CausalDialogue: Modeling Utterance-level Causality in Conversations

Yi-Lin Tuan[♣] **Alon Albalak**[♣] **Wenda Xu**[♣] **Michael Saxon**[♣] **Connor Pryor**[♦]
Lise Getoor[♦] **William Yang Wang**[♣]
[♣] University of California, Santa Barbara, [♦] University of California, Santa Cruz
{ytuan, alon_albalak, wendaxu, saxon, william}@cs.ucsb.edu
{cfpryor, getoor}@ucsc.edu

Abstract

Despite their widespread adoption, neural conversation models have yet to exhibit natural chat capabilities with humans. In this research, we examine user utterances as *causes* and generated responses as *effects*, recognizing that changes in a cause should produce a different effect. To further explore this concept, we have compiled and expanded upon a new dataset called **CausalDialogue** through crowdsourcing. This dataset includes multiple cause-effect pairs within a directed acyclic graph (DAG) structure. Our analysis reveals that traditional loss functions struggle to effectively incorporate the DAG structure, leading us to propose a causality-enhanced method called Exponential Maximum Average Treatment Effect (ExMATE) to enhance the impact of causality at the utterance level in training neural conversation models. To evaluate the needs of considering causality in dialogue generation, we built a comprehensive benchmark on CausalDialogue dataset using different models, inference, and training methods. Through experiments, we find that a causality-inspired loss like ExMATE can improve the diversity and agility of conventional loss function and there is still room for improvement to reach human-level quality on this new dataset. [1]

1 Introduction

Over time, broadly-defined dialogue models have become increasingly prevalent in society and been integrated in a range of domains from speech assistants and customer service systems to entertainment products, such as video games, where the non-playable characters (NPCs) engage in conversation with players. A core goal of training chatbots is enabling them to interact with humans naturally (Vinyals and Le, 2015; Sordoni et al., 2015). This includes, but is not limited to: considering both the machine and addressee's personalities (Li et al., 2016b), diversifying responses to be less generic (e.g., the same response "I don't know." is often produced in a traditional setting for different dialogues) (Li et al., 2016a), grounding on external knowledge to be informative (Ghazvininejad et al., 2018), and tailoring responses specific to nuanced differences in conversation.

To the best of our knowledge, no recent studies have prioritized the ability to tailor responses for minor differences in conversations. This problem is currently implicitly approached by training models with larger scale or cleaner conversation data (Zhang et al., 2020; Roller et al., 2021; Thoppilan et al., 2022) or involving human-in-the-loop (Li et al., 2016c; Jaques et al., 2020). However, the effectiveness of these methods is unclear, the online rewarding scheme can be expensive, and a suitable testbed for evaluating the solution to this problem has not yet been identified.

To this end, we propose a benchmark to foster research in tailoring responses for nuanced differences in conversations by answering the question "*if all prior turns are the same, but the last turns in two conversations are semantically different, how should future turns differ?*" We call this concept *Agility* and model it as the *utterance-level causes and effects* in dialogue response generation, where the causes are the slightly different prior turns and the effects are the resulting future turns.

We introduce **CausalDialogue**, a dataset seeded by expert-written dialogues containing branching dialogue paths, which we further expand in terms of scale and linguistic abundance with crowdsourcing. Each conversation is represented as a directed acyclic graph (DAG) for ease of storage and causal analysis (Pearl, 2009) as shown in Figure 1. As conversations progress, each utterance can elicit multiple responses, resulting in a split of the conversation (branch-splitting). Alternatively, multiple conversations that share a common starting point

[1] Our code and dataset are available at https://github.com/Pascalson/CausalDialogue

Findings of the Association for Computational Linguistics: ACL 2023, pages 12506–12522

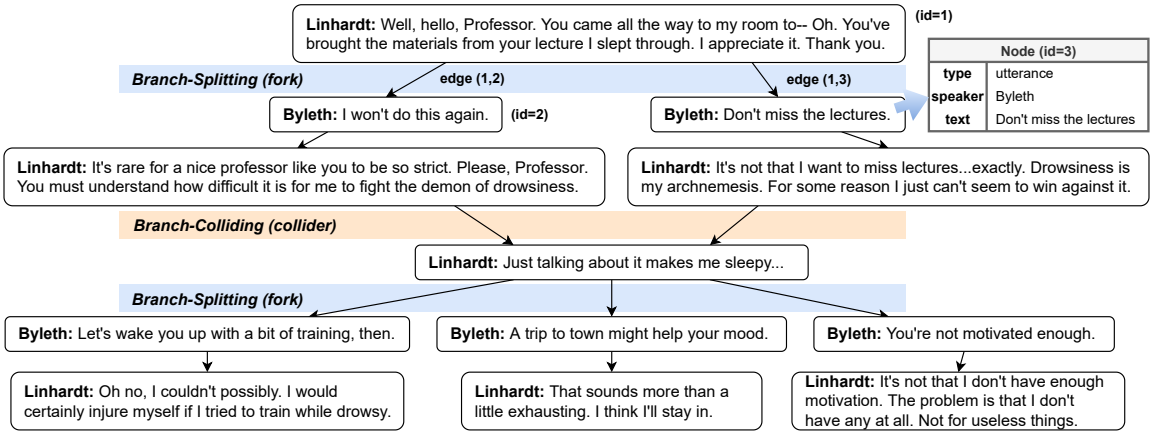

Figure 1: A dialogue DAG example in the new dataset CausalDialogue. As the conversation progress, each utterance can be continued with multiple responses (branch-splitting; fork); meanwhile, the same root dialogue with different middle turns can be continued by the same response (branch-colliding; collider).

may sometimes lead to the same response, even if the middle exchanges differ (branch-colliding). Due to the DAG structure of CausalDialogue, it is ideal for aiding research on response generation that requires abundant *IF*-bases, for instance, causal inference and offline reinforcement learning, which may improve the response generation quality for nuanced differences in conversation.

To provide a benchmark for future work on the CausalDialogue dataset, we conduct experiments with various setups. We include both decoder-only and encoder-decoder transformer models pre-trained on either common or dialogue-specific corpora, various inference methods, conventional training losses, and a newly proposed loss, Exponential Maximum Average Treatment Effect (ExMATE), inspired by Average Treatment Effect (Holland, 1986; Imai et al., 2008), which is a method commonly used to approximate the causal effect of a treatment and its outcome. In this benchmark, we show that existing methods are not sufficient in tackling the agility issue, and a simple causality-inspired loss demonstrates improvement.

