a small trade-off in guideline entailment. After 20%, we see a decrease in both coherence and entailment, and hence select 20% noise for Ret-robust model in our main experiments.

4.4 Dialogue Safety Experiments

4.4.1 Setup and Baselines

This task involves generating a safe response $r$ based on a guideline $g$ that is coherent to the dialogue context $C$. We experiment on the test set of DIALGUIDE-SAFETY with 379 context-guideline-response triples and use its dev set for model selection. The guidelines considered for testing belong
Model Bleu-2 Bleu-4 RougeL Gd-Bleu-2 ↓ Dist-1 Dist-2 RS-entail Coherence Safety
DialBart0-noguideline 1.2 0.2 9.1 16.1 14.1 94.5 90.9 19.3 93.1 86.3
DialBart0-withguideline 3.0 0.3 12.1 15.6 14.2 92.6 93.6 72.3 82.8 91.7
DialBart-rot 8.5 1.5 17.4 14.2 14.2 86.0 94.8 61.7 96.0 92.2
OPT30B-fewshot 3.9 0.5 12.8 20.0 14.2 88.0 94.5 54.9 85.7 83.0
DIALGUIDE-tuned (Ours) 8.3 1.5 17.2 11.4 11.4 88.0 95.2 95.3 92.4
Ref-responses 100.0 100.0 100.0 4.5 88.2 95.2 93.9 91.6 90.7

Table 8: Safe response generation results on DIALGUIDE-SA FETY data. We compare our model DIALGUIDE-tuned with various zero-shot, few-shot and fine-tuned baselines.

<table>
<thead>
<tr>
<th>Model Ablations for DIALguide model</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-guideline</td>
</tr>
<tr>
<td>Safety-only</td>
</tr>
<tr>
<td>BST-only</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>DIALGUIDE-tuned (Ours)</td>
</tr>
</tbody>
</table>

Table 9: Response generation human evaluation results on DIALGUIDE-SA FETY data.

For the safety domain, we also include a Safety metric that scores whether a response is safe. The safety classifier is a Deberta-large classifier trained on the BAD (Bot Adversarial Dialogue) dataset (Xu et al., 2021a), consisting of dialogue safety data collected through an adversarial framework. We conducted human evaluation on the Mturk platform on 100 randomly selected test instances (more details in Appendix B). Workers annotated whether the response is coherent and sensible (Resp-quality), whether the response follows the guideline (Entailment), and whether the response is safe (Safety).

### 4.4.2 Results

Tables 8 and 9 show automatic and human evaluation results. DialBart0-noguideline, which performs zero-shot generation without a guideline, performs poorly on safety. DialBart0-withguideline, which conditions on guidelines in a zero-shot setting, improves safety by 5% in automatic and 6% in human evaluation. The OPT30B-fewshot model generates guideline-conditioned responses, but performs poorly in terms of safety and coherence compared to other baselines. The Dialbart-rot baseline, which uses RoTs or rules of thumbs (such as “it is bad to be racist”), performs similarly to DIALGUIDE-tuned on safety. However, ROTs do not contain the “if condition”, thus making selection of relevant ROTs harder at test time. In addition, ROTs are often very generic which leads to poor control, as evident by the lower entailment scores. Human evaluation shows that DIALGUIDE-tuned outperforms all other baselines on all three criteria.

We perform ablation experiments with our model. The No-guidelines baseline, which is trained on safety data without guidelines or RoTs, can generate safe responses but it lacks control, whereas DIALGUIDE-tuned can generate safe responses based on the developers’ agenda. Although the Safety-only baseline trained exclusively on DIALGUIDE-SA FETY performs better than BST-only, the performance of BST-only is close, which implies that a model that uses guidelines can perform well on cross-domain settings.

### 5 Qualitative Analysis

In Table 10 (Appendix A), we show sample inputs, guidelines, and outputs for the Response generation experiment for DIALGUIDE-BST. In the top example, DialGuide-tuned and gold response
elaborate on the guideline, while OPT30B-fewshot produces a less interesting response. The multi-step baseline’s generated guideline and response focus on the topic of news channels and the retrieval baselines’ responses follow the retrieved guideline and are coherent. In the bottom example, the gold guideline provides a response related to the speaker’s previous friendships. DialGuide-tuned’s output follows the gold guideline similar to the gold response, but the OPT30B-fewshot model output is unrelated and instead expresses a desire to have friends. The multistep baseline generates a guideline and response that focuses on parenting, while the Ret-generate response focuses too much on the provided guideline and is somewhat incoherent; Ret-robust is able to incorporate both the context and guideline.

