Probing Commonsense Explanation in Dialogue Response Generation

Pei Zhou **Pegah Jandaghi** **Hyundong Cho** **Bill Yuchen Lin**
Jay Pujara **Xiang Ren**
Department of Computer Science and Information Sciences Institute
University of Southern California
{peiz,yuchen.lin,jpujara,xiangren}@usc.edu, {jandaghi,jcho}@isi.edu

Abstract

Humans use commonsense reasoning (CSR) implicitly to produce natural and coherent responses in conversations. Aiming to close the gap between current response generation (RG) models and human communication abilities, we want to understand *why* RG models respond as they do by probing RG model's understanding of commonsense reasoning that elicits proper responses. We formalize the problem by framing commonsense as a latent variable in the RG task and using explanations for responses as textual form of commonsense. We collect 6k annotated *explanations* justifying responses from four dialogue datasets and ask humans to verify them and propose two probing settings to evaluate RG models' CSR capabilities. Probing results show that models fail to capture the logical relations between commonsense explanations and responses and fine-tuning on in-domain data and increasing model sizes do not lead to understanding of CSR for RG. We hope our study motivates more research in making RG models emulate the human reasoning process in pursuit of smooth human-AI communication [1].

1 Introduction

Response generation (RG) systems, which have the basic goal of mimicking human conversation, have as of yet an unmeasured ability to understand communicative intents. In general, standard neural language models build correlative models of linguistic stimuli rather than deep understanding of human-level meaning (Bender and Koller, 2020). As such, there is reason to suspect that, while RG systems today have impressive performance on common metrics (Zhang et al., 2020b; Roller et al., 2021), they achieve this performance without truly understanding human communication. Commonsense reasoning (CSR), defined as "*the basic level of practical knowledge and reasoning concerning everyday situations and events that are commonly shared among most people*" (Sap et al., 2020), is critical in human communication. Specifically, CSR helps establish a common ground consisting of "*mutual knowledge*" between participants, which is key to smooth communication (Clark and Schaefer, 1989; Clark and Brennan, 1991).

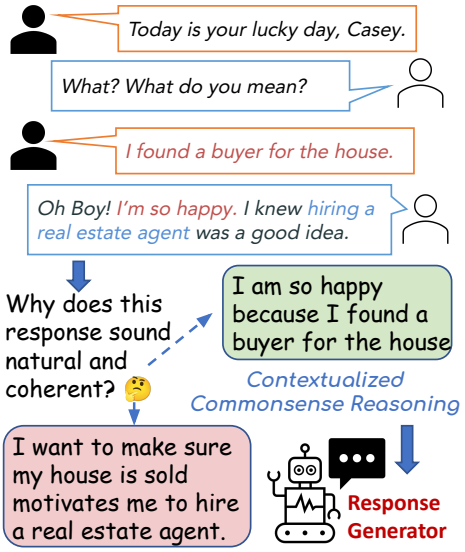

Figure 1: **A motivating example for our study.** We want to know whether RG models understand the implicit common sense that justifies dialogue responses.

For example, consider a conversation between two friends shown in Figure 1. The reason the person on the right (responder) is happy is not indicated explicitly, but it is common sense that finding a buyer for the house (that the responder is likely aiming to sell) makes one happy, which explains the response "*I'm so happy*". Motivated by how humans communicate, we ask a main research question: *do RG models understand the implicit CSR that explains why a response makes sense?* This will help us analyze whether the RG models that seem to produce human-like responses really understand the reasoning process that justifies the response, which is important to build a reliable and

[1] Our code and data are on our project page: https://sites.google.com/usc.edu/cedar.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4132–4146
November 7–11, 2021.

robust dialogue system. Furthermore, understanding implicit common sense behind RG can also help make models generate more natural and coherent responses.To answer this important research question, we present our initial findings from *annotating* commonsense explanations in dialogues and *evaluating RG models* for commonsense reasoning capabilities.

We first present a probing setup for evaluating common sense in RG, called *CEDAR*: **C**ommon s**E**nse in **D**i**A**logue **R**esponse generation. We start with formalizing CSR in RG by considering common sense as a latent variable that helps explain the observed variable "*response*" in the RG process – similar to how humans use common sense in communication (Hilton, 1990). To instantiate implicit common sense for probing, we use textual explanations of the response as the common sense embedded in the dialogue context. To understand whether RG models can comprehend implicit common sense, we *corrupt* explanations to break the logical coherence and compare model behaviors between a valid explanation and a corrupted one.

To operationalize the probing, we collect the first annotations on commonsense explanations that justify dialogue responses. Each annotation is a dialogue-specific explanation that explicitly describes what might cause the response in one of the five dimensions: event, emotion, location, possession, and attribute, inspired by human cognitive psychology (Kintsch and Van Dijk, 1978). We find through pilot studies that directly asking people to annotate result in explanations with high variation and subjectivity, to account for this, we first generate candidate explanations by adopting a large text-to-text language model trained on a story explanation dataset, namely GLUCOSE (Mostafazadeh et al., 2020), under the dialogue setting. Next, we conduct a carefully designed two-stage human verification process with a qualification test and the main annotation task. We present our findings from verifying 6k generated explanations on 1,200 dialogues sampled from four public dialogue datasets.

Using the annotated explanations, we probe state-of-the-art (SOTA) RG models for two CSR-related abilities: (i) the ability to understand whether the commonsense explanation can justify a response, and (ii) the ability to attribute logically-coherent explanations for dialogue responses. These are inspired by what showcases human understanding of common sense in conversational communication. Our probing setup contrasts valid explanations with corrupted version. Corruptions are generated via two methods: *logical corruptions* that disrupt logical coherence, and *complete corruption* where we disrupt the grammatical naturalness of the sentence.