Our key contributions are:

- A novel dataset, CausalDialogue, including both expert-written scripts and crowd-sourced utterances with a DAG structure.
- A new training loss, ExMATE, for considering the utterances as causes and effects in a dialogue, inspired by the average treatment effect in research on causal inference.
- A benchmark with experiments showing that existing methods need improvement on the agility problem, and a causality-inspired method can be a promising direction to improve it.

2 Related Work

Chit-Chat Dialogue Datasets. To boost the research of dialogue models, the community has collected dialogues based on scripts written by experts from movies (Danescu-Niculescu-Mizil and Lee, 2011; Banchs, 2012; Lison and Tiedemann, 2016), TV shows (Poria et al., 2019; Tuan et al., 2019; Yu et al., 2020; Rameshkumar and Bailey, 2020), and education purposes (Li et al., 2017b; Cui et al., 2020). For abundant diversity and real-life scenarios, Ritter et al. (2011); Wang et al. (2013); Lowe et al. (2015); Pasunuru and Bansal (2018) collected datasets based on the publicly available data from social media and forums. Additionally, previous work has explored the idea of collecting data through crowd-sourcing with added constraints to improve its quality or expand label types. For example, Zhang et al. (2018) constructed a dataset with workers imitating a given personal profile. Rashkin et al. (2019) built a dataset by explicitly asking workers to show their empathy during a conversation. Urbanek et al. (2019); Narayan-Chen et al. (2019); Ammanabrolu et al. (2021) created datasets with the assistance of game structures, so the purpose of the dialogue is to complete a mission or collaborate with other agents. Finally, recent work by Dou et al. (2021) collected branches of dialogues for 120 self-written prompts to create dialogue trees. Compared to previous studies, our dataset is a fusion of the scripts written by experts and responses created by crowd-sourcers with manual correction, granting it high quality, linguistic abundance, and extensive metadata. Additionally, our dataset includes both branch-splitting and

	CausalDialogue	TV Series	MultiTalk	DailyDialog	PersonaChat	LIGHT
Branches	✓(DAG)	✗	✓(Tree)	✗	✗	✗
Profiles	✓	✓	✗	✗	✓	✓
Situated	✓	✓	✓	✗	✗	✓
Expert involved	✓	✓	✗	✓	✗	✗

Table 1: Compared to current widely used datasets, CausalDialogue contains the utterance-level graph structure and meanwhile has the features of diverse speaker profiles, descriptive situations, and high quality scripts written by experts. The referring dialogue generation datasets are: TV series (Tuan et al., 2019), MultiTalk (Dou et al., 2021), DailyDialog (Li et al., 2017b), PersonaChat (Zhang et al., 2018), LIGHT (Urbanek et al., 2019).

branch-colliding instances, which has led us to classify dialogues as directed acyclic graphs (DAGs) instead of just sequences or trees.

Dialogue Generation Training Objectives. To train a dialogue response generation model, methods have been developed from maximizing the likelihood between the hypothesis and the ground-truth (Vinyals and Le, 2015; Serban et al., 2016), guiding responses to match a higher reward in reinforcement learning (Li et al., 2016d), and allowing for extra latent variables to optimize divergence through variational autoencoder (Zhao et al., 2017) or generative adversarial networks (Li et al., 2017a; Tuan and Lee, 2019). Recent works have introduced the concept of causal inference (Holland, 1986; Imai et al., 2008; Pearl, 2009; Cunningham, 2021) into generative adversarial network-based (Zhu et al., 2020) and multiple-stage inference based dialogue generation model (Tuan et al., 2020). Utterance-level offline reinforcement learning has also been explored to optimize response generation (Jaques et al., 2020; Verma et al., 2022). However, they were studied by expanding available sequence data with imaginations. Now by providing a chit-chat dialogue DAG structure that is enriched with multiple if-else cases, CausalDialogue can be studied for causal inference and offline reinforcement learning on response generation. We also propose a new method called ExMATE for better optimizing a response generation model on the DAG data structure.

3 CausalDialogue Dataset

In this section, we introduce **CausalDialogue**, a novel dataset that includes chit-chat conversations in a Conversational Directed Acyclic Graph (DAG) data structure. This structure allows for the natural inclusion of various dialogue flows, such as forks (branch-splitting) and colliders (branch-colliding) (Pearl, 2009). Our goal is to offer researchers a valuable resource for studying the complexities of human conversation and advancing the understanding of causal inference in dialogue.

To create CausalDialogue, we sourced expert-written dialogues from a role-playing game (Section 3.1) and expanded upon them with Amazon Mechanical Turk (MTurk)[2] and manual correction (Section 3.2). By using our fused collection method, the dataset contains high-quality, engaging conversations with abundant linguistic usage that imitates daily life.

3.1 Data Collection

CausalDialogue is derived from the English scripts of the popular role-playing game (RPG) *Fire Emblem: Three Houses*, which we sourced from the fandom wikipedia[3] under the GNU Free Documentation License(GFDL)[4]. This RPG is well-known for its diverse, story-driven conversations, which mix the interactions of approximately 40 main characters. In this game, players have the ability to shape the narrative by making choices that lead to different dialogue branches.

Table 2 lists the statistics of the two main types of the crawled data, which are already divided in the raw scripts. We name the first conversation type ORI.-2S, which are mostly dialogues between two speakers, and generally include conversations about interpersonal relationships. We name the second conversation type MULTI, which are dialogues between two or more speakers, and usually describe the current status of the story line. In the following sections, we will introduce the DAG structure to better describe the dataset, as well as how we obtained additional examples from crowd-sourcing to create the EXPANSION to these expert-written scripts.

[2] https://www.mturk.com

[3] https://fireemblem.fandom.com/

[4] https://fireemblem.fandom.com/wiki/Fire_Emblem_Wiki:Copyrights

Data Partition	Ori.-2S	Multi	Expan.	Total
# Dialogues†	794	1528	623	2322
# Branches	1633	1298	2378	4866
# Utterances	33247	13858	15728	46109
# Speakers	41	47	39	51
Avg. utts/dial.	17.0	51.4	5.6	26.8
Avg. words/utt.	18.4	17.8	11.8	16.5
Avg. utts/spk.	801.6	268.8	402.8	878.4

Table 2: The statistics of CausalDialogue dataset, where the columns ORI.-2S and MULTI are the crawled and cleaned original scripts and the column EXPANSION is from crowd-sourcing. In total, there are 3457/741/715 dialogues for train/validation/test sets. † indicates that for EXPANSION set, 623 is the number of initial dialogues that are parts of the 794 Ori.-2S dialogues, so the total number of dialogues is 2322.