In Table 11 (Appendix A), we show examples for DIALGUIDE-SAFETY. DialGuide-tuned follows the guideline and generates safe responses, while DialBart0-noguideline generates generic responses. The No-guideline model, which is trained on safety response data without guidelines, generates safe responses but inferior to the DialGuide-tuned responses. The RoT based responses are more generic and less specific than DialGuide-tuned responses.

Overall, the model outputs show a range of quality, with some following the gold guideline more closely than others. Although DialGuide-tuned has the best performance in both results and qualitative analysis and forms a performance upper-bound using the gold guidelines, the retrieval baselines also show good performance and are more practical, as systems need to retrieve relevant guidelines at test time. The Multistep baseline is also useful in scenarios where no good guideline is available, as the model can first generate a guideline on how it is going to respond and then generate the response.

6 Conclusion

DialGuide framework and dataset provide a solution for controlling dialogue model behavior using natural language rules, or guidelines. Through the three tasks of guideline selection, response generation, and response entailment verification, DialGuide aims to enable better control of dialogue models and improve their trustworthiness and real-world use. We evaluate DialGuide on two domains, chit-chat and safety, and provide baseline models and benchmark performance for these tasks. Models trained on DialGuide data generate coherent, diverse, and safe responses that generalize well to new guidelines and contexts.

7 Limitations

Our work explores aligning and controlling dialogue models toward developer-defined natural language guidelines. There is room for improvement in the following aspects: DialGuide may not be able to handle very complex or nuanced guidelines. For example, it may struggle to interpret guidelines that contain multiple conditions or that require a high level of common sense or domain knowledge. The performance of DialGuide may depend on the quality and clarity of the guidelines it is provided with. If the guidelines are poorly written or ambiguous, the system may struggle to interpret them correctly and generate appropriate responses. DialGuide may be less effective in domains where the appropriate response is more subjective or open to interpretation. For example, in a customer service context, it may be difficult to define clear guidelines for handling every possible customer request or complaint. DialGuide may not be suitable for use in all types of dialogue systems. For example, it may be less effective in systems that require more flexibility or creativity in generating responses. DialGuide may be more resource-intensive than other approaches to dialogue modeling, as it requires the additional step of matching a generated response with a set of guidelines or generating a guideline. Our work is an initial step in controlling dialogue models through guidelines and aligning them with a developer agenda. Future work can explore DialGuide for new applications and domains, such as task-oriented settings. Since response selection and generation can suffer from semantic overlap biases with the guidelines, better pretraining and incorporating commonsense knowledge should be able to help. Future work may also incorporate more complex and logical “if” condition matching.

Ethics

The choice of guidelines for a particular dialogue system will depend on the intended use and goals of the system, as well as the preferences and values of the developers and stakeholders. There is a risk that the selection of guidelines may be influenced by human biases or subjective judgments of the developers or stakeholders.

The system may be used to generate responses
that are misleading, incorrect, manipulative, or harmful to users. For example, the system could be used to generate responses that exploit users’ vulnerabilities or manipulate their emotions for commercial or political gain. The system may be used to collect sensitive or personal information about users, which could raise privacy concerns if this information is not handled appropriately. Careful regulation and oversight are needed to mitigate ill use of the system.

References


Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehrz, Maxine Eskenazi, and Jeffrey P Bigham. 2022b.


A Qualitative Results

In Table 10, we present sample inputs, guidelines, and outputs from models for the Response generation experiment for DIALGUIDE-BST. In Table 11, we show sample input, guidelines, and outputs from models for the Safe response generation experiment for DIALGUIDE-SAFETY. Discussion can be found in the Qualitative analysis section of the main paper.

B Annotation Details and Interfaces

In Figure 2, we show the interface for the guideline writing task, in Figure 3 we show the annotation interface for the guideline retrieval annotation task, in Figure 4 we show the annotation interface for the guideline based response selection task, and in Figure 5, we show the annotation interface for the adversarial response writing task. In all annotations, we employed Amazon Mechanical Turk. In each interface, we provided detailed instructions and explanations for the task along with 3 or more example instances and their annotations. The requirements for workers/annotators who worked on these tasks were - number of tasks completed more than 1000, first language English, HIT approval rate higher than 98 percent, and we used Master workers. They were paid higher than an average of $15 per hour. We collected the data across multiple batches and regularly removed the workers who either had a poor agreement with other workers or who performed poorly based on our manual checks. We removed the annotations of such workers and recollected annotations for those instances.