We find that the models fail to understand common sense that elicits proper responses according to performance on our probing settings and some models even do not distinguish gibberish sentences. Fine-tuning on in-domain dialogues and verified explanations do not help with understanding. We also find interesting cases that show potential statistical biases in RG models. We hope our annotated explanations and probing findings encourage more studies on making RG models communicates with deep understanding of human reasoning process.

2 Task Formulation and Challenges

This section first introduces how we incorporate common sense as a latent variable in the RG setting. Then we specify two challenges that arise in order to examine whether RG models can comprehend common sense to arrive at responses similarly as humans do. Lastly, we present our solutions to the challenges by instantiating common sense as textual explanation and proposing two probing settings to evaluate if models reason about common sense when generating responses.

2.1 Common Sense in Response Generation

Preliminaries We consider the classic dialogue response generation (RG) setup (Weizenbaum, 1966; Ritter et al., 2011; Sordoni et al., 2015): given a dialogue *history* H, generate an appropriate *response* R. Most state-of-the-art (SOTA) neural RG models generate a response given a dialogue history as a *conditional language modeling* problem. Specifically, given a *history* (H) consisting of a sequence of dialogue turns from the dialogue history $x_1, x_2, ..., x_n$ (each containing a sequence of tokens) and a *response* (R) sentence y comprised of a sequence of tokens $y_1, y_2, ..., y_m$, RG models aim to learn the conditional probability distribution by training on human dialogues:

$$P_\theta(R|H) = \prod_{i=1}^{m} P_\theta(y_i|y_{<i}, x_1, ..., x_n). \quad (1)$$

Common Sense as a Latent Variable As illustrated in Figures 1 and 2, when humans respond in a conversation, we use common sense implicitly to

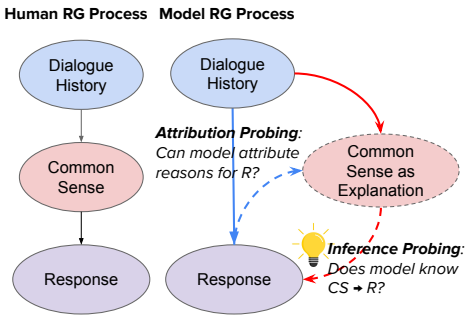

Figure 2: **Probing setting** illustrations. We draw inspirations from human reasoning process during communication and probe RG models' understanding of implicit common sense in RG in two ways (red and blue dotted lines).

establish *common ground* (Grice, 1975; Clark and Brennan, 1991), reach mutual understanding, and help produce natural responses for smooth communication. We consider common sense to be *latent* because it is infrequently stated due to the cooperative principle that states that participants should "*not make your contribution more informative than is required*" (Grice, 1975). However, the reasoning it enables is an integral part establishing common ground and critical for communication. To formalize this process, we consider *common sense* (CS) as an important *latent variable* in the modeling of a dialogue response when given the history – *i.e*, $P(R|H, CS)$. Other latent factors such as the environment in which the conversation happens and background information of the participants can also influence the dialogue, but here we focus on common sense.

2.2 Probing Setup

Current RG models generate responses in an end-to-end manner with only input from dialogue history (*i.e.*, H), making it non-trivial to examine if they understand the implicit common sense behind RG process (also see Figure 2 for an illustration). We instantiate implicit common sense in dialogues and then design probes to evaluate models' grasp of common sense. This leads to two key challenges: 1) *how to instantiate abstract and implicit common sense CS in dialogues?* and 2) *how to probe RG models' understanding of common sense in dialogue response generation?*

Instantiate Common Sense Using Explanations We use *natural language explanations justifying why a response makes common sense* as a proxy to instantiate common sense in RG. Traditional studies have tied common sense and the ability to provide explanations for events and actions closely (Hansen, 1980; Hilton and Slugoski, 1986), which also holds true in a conversational setting (Hilton, 1990). Specifically, as shown in Figure 1, "*I want to make sure my house is sold*" is a potential explanation about what leads to "*hiring a real estate agent*" in the response and this explanation requires understanding the commonsense relation that a real estate agent helps sell a house and the desire to sell a house motivates a person to hire an agent. Formally, we concretize the abstract latent variable common sense CS in textual form as an explanation E explaining what might cause the response R given the history H. We introduce our process of collecting such explanations for RG in Section 3.

Probe Models' Understanding in Two Settings We then draw inspiration from human reasoning process behind dialogue response generation to design two probing tasks. First, humans use common sense implicitly to produce natural and coherent responses in conversations (Clark and Schaefer, 1989). Common sense helps humans determine what responses make sense in certain context. We want to see if providing common sense in the form of explanation also helps RG models arrive at coherent and natural responses more easily. Second, humans can perform *causal attribution* on an event or an action by finding reasons that might cause it (Hilton, 1990). If the person producing the response is asked about why they are feeling happy, they can easily respond with reasons about their reasoning process. We are interested in examining can RG models also generate responses to justify a previous response when asked.

We probe RG models in a *contrastive* manner, by comparing model behaviors with a valid explanation E to the response and a *corrupted* E' that breaks logical coherence. We introduce the two settings in more detail as follows.

Inference Probing Here we directly measure if $P(R|E, H) > P(R|E', H)$ for RG models, *i.e.*, can models assign a higher probability to the response when provided with valid common sense in the form of explanations compared to logically-incoherent explanations? Since existing RG models are not trained to take explanations as additional input, the probing results may be confounded by the model's unfamiliarity with the probing set-

ting. To account for this issue, we 1) probe on a knowledge-grounded RG model that is used to taking in additional knowledge sentences as input and 2) fine-tune RG models on a proportion of our collected explanation and compare the effects. We discuss results and issues about probing models fine-tuned on explanations in Section 5.3.