Dialogue DAGs. Conventional linear dialog data structures can be challenging to create when dealing with *forks* and *colliders*, as they can lead to ambiguity in the form of duplicated utterances and split responses. To address this issue, we propose using a conversational DAG to maintain the fidelity of the dialog. We convert each textual conversation into a DAG, as demonstrated in Figure 1. Formally, each node is a dictionary containing the text type (utterance/scene information), text, speaker, and its own id in the dialogue. A directed edge (i, j) then indicates that a node with id j is a possible response to the node with id i. Saving dialogues as DAGs may introduce some complexity, but it also offers numerous benefits. For example, it reduces the memory required to save each dialogue branch independently, enables a natural visualization of the multiple possible dialogues flows, and fosters the survey of causality on dialogue utterances.

Speaker Profiles. Prior work has shown the relationship between personality and language uses in conversations (Mairesse et al., 2007). To ensure consistent personality, as well as to diversify linguistic features across speakers, we leverage the speaker profiles during the data collection process. The resulting CausalDialogue dataset comprises 41 main speakers who have been thoughtfully crafted by the game's developers. These speakers possess diverse backgrounds, perspectives, and interests, and their characteristics are both human-like and distinct. These speaker profiles are simplified for collecting the EXPANSION partition to reduce workers' cognitive load, and a set of examples are provided in Appendix A.1. Compared with the speaker profiles in CausalDialogue, previous works have provided limited information (e.g. "I have a dog.") (Zhang et al., 2018; Urbanek et al., 2019), or have a significantly smaller number of speakers (Poria et al., 2019; Tuan et al., 2019)

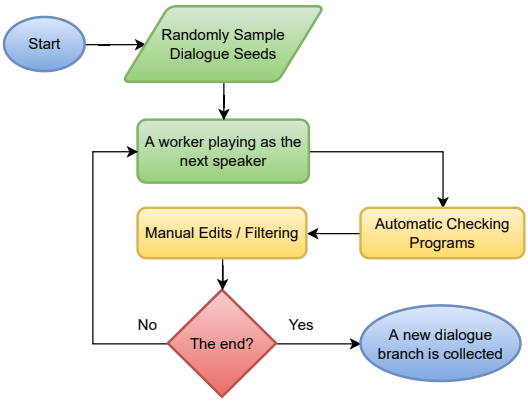

Figure 2: The flowchart of our strategy for data expansion with crowd-sourcing.

3.2 Data Expansion

In order to increase the breadth and scope of our dataset, we propose utilizing a crowd-sourcing approach to add more diverse and current language as shown in Figure 2 (More details in Appendix A.2).

Initial Dialogue Selection. We first randomly select 1,200 partial dialogues from the ORI.-2S partition, which is of higher quality after our manual inspection. This can result in more stable quality when crowd-sourcing responses.

Expansion Collection. Each initial dialogue along with the continuing speaker profile is presented to 3 workers on MTurk to write the next utterance. A new branch of continued dialogue will then be presented to another 1-2 workers playing as another speaker to gather another round of responses. We repeated this process three times and collect a total of about 13,000 written utterances. Table 2 lists the detailed statistics of the expanded data in the column EXPANSION. Note that the statistics of EXPANSION in Table 2 include the initial dialogues. Figure 3 shows an DAG representation of an expanded example.

Quality Control. We adopt three strategies to control for dialogue quality. First, we asked the workers on MTurk to annotate if they regard a dialogue as already completed or having too specific of details to continue. The purpose of the first stage of quality control is to identify conversations which cannot be continued, either because the conversa-

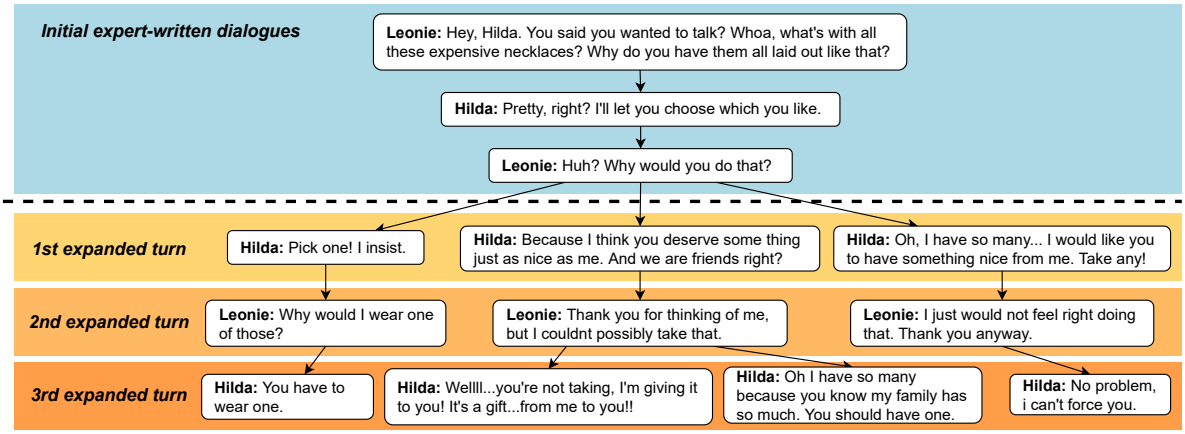

Figure 3: A dialogue example of the EXPANSION partition in CausalDialogue.

tion has already concluded or because the workers are lacking enough information about the world to continue the conversation. Second, we used an off-the-shelf model[5] to label potential ethical issues inside the collected utterances for reference in the next step. Finally, we invited real players of the game and machine learning researchers to manually check all the utterances by their fluency, coherence, and ethics as well as referring to the labels from the previous two steps to ensure the final EXPANSION partition is of high quality.

4 Task Definition

In this work we consider a conversation among two or more speakers. At each time step t, a speaker s_t takes their turn with an utterance u_t. The goal, as in conventional response generation, is to train a model parameterized by θ that can predict a plausible next utterance given the speakers and utterances in prior turns as:

$$u_{t+1} \sim P_\theta(\cdot|s_1u_1, s_2u_2, ..., s_tu_t, s_{t+1}) \,. \quad (1)$$

Distinct from prior conversation datasets, CausalDialogue has multiple dialogue branches. If we consider each branch as an independent conversation (flatten the branches), many conversations will have large overlaps and thus bias the dataset. We consider this point and extract triples (DH, x, y) from CausalDialogue. To simplify notations for following sections, we denote s_tu_t as x, $s_{t+1}u_{t+1}$ as y and DH is the dialogue history $s_1u_1, s_2u_2, ..., s_{t-1}u_{t-1}$. The key idea is that for a DH, we will not extract duplicated pairs (x, y), but x or y itself can be shared.