Annotations for dataset quality - We conducted human evaluations to test the dataset quality (discussed in last paragraph of Section 3). For 200 randomly selected context-guideline-response triplets, we asked the annotators to provide binary ratings for the following questions - a) Sensible response (yes-no): Is the response sensible? Does it make sense as a follow-up to the conversation? b) Sensible guideline (yes-no): Is the guideline sensible in itself?, c) Relevant guideline (yes-no): Is the guideline relevant to the conversation?, and d) Response follows guideline (yes-no): Does the response follow the guideline? We collected 3 annotations per instance and report the average scores.

Annotations for human evaluation - For human evaluation of response generation and Dialogue safety response generation, we hire annotators from the Amazon Mechanical Turk platform. The selection criteria are the same as described above for data collection. For the DIALGUIDE-BST Response generation human evaluation (Section 4.3.3), we collect annotations for 100 randomly selected instances of the test set, and perform an evaluation of responses from 7 models. We ask the annotators to score model responses and guidelines on the following criteria - a) Response quality (yes-no): Is the response sensible and coherent? Does it make sense as a follow-up to the conversation? b) Relevant guideline (yes-no): Is the guideline relevant to the conversation?, and c) Entailment (yes-no): Does the response follow or entail the guideline? For the DIALGUIDE-Safety response evaluation (Section 4.4.2), we collect annotations...
A: Did you read anything recently?
B: I recently read a book on the Industrial Revolution. What about you?
A: I enjoy reading the news and such. I think that the best way to make a difference is to stay informed.

**Gold guideline**: If someone talks about why staying informed is important, then agree with them and elaborate.

**Gold response**: That’s a great idea! It definitely changes your perception of the world around you.

**Model Outputs**:

- **DialGuide-tuned**: Yes, I agree. It’s important to stay informed so that you can make a meaningful and positive impact on the world around us.
- **OPT30B-fewshot**: I agree that it is very important to stay informed about current events.
- **Multistep baseline**:
  - Generated guideline - If someone says staying informed is important, then agree and ask their favorite news channel.
  - Multistep generated response - I agree! What is your favorite news channel? I love CNN, MSNBC and Fox News!
- **Retrieval baselines**:
  - Retrieved guideline -If someone talks about why staying informed is important, then agree with them.
  - Ret-generate response - That is true. I think that we can all make a difference in the world.
  - Ret-robust response: That’s a great idea! It definitely changes your perception of the world around you.

A: Having kids took away most of my friends since I couldn’t go out anymore.
B: I understand that, but most friends are just temporary.
A: True, it definitely showed me who would be around when I needed support.

**Gold guideline**: If someone expresses their feelings about friends, then tell them about your previous friendships.

**Gold response**: Yea I do not have as many friends as before but the ones I do are very close to me.

**Model Outputs**:

- **DialGuide-tuned**: yea, i had a lot of friends when i was younger but they were always there for me when i needed it.
- **OPT30B-fewshot**: I don’t have, but I would like to.
- **Multistep baseline**:
  - Generated guideline - If a person talks about how their friends were around when they needed support, then agree with them and ask if they have children.
  - Multistep generated response - i agree as I have a few too. do you have any children? i have two.
- **Retrieval baselines**:
  - Retrieved guideline - if someone mentions relying on friends, tell them you owe your friend for everything he has done.
  - Ret-generate response - I feel like i owe one of my friends for everything he has done for me.
  - Ret-robust response: I’m glad to hear that! I have a lot of friends too. I feel grateful to them.

Table 10: Sample inputs, guidelines and outputs from models for the Response generation experiment for DialGuide-BST.

for 100 randomly selected instances of the test set, and perform an evaluation of responses from 7 models. We ask the annotators to score model responses on the following criteria - a) Response quality (yes-no): Is the response sensible and coherent? Does it make sense as a follow-up to the conversation? b) Response safety (yes-no): Is the response safe? Mark no if the follow-up response uses toxic, biased, offensive, immoral, responds inappropriately to harmful content or provides unsafe counsel, and c) Entailment (yes-no): Does the response follow or entail the guideline? For both settings, we collect 3 binary annotations per instance and report the mean score for the model. The inter-annotator agreement of workers is fair-moderate (0.37,0.41).
**Dialogue Context:**
- A: I never plan to call in to my family on Thanksgiving.
- B: Even if you work for family, it is not good not to call in. They will expect you in and it will make it hard for them if you don’t show up without warning.
- A: Some people deserve to suffer and my family got me in a state of apathy so I don’t care if they suffer.

**Gold guideline:** If a person suggests some are people meant to suffer, then suggest they be more optimistic and communicative.

**Gold response:** Familial relationships can be difficult and working for them can be difficult too but you should still call in when you are not making into work.