If the model assigns a similar or lower probability to the response given a valid explanation compared to a logically-incoherent explanation, it indicates that the reason why this response makes sense is not clear for models.

Attribution Probing Here we examine if $P(E|H,R) > P(E'|H,R)$, *i.e.*, can RG models perform *causal attribution* as humans by assigning a higher probability to a valid explanation of the response (that makes sense) compared to a corrupted explanation, given the dialogue history and the response? To address the unfamiliarity of models, we make the probing setting close to real dialogues by continuing the conversation (consisting of H and R) with "*why*" to prompt the models to generate an explanation. We also conduct fine-tuning on a proportion of our collected explanations similarly to the first setting discussed in Section 5.3.

If the model prefers the attribution of the response that is incoherent with the response by giving it a higher probability, it indicates that the model fails to generate valid reasons for responses, which requires understanding the implicit common sense behind dialogues.

3 Generating Commonsense Explanations for Dialogue Responses

To get explanation annotations for dialogue responses, we first automatically generate commonsense causal explanations and then manually verify via crowdsourcing. We use a text-to-text model trained on commonsense story explanation dataset GLUCOSE (Mostafazadeh et al., 2020) as the generator and conduct 2-stage human verification on generated explanations. We first introduce the model we use, the adaptation of the model on dialogue data, and our verification process to ensure the quality of generated explanations.

3.1 Generating Commonsense Explanations

GLUCOSE is a large-scale dataset of implicit commonsense causal explanations grounded in a story context (Mostafazadeh et al., 2020). Given a short story and a sentence X in the story, GLUCOSE contains human annotations of five dimensions of causal explanation related to X (an event/emotion/location/possession/attribute leads to X), each in a semi-structured form "antecedent *connective* consequent." Using the collected explanations, the authors train state-of-the-art neural models and find that the trained models are able to produce commonsense inferences on unseen stories. More details about how models are trained on GLUCOSE are included in Appendix A.

Examples of Verified Explanations:
Event: *I found a buyer for the house causes I am so happy*
Emotion: *I want to make sure my house is sold motivates I hire a real estate agent*
Attribute: *I have an agent enables I knew hiring a real estate agent was a good idea*

Figure 3: Examples of human-verified commonsense explanations for the dialogue shown in Figure 1

We consider using a model trained on GLUCOSE to automatically generate commonsense explanations in dialogues for several reasons. First, it generates *contextual commonsense explanations* that provides causal knowledge about what justifies a sentence. Second, it provides fine-grained causal explanations along different dimensions. Last but not least, we have conducted multiple rounds of pilot studies to directly ask workers to *write out* commonsense explanations for a response, but the subjectivity of this open-ended task led to large variations in quality. Instead we ask workers to *verify* explanations generated from a model.

We sample 1,200 dialogues from 4 dialogue datasets (300 from each): DailyDialog (Li et al., 2017), EmpatheticDialogues (Rashkin et al., 2019), MuTual (Cui et al., 2020), and SocialIQA-prompted dialogues (Zhou et al., 2021). We generate 6k commonsense causal explanations (5 dimensions for each dialogue), using the last turn as the response and the previous turns as dialogue history (after filtering short turns). We follow Zhou et al. (2021)'s approach to select dialogues that contain at least a one-hop triple from ConceptNet (Liu and Singh, 2004). We use the same hyperparameters and weights from the best-performing 770M T5 model from Mostafazadeh et al. (2020).

3.2 Verification

To ensure the quality of generated explanations, we carefully design a two-stage human *verification* process with a qualification test and the main task.

Workers must first pass a *qualification test* (QT) that tests their understanding of the CS criteria necessary for our main annotation tasks (more details in Appendix B). We consider three criteria, requiring generated explanations to pass all three to be considered a valid commonsense explanation for a response. We ask *three* workers on Amazon Mechanical Turk (MTurk) to annotate the *three* criteria for each explanation.

Criteria 1). **Relevant**. A good causal explanation has to focus on explaining what could cause the *response* in the dialogue context (Hilton, 1990). An example of an irrelevant explanation for the example shown in Figure 1 is "*I possess a house enables I live in a house*" since "*living in a house*" is not what the response is about, so it doesn't help explain the response. 2). **Non-trivial**. We observe that sometimes the model simply duplicates a previous dialogue turn as the cause, which trivially associates history and response. We are interested in implicit and specific commonsense so we filter out explanations that parrot a previous turn. For example, "*I found a buyer for the house motivates Oh Boy! I'm so happy. I knew hiring a real estate agent was a good idea.*" 3). **Plausible**. We ask humans to verify if the generated explanation plausibly identifies a likely cause for the response. An example of an implausible explanation is "*I am in a house enables I am so happy*" since "*I am in a house*" is not the direct cause why the person producing the response is feeling so happy, "*found a buyer for the house*" is. This is the hardest criterion for humans to decide due to its subjectivity nature.

Results We present results of our verification of 6k explanations from three in-house annotators. To filter ambiguous explanations and be strict about the quality of verified explanations, we only consider explanations valid if *all* three annotators have agreed that they satisfy *all* three criteria, i.e., 100% agreement for all verified explanations. For the annotated explanation, passing rates (agreed by 3 workers) for criterion (relevant, non-trivial, plausible) are (55%, 73%, 37%) – yielding an overall passing rate of 26% (1,560 explanations). Passing rates for the five dimensions (event, emotion, location, possession, attribute) are (31%, 33%, 13%, 24%, and 29%), with more details in Appendix B. Figure 3 presents examples for different dimension, full data is included in the supplementary material.