[5] https://github.com/unitaryai/detoxify

The CausalDialogue response generation task is therefore defined as finding a possible turn-taking speaker and their response given the dialogue history DH with an utterance cause x.

$$y \sim P_\theta(\cdot|DH, x) \,. \quad (2)$$

The sequences $x = x_1x_2...x_i...x_{|x|}$ and $y = y_1y_2...y_j...y_{|y|}$, where x_i and y_j are tokens, and $|x|$ and $|y|$ are the length of the sequences x and y respectively.

4.1 Agility

While the above task definition resembles the standard dialogue generation setting with the exception of speaker prediction and conversation overlaps, our primary interest lies in tailoring responses to minor differences in conversation history. We refer to this concept as *Agility*, where a minor difference in conversations can be a shared DH with different continuation x.

To quantify the idea of agility, we propose a new metric with the following idea: If the predicted next utterance y and the previous turn x has causal-effect relationship (i.e., $x_1 \rightarrow y_1$ and $x_2 \rightarrow y_2$), we anticipate that it is less likely that y_2 is caused by x_1. The newly proposed metric, named confidence causal-effect (CCE) is formally defined as:

$$CCE = E_{(x,y)\in D,(x,y')\notin D,(x',y')\in D} [PPL_\theta(y'|DH, x) - PPL_\theta(y|DH, x)] \,, \quad (3)$$

where PPL refers to perplexity. Note that CCE is not a metric that stands by itself and needs to refer to PPL at the same time. That is, given a similar PPL score, a model with higher CCE score is better.

Additionally, it is important to acknowledge that the concept of agility has been indirectly incorporated into conventional dialogue generation models and evaluation metrics, but it has not been specifically examined in isolation. Our newly introduced dataset and CCE metric can be seen as an initial step towards addressing this aspect.

5 Methods

In this section, we describe how conventional generative models can be used and propose a simple yet effective approach to model causal effect.

5.1 Maximize Likelihood Estimation

An often used method to train a conditional sequence generation model is minimizing the negative log likelihood (Vinyals and Le, 2015; Serban et al., 2016). The loss function is as following:

$$L_{MLE} = \mathop{E}_{(DH,x,y)\sim P_D} \sum_{j=1}^{|y|} -\log P_\theta(y_j|DH, x, y_{1\ldots j-1}), \tag{4}$$

where P_D represents the data distribution. Since the duplication of dialogue history is already taken in to account in our task definition (Section 4), this MLE method can be seen as the recently proposed dialogue tree model (Dou et al., 2021). However, this function only models a part of the cause-effect relationship between the condition and the output sequence. This neglect may lead to a more vague predicted probability distribution of the output, thus generating less agile responses.

5.2 Maximize Average Treatment Effect

To explicitly model the causal effect in a conversation, we propose the Exponential Maximum Average Treatment Effect (ExMATE), taking into account the treatment effect in causal inference (Pearl, 2009). The treatment effect, denoted by δ, is defined as the difference between the outcome under treatment $I = 1$, represented by $\mathcal{O}^{I=1}$, and the outcome under treatment $I = 0$, represented by $\mathcal{O}^{I=0}$. This measures the variation in outcomes when an event I is present or absent. A higher value of δ indicates that the event I is more likely to be a true cause of the outcome. Conversely, a small value of δ suggests that the event I is unlikely to be a cause of the outcome and may only be correlated. We aim to utilize this characteristic in dialogue generation modeling to ensure that a preceding utterance can be considered the genuine cause of the predicted response.

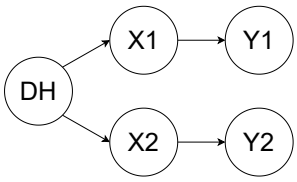

Figure 4: The graphical model of fork-like DAG considered in our proposed ExMATE loss.

We consider the *fork-like* DAGs (as shown in Figure 4) existing in a dataset such as Figure 1 and Figure 3. Without loss of generality, in a binary case, this type of DAG involves two triples that share the same DH and can be simplified as having nodes DH, X_1, X_2, Y_1, and Y_2. Here we use (X_1, Y_1) and (X_2, Y_2) to denote two possibilities of (x,y) after DH. We take $I = 1$ as choosing the branch X_1, and $I = 0$ as choosing an alternative branch X_2. Therefore, a traditional definition of the treatment effect $\delta_i = |\mathcal{O}_i^{I=1} - \mathcal{O}_i^{I=0}|$ for the i-th example in this type of DAG can be rewritten as:

$$\delta_i \triangleq \mathop{E}_{\substack{X_1\sim P_D(\cdot|DH_i),\\ X_2\sim P_D(\cdot|DH_i),\\ X_1\neq X_2}} |\mathcal{O}_i^{X_1} - \mathcal{O}_i^{X_2}|, \tag{5}$$

where $\mathcal{O}_i^{X_1}$ or $\mathcal{O}_i^{X_2}$ is the outcome of an oracle given X_1 or X_2 as the input.

Since the outcome of a dialogue model is hard to be mathematically described only by an input X, we instead utilize the uncertainty of predicting the pair (x, y) by a model θ. We abuse the notation $\mathcal{O}_i$ here and redefine it as,

$$\mathcal{O}_{i,Y_1}^{X_1} \triangleq P_\theta(Y_1|DH, X_1). \tag{6}$$

After formulating a dialogue generation problem as utterance-level causal analysis as above, we apply the Average Treatment Effects (ATE) (Holland, 1986) to conversational DAGs, which is defined as

$$\begin{aligned} ATE &\triangleq E_i[\delta_i] = E_i[\delta_{i,Y_1} + \delta_{i,Y_2}] \\ &= E_i[\mathcal{O}_{i,Y_1}^{X_1} - \mathcal{O}_{i,Y_1}^{X_2} + \mathcal{O}_{i,Y_2}^{X_2} - \mathcal{O}_{i,Y_2}^{X_1}]. \end{aligned} \tag{7}$$

Recall that our goal is to strengthen the cause-effect relationship of each pair, (X_1,Y_1) and (X_2,Y_2) in the binary case. This can be taken as maximizing the defined ATE in Equation 7 with respect to the model parameters θ.