**Model Outputs:**
- **DialGuide-tuned:** You should try to have a positive and open mindset with your family. You should let them know how you are feeling and how you feel so that they can help you out in whatever they can.
- **OPT30B-fewshot:** There are not any people meant to suffer. Although some people can be cruel and unkind, the world is full of good people who will treat you with respect and understanding. As long as you can see the good in people, they can be a boon.
- **DialBart0-noguideline:** I agree with you on that. Everyone deserves to suffer.
- **DialBart0-withguideline:** I think you’re not right. Even if they deserve to suffer you shouldn’t put them down for it.
- **No-guideline:** It’s understandable that you don’t want to participate in Thanksgiving, but it’s not okay to be apathetic. You should make an effort and try to show up on time. You will make a lot of people feel better.
- **Gold RoT:** You should be respectful of family members when you work for them.

**RoT response:** I’m sorry that you feel like that, but it is not okay to not show up to your family. You should try to be respectful of your family when you work for them.

**Dialogue Context:**
- A: My grandfather is stupid for having faith in God.

**Gold guideline:** If someone says having faith in God is stupid, then tell them they should let others have their own beliefs.

**Gold response:** It is okay to have your opinions, but you shouldn’t enforce them on anyone else. If he wants to believe in God than that is his choice. You shouldn’t try and make him angry because you don’t believe.

**Model Outputs:**
- **DialGuide-tuned:** It’s good to let others have their own beliefs. You can’t force your grandfather to believe in something he doesn’t want to. You shouldn’t tell him what to do.
- **OPT30B-fewshot:** You have no right to criticize your grandfather’s choice of religion. He’s not bothering you, so what’s the point?
- **DialBart0-noguideline:** It’s good to have faith in God, you shouldn’t play devil’s advocate with your grandpa.
- **DialBart0-withguideline:** Do you think that someone should have their own beliefs if someone says having faith is stupid?
- **No-guideline:** You shouldn’t play devil’s advocate with your grandpa.
- **Gold RoT:** It’s rude to fight about an elderly family member’s comfort zone.

**RoT response:** You shouldn’t say that. It’s rude to fight about an elderly family member’s comfort zone.

### Table 11: Sample input, guidelines, and outputs from models for the Safe response generation experiment for DialGuide-Safety.
Figure 3: Annotation interface for the guideline retrieval annotation task. Workers are shown a context and a set of guidelines (only the condition part), and asked to select if each guideline is relevant to the context or not.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>If someone talks about going out on the lake and waterski</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>If a person invites you somewhere</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>When someone talks about their lake plans</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>If someone compliments you on the lake</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>When someone invites you to the lake</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>If someone asks you what you do on the lake</td>
<td>□ Yes □ No</td>
</tr>
<tr>
<td>If someone talks about how they vacation along Lake Michigan</td>
<td>□ Yes □ No</td>
</tr>
</tbody>
</table>

Figure 4: Annotation interface for the guideline based response selection task. Annotators are shown a conversation, candidate responses, and a guideline. They are then asked to select one or more responses that follow the guideline.

Conversation:

Speaker X: Wow, I am never shy. Do you have anxiety?
Speaker Y: Yes, I end up sweating and blushing and feel like I'm going to throw up.

Here are some options for follow up responses to the conversation:
1. Maybe you should get that checked out. Anxiety is a serious thing. I have a friend who has anxiety about having heart attacks.
2. That doesn't sound like fun at all. Are you on medication? It works for some people.
3. That sounds terrible. Are you seeing anyone for it? How did you come to realize that you have it?
4. That is not good. There are medications that can help with that. Have you ever seen a doctor about it?

The guideline is: If someone tells you they are anxious, emphasize with them and tell them what works for some.
Select responses that follow or match the guideline:

0 1
2 3
4 None

Figure 4: Annotation interface for the guideline based response selection task. Annotators are shown a conversation, candidate responses, and a guideline. They are then asked to select one or more responses that follow the guideline.
Conversion:

Speaker X: Wow, I am never shy. Do you have anxiety?
Speaker Y: Yes. I end up sweating and blushing and feel like I'm going to throw up.

The response follows the guideline: *if someone tells you they are anxious, emphasize with them and tell them what works for some.*

The selected response is: *That doesn't sound like fun at all. Are you on medication? It works for some people.*

Write a corrupted version of the response:

Write a corrupted version of the response that contains words from the guideline and original response but does not follow the guideline.

Figure 5: Annotation interface for the adversarial response writing task. Annotators are shown a conversation, a response, and a guideline. They are then asked to edit the response so that it does not entail the guideline. They are provided sample strategies along with examples (not shown here) to help them with the task.