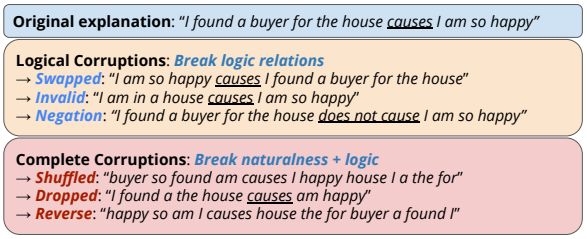

Figure 4: Examples of different **corruption types** to a commonsense causal explanation.

4 Probing Setup

We probe RG models' capability of understanding and using the explanation E in a *contrastive* manner (Sec. 2.2). This section first introduces the corruption types under two categories, then we introduce evaluation metrics, and finally we discuss several SOTA RG models with different neural architectures that we probe.

4.1 Corruption Types

We use verified explanations generated from GLUCOSE T5 model as valid explanations and define two categories of *corruptions* to corrupt the explanation to be logically-invalid and/or grammatically unnatural. We consider three *logical corruptions* that invalidate the logical connection between explanation and response, as well as three *complete corruptions* that break both logical coherence and naturalness of the sentence. Examples covering showing corruption types of a valid explanation are shown in Figure 4, for which "*I found a buyer for the house*" is the antecedent, "*causes*" is the connective, and "*I am so happy*" is the consequent.

Logical Corruptions We consider three ways to invalidate the logic of the explanation: 1) *Swapped* that swaps the antecedent and consequent of the explanation, 2) *Negation* that negates the connective word of the explanation, 3) *Incorrect* that uses an explanation from the same dialogue history-response instance that is rated as incorrect (if any) during the verification.

Complete Corruptions Inspired by Sankar et al. (2019) who design perturbations to apply on dialogue history and analyze sensitivity of RG models by measuring the perplexity of the response, we consider three operations that completely break the naturalness of the explanation: 1) *Shuffle* that randomly shuffles the words of the explanation, 2) *Dropped* that drops 30% of the words uniformly, 3)

Models	Logical Corruption Average [Accuracy/Δ NLL]				Complete Corruption Average [Accuracy/Δ NLL]			
	DD	ED	MuTual	SocialIQA	DD	ED	MuTual	SocialIQA
Inference Probing								
DialoGPT (l2r)	0.57/-0.01	**0.60/0.03**	**0.62/0.03**	**0.64/0.03**	0.71/0.15	0.77/0.25	**0.79/0.22**	**0.87/0.40**
TopicalChat-GPT2 (l2r)	0.49/-0.00	0.50/-0.00	0.49/-0.00	0.50/-0.00	**0.76/0.23**	**0.79/0.24**	0.78/0.24	0.81/0.27
BlenderBot (s2s)	0.46/0.00	0.55/0.02	0.51/0.02	0.50/0.01	0.45/-0.02	0.43/-0.05	0.49/-0.03	0.41/-0.03
BART-base (s2s)	**0.53/0.07**	**0.60/0.19**	0.57/0.07	0.54/0.09	0.36/-0.38	0.41/-0.23	0.43/-0.27	0.43/-0.21
BART-large (s2s)	0.51/-0.03	0.52/-0.01	0.48/-0.06	0.52/0.00	0.49/-0.05	0.55/0.06	0.52/0.01	0.57/0.11
DialoGPT-ft (l2r)	0.50/-0.05	0.39/-0.54	0.44/-0.33	0.43/-0.25	0.63/0.11	0.76/0.24	0.66/0.15	0.78/0.31
BART-base-ft (s2s)	0.59/0.02	0.58/0.01	0.58/0.02	0.60/0.03	0.57/0.04	0.72/0.07	0.59/0.04	0.70/0.09
BART-large-ft (s2s)	0.57/0.02	0.44/-0.01	0.53/0.01	0.48/0.00	0.35/-0.06	0.54/0.02	0.37/-0.04	0.48/-0.00
Human	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0

Table 1: **Inference probing results** for different response generation models on 4 dialogues datasets against two categories of corruptions. Accuracy is the bianry accuracy of giving a lower loss to the valid explanation than the corrupted one and Δ NLL is the average difference of per-token NLL between the loss of a corrupted inference and a valid inference (the more positive the better).

Reversed reverses the ordering of all the words in the explanation.

Evaluation Protocol and Metrics We use two metrics to measure RG models' capability to distinguish valid commonsense causal explanations from invalid explanations. The standard way of modeling P_θ in Equation (1) in generative models is using Maximum Likelihood Estimation (MLE) approach and minimize the conditional negative log-likelihood loss (NLL), i.e., $L(P_\theta, R, H) = -\sum_{i=1}^{m} \log P_\theta(y_i|y_{<i}, x_1, ..., x_n)$.

Since NLL is a direct measure of the probability distribution learned by the models, we use the same NLL measure for probing RG models' behavior. To measure performance of RG models, we directly compare the average per-token NLL when given a valid explanation and when given an invalid explanation to the response.

We first consider binary accuracy of giving a lower loss (higher probability) to the valid explanation than the corrupted one. A random-guessing baseline for the accuracy is 0.5. To further measure how *confident* the model is in determining the validity of commonsense explanations, we also compute the average difference ΔNLL by subtracting the loss of the valid inference from the invalid inference loss. The closer to zero the difference is, the less confident the model is.

4.2 Response Generation Models

We experiment with multiple models from two neural architectures: GPT-2-based (Radford et al., 2019) unidirectional transformer language model and Seq2Seq-based transformer (Vaswani et al., 2017) models. For GPT-2-based models, we use *DialoGPT* that is trained on 147M multi-turn conversation-like exchanges extracted from Reddit (Zhang et al., 2020b) and GPT-2 trained on *TopicalChat* (Gopalakrishnan et al., 2019) as the knowledge-grounded RG model. For seq2seq models, we use BlenderBot (Roller et al., 2021) and BART (Lewis et al., 2020). More details about these RG models are included in Appendix C.