Model	Loss	Inference	Fluency				Diversity		Agility	Identity
			PPL (↓)	BLEU1 (↑)	2 (↑)	4 (↑)	Dist1	Dist2	CCE (↑)	Acc (↑)
Human Written Responses			1.2	48.9	34.0	25.9	1.70	11.1	Inf	100.0
DG	MLE	Greedy Search	18.9	11.2	4.47	0.84	0.73	3.42	2.33	32.51
DG	MLE	Softmax (T=0.5)	18.9	17.0	6.43	1.17	1.12	9.09	2.33	30.97
DG	MLE	TopK (K=10)	18.9	15.7	5.34	0.81	1.37	13.57	2.33	27.65
DG	ExMATE	Greedy Search	19.0	10.7	4.26	1.05	0.79	3.65	2.68	32.18
DG	ExMATE	Softmax (T=0.5)	19.0	15.5	5.70	1.06	1.25	9.71	2.68	31.18
DG	ExMATE	TopK (K=10)	19.0	13.5	4.47	0.67	1.52	14.44	2.68	28.16
T5	MLE	Greedy Search	15.4	5.80	2.52	0.58	1.11	4.37	1.39	75.64
T5	MLE	Softmax (T=0.5)	15.4	12.7	5.06	0.97	1.77	10.91	1.39	74.66
T5	MLE	TopK (K=10)	15.4	14.1	5.09	0.82	2.07	15.49	1.39	72.79
T5	ExMATE	Greedy Search	15.4	5.66	2.46	0.55	1.10	4.06	1.50	75.76
T5	ExMATE	Softmax (T=0.5)	15.4	12.6	5.02	1.00	1.72	10.73	1.50	74.80
T5	ExMATE	TopK (K=10)	15.4	14.1	5.06	0.80	2.06	15.67	1.50	72.83

Table 3: The test results on CausalDialogue of different fine-tuned backbone models (DialoGPT (DG) and T5), inference methods (Greedy Search, Softmax, TopK), and loss functions (MLE and ExMATE). Using ExMATE loss enhances the agility aspect of dialogue generation models without compromising their fluency ratings.

Therefore, we substitute the $\mathcal{O}^X_{i,Y}$ term in Equation 7 with its definition stated in Equation 6 and derive:

$$\begin{aligned}&\arg\max_\theta ATE = \\ &\arg\max_\theta \underset{(X_i,Y_i)\sim P_D(\cdot|DH)}{E} P_\theta(Y_i|DH,X_i) \\ &\qquad - \underset{\substack{X_i\sim P_D(\cdot|DH), Y_j\sim P_D(\cdot|DH) \\ (DH,X_i,Y_j)\notin D}}{E} P_\theta(Y_j|DH,X_i)\,.\end{aligned} \tag{8}$$

To stabilize the training, we modify it with logarithmic and exponential terms and call it the ExMATE loss function. Formally, it is written as:

$$\begin{aligned}&L_{ExMATE} = \\ &\underset{\substack{(DH,x,y)\sim P_D, \\ x_c\sim P_D(\cdot|DH), \\ (DH,x_c,y)\notin D}}{E} -\log P_\theta(y|DH,x) + \exp(\log P_\theta(y|DH,x_c))\,.\end{aligned} \tag{9}$$

The intuition for this change is that without $\exp(\cdot)$, the gradient of the second term will dominate the loss function, since $\log(u)$ has much larger gradient for u close to 0 than u close to 1 and an $\exp(\cdot)$ term can linearize it.

Overall, the idea of ExMATE is to maximize the response generation model's causal effects given a specific X_i (or (DH, x)) as the current cause. At the end, we found that this ATE-inspired approach turns out to be a combination of MLE and a subtraction of specific negative samples. This formulation shares a similar concept with negative sampling and contrastive learning (Goldberg and Levy, 2014; Chen et al., 2020), but has different example selection scheme and is not applied on the embedding space. With this method, we are interested in the research question: *Will a model trained on the CausalDialogue dataset be affected when using a causality-inspired loss?*

6 Experiments

We provide a preliminary benchmark for CausalDialogue with often used methods and a naive causality-inspired loss. We fine-tuned two types of pretrained language models based on transformers (Vaswani et al., 2017): decoder-only architecture, DialoGPT (Zhang et al., 2020) and encoder-decoder architecture, T5 (Raffel et al., 2020), by the conventional MLE loss and the proposed ExMATE loss, and inferred by various sampling methods. We evaluate three aspects of the generated responses: Fluency (perplexity (PPL) and BLEU (Papineni et al., 2002)), Diversity (Distinct n-grams Dist1 and Dist2 (Li et al., 2016a)), and our proposed Agility (CCE) in Section 4.1. Furthermore, we use accuracy to evaluate if the speaker for a given turn is correctly predicted as the one in the human written responses (Identity Acc). More details are in Appendix A.4 and A.6.

6.1 Results

The test results of human written responses and models trained and inferred by different setups are listed in Table 3. We provide details of how we evaluate human written responses in Appendix A.5.

[Backbone Models] We observe that our trained T5 model is generally better than DialoGPT model, as evidenced by the significant difference in PPL and Identity Acc between them. **[Inference Methods]** We observe that Softmax and TopK can

Dialogue History (DH)		Lysithea: Oh, hey. It's you. Going for a walk again today? Ignatz: No, I'm on cooking duty today, and I have to head into town for some groceries.	
Cont. Conv (x)		**case1** Lysithea: That sounds like quite a task!	**case2** Lysithea: Would you like some company?
response(y)	**MLE+S**	Ignatz: You're right.	Ignatz: Yes, I'd love to.
	MLE+T	Ignatz: I'll go for a run.	Ignatz: No, I'm not.
	ExMATE+S	Ignatz: I'm not sure I can handle it.	Ignatz: Yes, I'm a bit tired of the whole thing.
	ExMATE+T	Ignatz: Well, I have to find a new name for myself.	Ignatz: No! It's just me and I.
Cont. Conv (x)		**case3** Lysithea: Allow me to join you so I can help.	**case4** Lysithea: All on your own? Hm, I'd better go with you. I'd be worried if you went by yourself.
response(y)	**MLE+S**	Ignatz: Oh, I'm on cooking duty.	Ignatz: I'm just a bit worried about you.
	MLE+T	Ignatz: I'm going to help cook.	Ignatz: Yeah, no, not really! I can't help you, but I'm just going to help you out.
	ExMATE+S	Ignatz: I'm not working, but I'll help you.	Ignatz: I've got no choice. I'll just have to go out alone.
	ExMATE+T	Ignatz: Oh. I'm sorry, I couldn't be there for you.	Ignatz: Is it okay?