5 Probing Results and Analysis

We present results and findings for our two probing settings using different dialogue RG models across four datasets for which we collected verified explanations. For each human-validated explanation, we generate a corrupted version using one of our six corruption types, and compare the NLL for the probe target according to our two settings.

In Tables 1 and 2, we show both binary accuracy and average difference in NLL for dialogues from four datasets under the two settings and aggregate the six corruption types into two categories. We also sample 5% of the dialogues for human verification under the same two probe settings.

5.1 How Does Probability of Response Change Given Explanations?

All models are insensitive to the relation between explanations and responses. As shown in the left portion of Table 1, we find that when comparing a valid explanation with a *logically corrupted* (LC) one, all models, regardless of left-to-right or seq2seq model architecture, have accuracy around 50-60%, near a random guessing baseline, with extremely small differences in NLL (some even negative). This suggests that the RG models do not understand the causal relation between the explanation and the response since they give similar probabilities to the response when conditioned on a valid explanation and on a incoherent expla-

Models	Logical Corruption Average [Accuracy/Δ NLL]				Complete Corruption Average [Accuracy/Δ NLL]			
	DD	ED	MuTual	SocialIQA	DD	ED	MuTual	SocialIQA
Attribution Probing								
DialoGPT (l2r)	0.46/-0.07	0.47/-0.04	0.48/0.03	0.49/0.00	0.91/1.60	0.93/2.32	0.92/1.90	0.93/2.36
TopicalChat-GPT2 (l2r)	0.57/0.05	0.55/0.10	0.57/0.10	0.55/0.09	**0.97/2.75**	**0.97/3.08**	**0.96/2.93**	**0.96/2.93**
BlenderBot (s2s)	**0.60/0.04**	**0.59/0.05**	**0.60/0.05**	**0.58/0.06**	0.83/0.45	0.87/0.72	0.86/0.58	0.84/0.55
BART-base (s2s)	0.39/-0.19	0.41/-0.14	0.44/-0.10	0.42/-0.13	0.52/0.08	0.50/0.01	0.52/0.14	0.51/0.10
BART-large (s2s)	0.42/-0.15	0.41/-0.19	0.41/-0.18	0.40/-0.18	0.88/1.37	0.91/1.30	0.91/1.40	0.94/1.44
DialoGPT-ft (l2r)	0.43/-0.09	0.41/-0.04	0.47/0.01	0.46/0.00	0.93/2.01	0.96/2.60	0.93/2.22	0.95/2.70
BART-base-ft (s2s)	0.37/-0.16	0.36/-0.14	0.37/-0.19	0.37/-0.13	0.63/0.37	0.77/0.62	0.60/0.26	0.58/0.27
BART-large-ft (s2s)	0.36/-0.28	0.41/-0.13	0.35/-0.30	0.37/-0.23	0.45/0.02	0.83/1.04	0.54/0.30	0.63/0.41
Human	1.0	1.0	0.9	1.0	1.0	1.0	1.0	1.0

Table 2: **Attribution probing results** for different response generation models on 4 dialogues datasets against two categories of corruptions.

nation, while humans can easily identify the valid explanation.

Even gibberish does not change response probability much. Surprisingly, we find that even when corrupting the explanation so completely that it becomes unnatural English, most seq2seq RG models still generate responses with a roughly equal likelihood (left2right models perform better but still lag human performance) as shown in the right portion of Table 1. Sankar et al. (2019) find that the increase in perplexity of the response is tiny when they perturb the dialogue context, but here we find that there might even not be any increase in perplexity when conditioned on gibberish compared to a valid explanation expressed in English, while humans can identify the natural explanation perfectly.

5.2 Can RG Models Attribute Valid Reasons for the Responses?

Logically incoherent attribution confuses the models. Similar to the inference probing setting, for *logically corrupted* one, all models have accuracy around 50-60% and tiny differences in NLL from the left part of Table 2. This indicates that the RG models cannot identify a logically-valid reason for a response from a reason that is similarly natural in terms of grammar but with totally different and invalid logical implications for the dialogue. Humans, from our sampled dialogues, again show much higher accuracy in this setting.

Models can confidently distinguish valid attribution from unnatural ones. For complete corruptions (CC), we find that except for BART, RG models perform much better in identifying a valid explanation compared to a completely corrupted one with most accuracy being close to 1 and relatively larger NLL differences. We conclude that these RG models find generating a valid explanation more natural than completely corrupted ones, which is expected since they are trained to generate natural sentences. However, combining this finding with the previous observation, we find that these RG models can discern unnatural sentences by giving a low probability, but fail to determine the logical validity of the reasons for responses, posing doubts on whether they understand CSR behind a response.

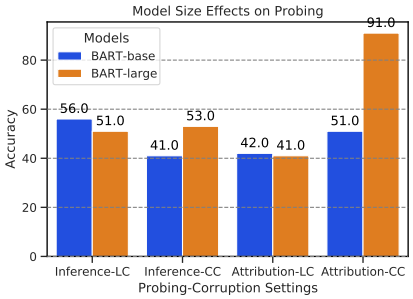

Figure 5: **Model size Effects** on the two probing settings for BART aggregated across four datasets and types of corruptions. We find that except for the attribution setting against complete corruptions, increasing size does not impact much on probing performance.

5.3 Analysis of Probing Results

Unfamiliarity with probing format is not the bottleneck. Since these RG models are *not* trained directly to take additional knowledge as input to generate responses or generate explanations for responses (although explaining happens often in dialogues), these poor results may be due to the probing setup. We thus fine-tune BART-base on 50% of our verified explanations in the same format as our two settings and probe on the rest. We find even when the model is accustomed to the

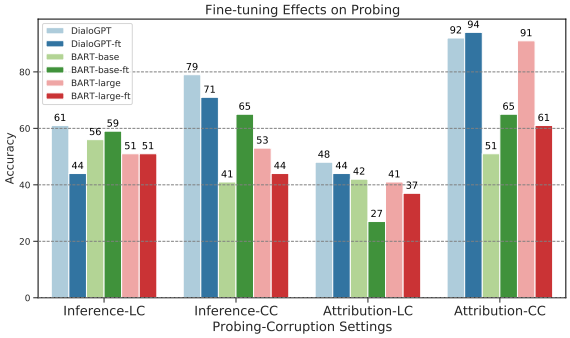

Figure 6: **Fine-tuning (on in-domain dialogues) Effects** on the two probing settings for three models aggregated across four datasets and types of corruptions. We find that in general fine-tuning does not help and sometimes even hurt the probing performance.

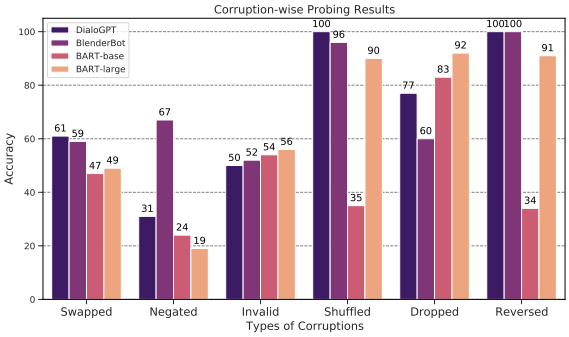

Figure 7: **Corruption results** breakdown on the attribution probing settings aggregated across four datasets.

tasks, the accuracy against logical corruptions for both settings is still around 60%. Although it is possible that with more data the performance can be improved, we also note that training with explanations also makes the model biased to prefer explanations over corrupted ones due to pattern matching. For example, no explanations contain negated connectives, which might be used to gain an advantage unrelated to understanding common sense when compared against negated corruption.

To probe a model that is accustomed to the task but not exposed to explanation patterns, we consider a GPT-2-based (Radford et al., 2019) model trained on TopicalChat (Gopalakrishnan et al., 2019), a knowledge-grounded dialogue dataset. The model is trained on given input of dialogue history concatenated with a knowledge sentence that the response needs to use. We treat the commonsense explanation as the knowledge sentence as they both provide necessary information that leads to the response. We find that the model performs similarly to DialoGPT on our probing setting for logical corruptions, providing evidence that the reason why these RG models cannot identify causal relations behind dialogue responses is not because the model is not used to taking explanation as input.

Model size does not help with understanding common sense. Comparing BART-base and BART-large in Figure 5, we find that except for the attribution setting with complete corruptions, size does not change probing results (even lower accuracy against logical corruptions), indicating that the size of RG model is not the key to understand commonsense explanations for dialogue responses.

Fine-tuning on in-domain dialogues sometimes have opposite effects. Since these RG models are trained on different dialogue datasets that are not necessarily in the same domain as probing dialogues, we also explore the effects of fine-tuning on *in-domain* dialogues, dialogues from the 4 datasets we use for probing. Three pairs of model (before and after fine-tuning) results are shown in Figure 6 and we do not find significant differences. We even find sometimes fine-tuning hurts the probing results, which might be due to models picking up statistical patterns while training on similar dialogues, relying less on "*reasoning*", if any.

Potential biases on certain perturbation types. The observations above are general trends of the models performance, but we also find interesting corner cases indicating potential biases in the models when we breakdown performance for six corruption types shown in Figure 7 on the attribution probing setting. For the *Negated* corruption type, DialoGPT and BART have accuracy around 30%, meaning that for 70% of the time, they *prefer* generating explanations with negated relations in it.

6 Related Work

Commonsense Reasoning The majority of recent CSR benchmarks (Zellers et al., 2018; Talmor et al., 2019; Bisk et al., 2020; Sap et al., 2019; Lin et al., 2021c,a, 2020) test a model's ability to choose the correct option given a context and a question. Recent work also aims to probe models in these tasks to see if reasoning is actually achieved (Richardson and Sabharwal, 2020; Richardson et al., 2020; Zhou et al., 2020; Lin et al., 2021b). Arabshahi et al. (2020) focuses on if-then-because reasoning in conversations and design a theorem prover. In RG, several works have tried

to incorporate commonsense (Zhou et al., 2018; Zhang et al., 2020a) using ConceptNet, a commonsense knowledge graph (Liu and Singh, 2004) to make responses more natural-sounding.

Dialogue Response Generation Recent work focused on fine-tuning large pre-trained transformer models (Radford et al., 2019; Zhang et al., 2020b) on dialogue data. Many dialogue datasets have been collected with different focuses such as incorporating knowledge (Gopalakrishnan et al., 2019; Dinan et al., 2019b), empathy (Rashkin et al., 2019), personality (Zhang et al., 2018) and reasoning (Cui et al., 2020) within dialog systems. There has also been work on combining a variety of datasets to exhibit multiple attributes (Roller et al., 2021).

7 Conclusion

We study commonsense reasoning in dialogue response generation aiming to close the gap between current RG models and human communication abilities. Specifically we formalize the problem by framing commonsense as a latent variable in the RG task and using explanations for responses as textual form of commonsense. We design an explanation collection procedure for RG and propose two probing settings to evaluate RG models' CSR capabilities. We hope our study motivates more research in making RG models emulate human reasoning process in pursuit of smooth human-AI communication.