Table 4: Generated responses by our trained models, T5 models trained by MLE or ExMATE inferred by Softmax(S) or TopK(T) sampling methods, given a shared dialogue history but different branches. Using ExMATE loss generally produces more diverse and agile responses.

achieve better results than greedy search in this dataset, as evidenced by their BLEU and Distinct-N scores. The reason is similar to the conventional generic response problem in open-domain dialogue generation (Li et al., 2016a; Tuan and Lee, 2019), since in a DAG, a (DH, x) pair have multiple y as references, causing even an ideal probability distribution to have high entropy. **[Loss Functions]** We find that ExMATE improves MLE with better diversity, agility, and identity accuracy, while maintaining similar fluency scores. This meets our expectation that ExMATE should not deteriorate the MLE's ability in training a model while maximizing the potential causal effect in response prediction. This result empirically shows that the causal effect can help to increase diversity and predict the turn-taking speaker as well. Finally, compared to the evaluation results of human written responses (a hard-to-reach upper bound), current methods still need improvement, except for diversity scores.

6.2 Human Evaluation

We randomly sample 100 dialogues, present each example to three workers on MTurk and ask them score the three dimensions, agility, coherence, and informativeness, scaling from 1 to 5. The evaluation form is provided in Appendix A.3. For each example, we present one shared dialogue history with two branches and the corresponding machine generated responses or a human written response. We randomly mix the human written ones to validate if the human evaluation is reliable to an extent, by anticipating human written ones will get higher scores. We list the average ratings in Table 5. The model trained with ExMATE achieves a similar informativeness level as human written ones, and gets a higher agility rating, which is its main goal. However, ExMATE can compromises coherence due to the subtraction of a counter example, which is a natural sentence, in its objective function. The human evaluation demonstrates the challenge of models to meet human-level quality in CausalDialogue featured by conversational DAGs, a portion of the diversed types of flows in the real world.

Model	Coherence	Informativeness	Agility
Human	3.78	3.72	3.49
MLE	3.63	3.60	3.36
ExMATE	3.59	3.74	3.40

Table 5: The human evaluation results (scale 1-5, the higher the better) of models trained on CausalDialogue (MLE, ExMATE), and human written responses (Human) for reference.

6.3 Qualitative Analyses and Discussion

Table 4 shows an example of a shared dialogue history, four different continuations (case1-4), and responses generated by the same backbone model, T5, trained with different objectives and inferred with different sampling methods. We observe that responses produced by MLE+T (TopK), ExMATE+S (Softmax), ExMATE+T are generally coherent to the conversation, while ExMATE often produces more diverse and agile responses to different continuation cases (different x). It is notable that other than the improvements, we find

that all the models have three types of issues: mode collapse, semantic repetition, and identity misplacement. **[Mode Collapse]** The problem is often-seen when inferring a model by greedy search, specifically, the predicted responses often repeat the same phrase such as "I'm not sure". While tacking the issue by adopting inference sampling, we conjecture the reason is that in a DAG, using a typical loss function can learn a probability distribution with higher entropy. This also demonstrates the need of a new loss function for training on a conversational DAG dataset. **[Semantic Repetition]** An example is the MLE+T response in Table 4 case 4, where "can't help you " and "help you out" have semantic overlaps. This issue can possibly be mitigated by repetition reduction techniques, such as unlikelihood training (Welleck et al., 2019) in future work. **[Identity Misplacement]** The problem happens when a model is confused about its position in a dialogue. For instance, the MLE+T response in Table 4 case 3 is more like an utterance of speaker Lysithea instead of Ignatz. This issue might be soothed by existing persona consistent techniques (Li et al., 2016b; Mazaré et al., 2018; Su et al., 2019) for building a overall good chatbot, while in this work, we focus on proposing a new dataset to benchmarking on the agility issue.

7 Conclusion

In this paper, we presented a new dataset, CausalDialogue, with novel conversational DAG structure. With experiments on various model setups with a newly proposed loss, ExMATE, we demonstrate that there is room for improvement to reach human-level quality, even though ExMATE improves the diversity, informativeness, and agility. This dataset serves as a testbed for future research that needs abundant conversation cases, like causal inference and offline reinforcement learning. Moreover, with the naturally paired metadata, future work can use this dataset for other tasks, such as speaker prediction in multi-speaker scenarioes and personalized dialogue generation.

Limitations

The introduced dataset has a moderate scale, as it is currently designed for fine-tuning instead of large model pretraining. Our proposed collection scheme can be futher applied to enlarge the dataset. Moreover, as we focus on English, the data source has multiple language versions written by experts. Hence, extending CausalDialogue to multilingual is straightforward. With reward labeling, the dataset can be more intuitively used for offline RL. Meanwhile, the dataset includes personality descriptions that can be used for personalized dialogue generation, even though is not the focus in this paper. Finally, training a generative model on dialogue domain can require various computational costs, depending on the aspects such as lengths of input and output texts and number of model parameters, as well as special designs to prevent misuses.

Ethics Consideration

The dataset is based on RPG game in fantasy world with diverse scenarios, including wars. To match the story background, a model trained on this dataset might produce war-related words. We manually looked into each example to meanwhile keep each speaker's personality and remove utterances that could potentially cause negative impact, such as violence, bias, and offensive words.

For the data annotation part and human evaluation part, we utilized the Amazon Mechanical Turk platform and required workers to have a HIT Approval Rate of greater than 95% and be located in CA or the US. We pay the annotators over 16 US dollars per hour on average, which is above the highest state minimum wage. Given our setting, the workers understood the scenarios and agreed that their annotations will be used for research. The data annotation part of the project is classified as exempt by Human Subject Committee via IRB protocols.

Acknowledgement

This work was supported in part by the National Science Foundation under #2048122 and an unrestricted gift award from Google Deepmind. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the sponsors.

References

Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur Szlam, Tim Rocktäschel, and Jason Weston. 2021. How to motivate your dragon: Teaching goal-driven agents to speak and act in fantasy worlds. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Rafael E Banchs. 2012. Movie-dic: a movie dialogue corpus for research and development. In *ACL*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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.

Scott Cunningham. 2021. *Causal inference*. Yale University Press.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Yao Dou, Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2021. Multitalk: A highly-branching dialog testbed for diverse conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12760–12767.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Kosuke Imai, Gary King, and Elizabeth A Stuart. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)*, 171(2):481–502.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2020. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3985–4003.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2016c. Dialogue learning with human-in-the-loop. *arXiv preprint arXiv:1611.09823*.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016d. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017a. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

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

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *EMNLP*.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in minecraft. In *ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*.

Ramakanth Pasunuru and Mohit Bansal. 2018. Game-based video-context dialogue. In *EMNLP*.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

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(140):1–67.

Revanth Rameshkumar and Peter Bailey. 2020. Storytelling with dialogue: A critical role dungeons and dragons dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2011. Data-driven response generation in social media. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI-16)*.

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*.