Acknowledgments

We thank anonymous reviewers for providing insightful feedback along with Brendan Kennedy, Peifeng Wang, and members from INK and JAUNTS lab. This research is supported in part by the DARPA MCS program under Contract No. N660011924033, the Defense Advanced Research Projects Agency with award W911NF-19-20271, NSF IIS 2048211, and NSF SMA 182926.

Ethics and Broader Impact

Our work aims to examine RG model's ability to understand common sense for dialogue responses. Sheng et al. (2021) have found biases in DialoGPT responses and Mehrabi et al. (2021) have found representational harms in common sense resources. We acknowledge that the generated responses from models we use in probing experiments might contain biases. All of the dialogue datasets and models are in English, which benefits English speakers more. We have conducted human verification using Amazon Mechanical Turks. We pay turkers around $14 per hour, well above the highest state minimum wage and engage in constructive discussions if they have concerns about the process. We also give each annotation instance enough time so that we do not pressure annotators.

References

Forough Arabshahi, Jennifer Lee, Mikayla Gawarecki, Kathryn Mazaitis, Amos Azaria, and Tom Mitchell. 2020. Conversational neuro-symbolic commonsense reasoning. *ArXiv preprint*, abs/2006.10022.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019a. The second conversational intelligence challenge (convai2). *ArXiv preprint*, abs/1902.00098.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Ranald D Hansen. 1980. Commonsense attribution. *Journal of Personality and Social Psychology*, 39(6):996.

Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65.

Denis J Hilton and Ben R Slugoski. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*, 93(1):75.

Walter Kintsch and Teun A Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021a. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. 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 1274–1287, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren, and William Cohen. 2021b. Differentiable open-ended commonsense reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4611–4625, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021c. RiddleSense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1504–1515, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. *In EMNLP*.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog 1.8 (2019): 9.*

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

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.

Kyle Richardson, Hai Hu, Lawrence S Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *AAAI*, pages 8713–8721.

Kyle Richardson and Ashish Sabharwal. 2020. What does my QA model know? devising controlled probes using expert knowledge. *Transactions of the Association for Computational Linguistics*, 8:572–588.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. 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.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. "nice try, kiddo": Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767, Online. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020a. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4623–4629. ijcai.org.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang

Liu, and Dilek Hakkani-Tur. 2021. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–132, Singapore and Online. Association for Computational Linguistics.

Pei Zhou, Rahul Khanna, Seyeon Lee, Bill Yuchen Lin, Daniel Ho, Jay Pujara, and Xiang Ren. 2020. Rica: Evaluating robust inference capabilities based on commonsense axioms. *ArXiv preprint*, abs/2005.00782.

A GLUCOSE Detail

GLUCOSE contains human annotations of ten dimensions of causal explanation related to X. Five of the dimensions are about events and states happening before X and five are about those happening after X. Specifically, inspired by cognitive psychology, the authors of GLUCOSE consider events, emotions, location states, possession states, and other attributes as the five dimensions of causal inferences. According to their evaluation, the best-performing model is T5 (Raffel et al., 2020) (with 770M parameters) with the input formulated as $\#d : S^*[X]$, where d is the dimension and $S^*[X]$ is the story S with sentence X surrounded by asterisks". An illustrated example of inputs of outputs of the T5 model trained on GLUCOSE is shown in Figure 8.

To adopt the T5 model trained on GLUCOSE to our task: generating explanations about what might cause producing a response given a dialogue history, we append the dialogue history turns together, enclose the response we are interested in explaining with asterisks, and fill in dimension number 1 to 5 to ask for what event, emotion, location, possession, and attribute could cause, motivate, or enable the response. In other words, we formulate our queries as $\#d : H^*[R]$, where d is the dimension 1 to 5 and $H^*[R]$ is the dialogue history H appended with the response R surrounded with asterisks.

B Verification Detail

Table 3 shows the general pass rate for each criterion and the overall pass rate (need to pass all three criteria). Figure 9 shows distribution of valid and invalid explanations separated by the five causal dimensions. We find the explanations about a *location* state that causes the response have a lower valid rate (13%) than others. This might be due to that in some dialogues the location information is not important in explaining the response and thus it is difficult to come up with a plausible reason about a location that leads to the response. All other dimensions have a similar rate of 25-30%.

To ensure annotation quality, the workers first need to pass a *qualification test* (QT) that tests their understandings of the criteria to be able to do our main annotation tasks. Our QT contains eight questions, each contains a dialogue history, a response, and an explanation and we ask them to choose whether this explanation satisfies a specific criterion from the three above. The eight questions

Input:
"*1: Today is your lucky day Casey. What? What do you mean? I found a buyer for the house. *Oh Boy! I'm so happy. I knew hiring a real estate agent was a good idea.**"

Output:
1→ Event: I found a buyer for the house >Causes/Enables> I am so happy
2→ Emotion: I want (s) to make sure my house is sold >Motivates> I hire a real estate agent
3→ Location: I am in a house >Enables> I am so happy
4→ Possession: I possess (e s) a house >Enables> I am so happy
5→ Attribute: I have an agent >Enables> I knew hiring a real estate agent was a good idea

Figure 8: **Example input and output** from GLUCOSE-trained T5 model on a dialogue.