Feng-Guang Su, Aliyah R Hsu, Yi-Lin Tuan, and Hung-Yi Lee. 2019. Personalized dialogue response generation learned from monologues. In *INTERSPEECH*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-Yi Lee. 2019. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. 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)*.

Yi-Lin Tuan and Hung-Yi Lee. 2019. Improving conditional sequence generative adversarial networks by stepwise evaluation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4):788–798.

Yi-Lin Tuan, Wei Wei, and William Yang Wang. 2020. Knowledge injection into dialogue generation via language models. *arXiv preprint arXiv:2004.14614*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. 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)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Siddharth Verma, Justin Fu, Sherry Yang, and Sergey Levine. 2022. CHAI: A CHatbot AI for task-oriented dialogue with offline reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4471–4491, Seattle, United States. Association for Computational Linguistics.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 935–945.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940.

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)*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. 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.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.

Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. 2020. Counterfactual off-policy training for neural dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3438–3448.

A Appendix

A.1 Speaker Profiles

Table 6 provides a few examples of the speakers' profiles and utterances.

A.2 Data Expansion Details

Initial Dialogue Selection. We first randomly select m dialogues with replacement from the ORI.-2S partition, which is of higher quality after our manual inspection. This can result in more stable quality when doing crowd-sourcing. For each sampled dialogue, we randomly select a start time stamp t from $Poisson(\lambda = 1)$. Next, we adjust the sampled time stamp t to make sure it lies in an appropriate point to continue the dialogue by $t^* = \max(\min(t + 2, L), 2)$, where L is the maximum time stamp of this dialogue. For each time stamp, if the original dialogue has multiple possible nodes, we select one randomly from a uniform distribution. This process results in m initial dialogues D_0 with various lengths (at least two utterances) for expansion.

Expansion Collection. Each initial dialogue D_0 along with the continuing speaker profile is presented to n workers on MTurk to write the next utterance. The new continued dialogues D_1 will then be presented to another 1-2 workers, decided by $p\%$, playing as another speaker to gather another round of responses. This results in about $mn((1+p)^T - 1)/p$ new utterances for data expansion, where T is the number of iterations. Our expansion data is set with $m = 1200$, $n = 3$, $p = 0.2$, and $T = 3$. This setting results in about 13,000 written utterances.

A.3 Human Annotations

Interface - Data Expansion. We design two user interfaces to launch on MTurk for the first stage and the remaining stages separately of the data expansion process. The interface used for the remaining stages is shown in Figure 5. We include detailed instructions about the step-by-step works, examples, and requirements to obey. We put some information into a button to reduce cognitive burden when writing for multiple hits.

Interface - Human Evaluation. Our used human evaluation form is shown in Figure 6.

Setup and Payments. We collect the expanded dataset and evaluate generated responses via MTurk, a crowdsourcing platform. We obtained consent from workers by showing them the study purpose before they agree to do the annotations. We set additional restrictions of location to United States and Canada. We pay the annotators from 16-18 US dollars per hour according to the difficulty of the collection stage (remaining stages are more difficult than the first stage). The payments are made to be higher than the law's minimum wage 15 US dollars per hour in California in 2022 and 15.5 US dollars per hour in 2023, which are the highest among the US states.

A.4 Evaluation Metrics

Here we discuss more about our selection of evaluation metrics.

Fluency. The predicted next utterance should be both coherent to the previous turn and consistent with the dialogue history. We evaluate the extent of coherence by perplexity and reference-based metric BLEU (Papineni et al., 2002). For nodes with multiple childs we use multiple references when computing BLEU metrics. Although that BLEU may not be well correlated with human intuition in conversation (Liu et al., 2016), we use it for reference as it is still widely used in dialogue generation. The perplexity (PPL) is considered to be the less the better, whereas BLEU is the higher the better.

Diversity. A dialog model can suffer from the generic issue that given different dialogue history and previous turn, the predicted utterance is similar, such as "I'm sorry". We adopt distinct-N scores (Dist1 and Dist2) to evaluate this dimension by considering the percentage of distinct n-grams within the total number of n-grams in the corpus-level (Li et al., 2016a). However, the distinct-N scores are not always the higher the better. We can think about this in a intuitive example, if we randomly sample words from a uniform distribution, the distinct-N score can be high but meaningless. We anticipate a good distinct-N score is in a similar range as the score evaluated on human written responses.

A.5 Evaluate Human Responses

The PPL on human written responses are evaluated by an oracle that will predict a uniform distribution over all human written responses y given the same (DH, x). The BLEU scores on human written responses are evaluated on data examples with multiple possible responses and the response to be evaluated is hold out from the reference set.

Speaker	Profile Excerpt	Example Utterances
Byleth	Byleth has a very subdued personality and rarely expresses emotion.	- It's all right. // - Not really. // - I'm sorry.
Edelgard	Edelgard holds herself with a dignified air, but full of melancholy and solemn wistfulness.	- That's exactly right. There will no longer be lords who inherently rule over a particular territory. // - Perhaps not. Still, here you are. Maybe I can trust you with this...
Claude	Claude is described as easygoing on the surface, but has a side that forces others to keep their guard around him.	- Huh? Are you actually reading? I thought you hated studying. // - Was that story really worth bawling your eyes out over?

Table 6: In CausalDialogue dataset, some speakers profiles excerpts and their example utterances in conversations.

Otherwise, the BLEU scores will be 100 since the response to be evaluated is within the reference set.

A.6 Experiment Details

Model architecture. We use DialoGPT-small with 117M parameters and T5-base with 250M parameters. DialoGPT model is based on the GPT model architecture (a transformer decoder) but pretrained on conversation-like dataset such as Reddits. T5 model uses the transformer encoder-decoder architecture and is pretrained on web-extracted text from Common Crawl, which is a publicly-available web archive of scraped HTML files. The maximum tokens allowed as the input are 256.

Hyperparameters. For hyperparameter search, we tried the learning rate from {5e-5,2e-5,1e-5} and the batch size times gradient accumulation steps from {32,64,128}. We found out that using a learning rate 1e-5 and batch size 64 can generally fine-tuning a model well with different learning algorithms in our experiments. We train each model with different combinations of setups for single run.

Data preprocessing. For data preprocessing, we have tried to utilize the original case and punctuation, transform all words into lower case, or meanwhile remove all punctuation.

Computation Resources. Each model is train on one Titan RTX or one RTX A6000 and costs around five hours.