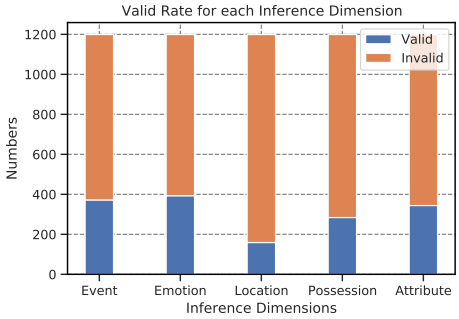

Figure 9: **Verification results** on 6k explanations from 1,200 dialogeus separated by five causal dimensions. The valid rates are 31%, 33%, 13%, 24%, and 29% for the five dimensions.

are formed into 4 pairs each consisting of a *training* question and a *testing* question and each pair focuses on the same criterion. For the *relevance* and *non-trivial* criteria, we have one pair for each and for the *plausible* criterion, we have two pairs since it is trickier to determine than the other two. We provide the right answer with explanation for the training question whether they answer it correctly or not and use the testing questions for assessment of their understanding.

C Model Detail

DialoGPT extends the GPT-2 architecture that adopts the generic transformer language model (Vaswani et al., 2017) by training on 147M multi-turn conversation-like exchanges extracted from Reddit. We use the 345M DialoGPT

Criterion	*Passing Rate*
Relevant	55%
Non-trivial	73%
Plausible	37%
All three	26%

Table 3: Passing rates for the three criteria and the overall valid rate (need to pass all three) from verification.

Figure 10: Example of qualification test question with shown explanation.

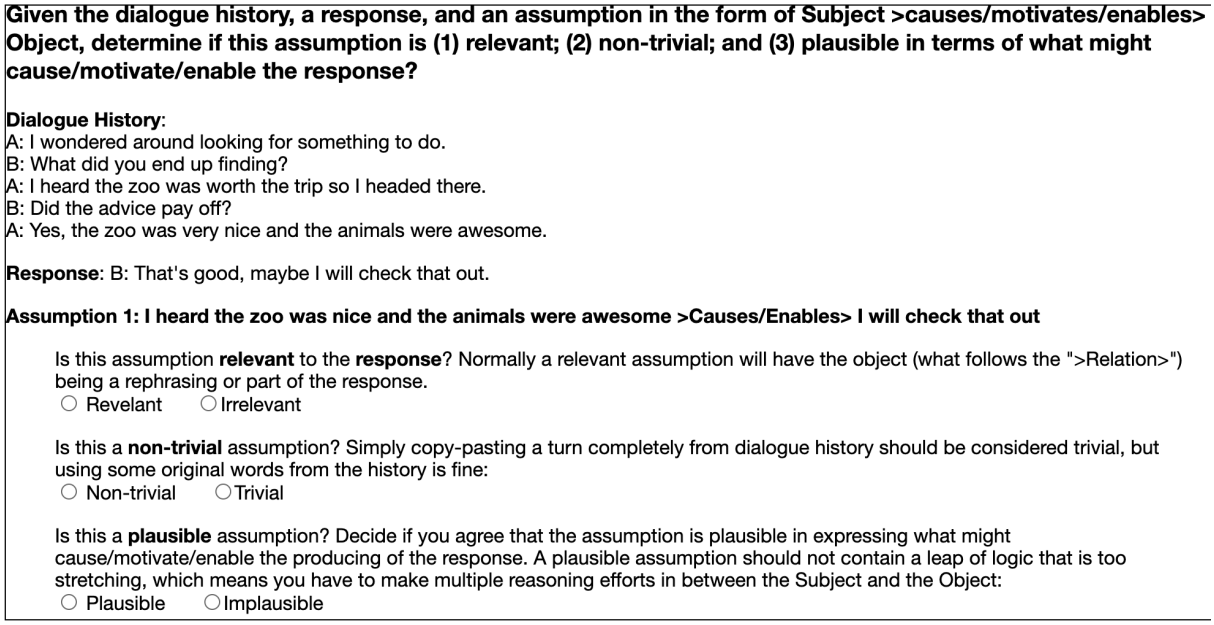

Figure 11: Example of the verification task question with three criteria for verifiers to choose.

model[2] (Zhang et al., 2020b).

BlenderBot is proposed by Roller et al. (2021) using a standard seq2seq transformer architecture (Vaswani et al., 2017). The model aims to blend multiple conversational skills. Human evaluations show their best models beat existing approaches in multi-turn dialogue in terms of engagingness and humanness. We use the 400M BlenderBot model distilled from 2.7B parameter model[3].

BART is proposed by Lewis et al. (2020) using a standard seq2seq architecture with a bidirectional BERT encoder and a left-to-right GPT decoder. It uses denoising pre-training objectives and has shown to outperform previous models in multiple language generation tasks including ConvAI2 (Dinan et al., 2019a). We use both BART-base and BART-large with 139M and 406M parameters, respectively[4].

[2] https://huggingface.co/microsoft/DialoGPT-medium

[3] https://huggingface.co/facebook/blenderbot-400M-distill

[4] https://huggingface.co/models?search=bart

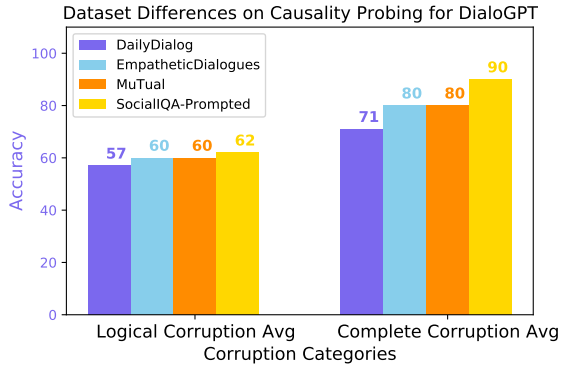

Figure 12: **Dataset differences** on the causality probing setting for DialoGPT model, we find that there is no drastic differences in probing performances across four datasets for logical corruptions, i.e., the conclusion that RG model fails to understand causality holds true for all datasets. We see difference in accuracy ranging from 70% to 90% for complete corruptions.