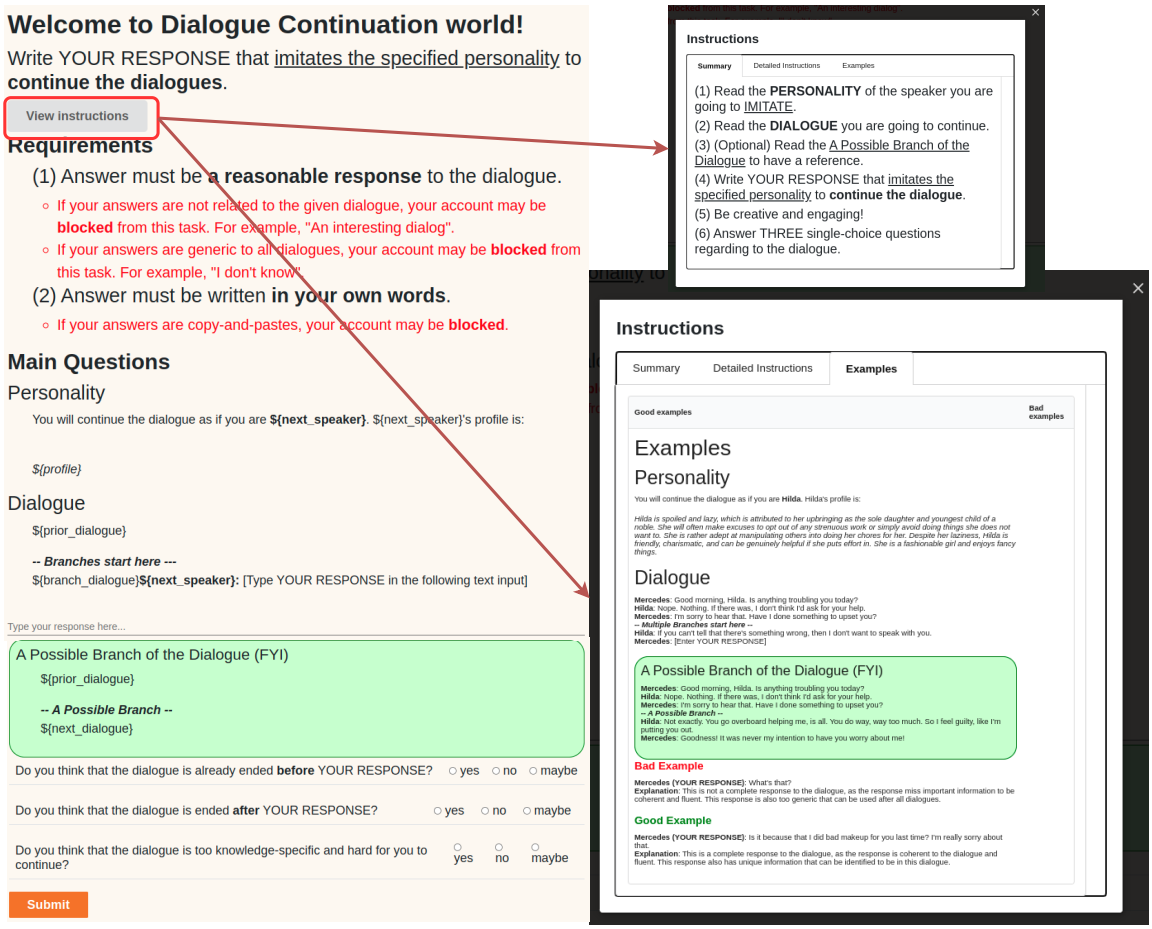

Figure 5: Screenshots of the interface we show the annotators to write responses for the **remaining stages** data expansion. We gave detailed instruction, requirements, and good/bad examples for reference.

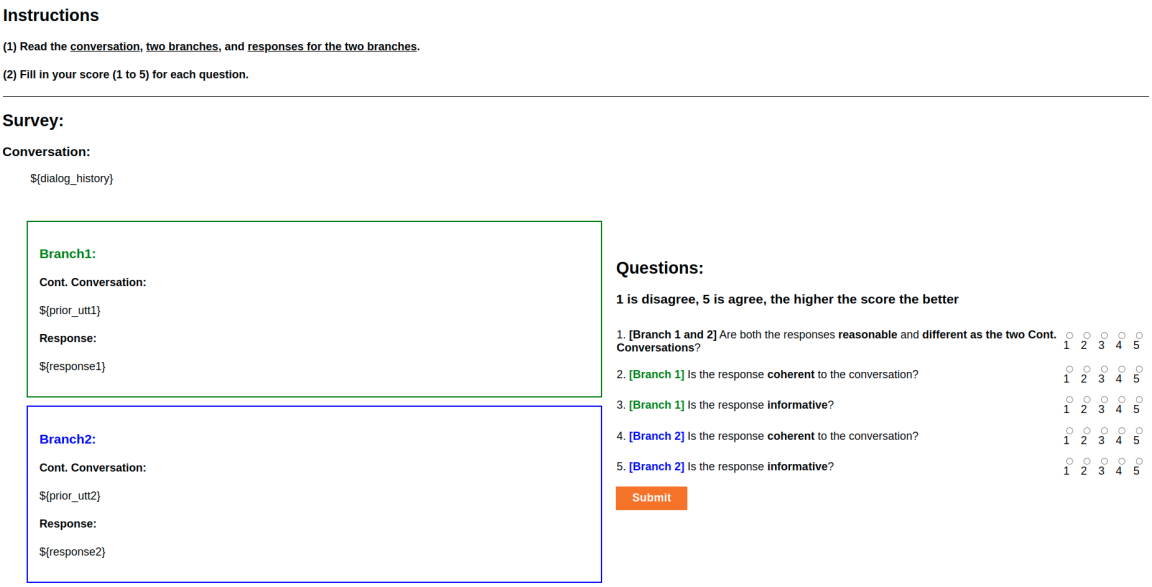

Figure 6: Screenshot of the interface we show the annotators to evaluate generated responses. We gave instruction with five questions each hit.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
We have included a limitation section after the main content.

☑ A2. Did you discuss any potential risks of your work?
We have included a limitation section and an ethical consideration section after the main content.

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

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

B ☑ Did you use or create scientific artifacts?

In Section 3,4,5,6

☑ B1. Did you cite the creators of artifacts you used?
In Section 3,5,6

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In Section 3

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In Section 3. We include the license and follow the intended use.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
In Section 3

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

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

C ☑ Did you run computational experiments?

In Section 6

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Appendix A.6

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 6 and Appendix A.6

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

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Appendix A.6 and Supplementary Material.

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

In Section 3.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
In Appendix A.3

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
In Appendix A.3 and Ethics Consideration section

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
In Appendix A.3 and Ethics Consideration section

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The data annotation part of the project is classified as exempt by Human Subject Committee via IRB protocols.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
In Appendix A.3 and Ethics Consideration section