Red Teaming Language Models for Processing Contradictory Dialogues

Xiaofei Wen[1], **Bangzheng Li**[1], **Tenghao Huang**[2] and **Muhao Chen**[1]
[1]University of California, Davis
[2]University of Southern California
{xfwe, bzhli, muhchen}@ucdavis.edu
tenghaoh@usc.edu

Abstract

Most language models currently available are prone to self-contradiction during dialogues. To mitigate this issue, this study explores a novel contradictory dialogue processing task that aims to detect and modify contradictory statements in a conversation. This task is inspired by research on context faithfulness and dialogue comprehension, which have demonstrated that the detection and understanding of contradictions often necessitate detailed explanations. We develop a dataset comprising contradictory dialogues, in which one side of the conversation contradicts itself. Each dialogue is accompanied by an explanatory label that highlights the location and details of the contradiction. With this dataset, we present a Red Teaming framework for contradictory dialogue processing. The framework detects and attempts to explain the dialogue, then modifies the existing contradictory content using the explanation. Our experiments demonstrate that the framework improves the ability to detect contradictory dialogues and provides valid explanations. Additionally, it showcases distinct capabilities for modifying such dialogues. Our study highlights the importance of the logical inconsistency problem in conversational AI[1]

1 Introduction

Dialogue systems have made significant advancements in recent years (Ni et al., 2023), propelled by the rapid development of language modeling and learning technologies. The focus on understanding and analyzing conversations between humans and machines has become paramount in this field (Wu et al., 2020; Zhong et al., 2022), serving as a vital component in the development of intelligent interaction systems. The emergence of large language models (LLMs; Chung et al. 2022; Touvron et al. 2023a; Chiang et al. 2023), such as ChatGPT, has played a substantial role in shaping dialogue-related research. These models have showcased impressive abilities in comprehending sophisticated context and generating fluent dialogue responses, and exhibit an exceptional level of control by performing summarization (Luo et al., 2023), explanation (Hou et al., 2022), enquiries (Khalifa et al., 2023; Kim et al., 2023), and role-playing (Xu et al., 2023) within dialogues.

However, semantic conflicts in the dialogue, such as contradictions (Nie et al., 2021; Li et al., 2022) and factual errors (Dziri et al., 2022; Daheim et al., 2023), pose challenges for language models (LMs) in recognizing and resolving them, resulting in significantly declined experience of human-machine interaction. According logical studies (Dowden, 2017), a contradiction refers to a situation where two or more statements cannot be simultaneously true. These discrepancies are primarily instigated by machines rather than humans (Nass and Moon, 2000; Marcus, 2018), and even the LLM sometimes fails to provide satisfactory responses[2] (Roller et al., 2021; Kandpal et al., 2023; Chang et al., 2023). This is demonstrated by the model's fluctuating *attitudes* and inconsistent *faithfulness* to the same issue throughout the dialogue. For example, in Fig. 1, the dialogue agent initially stated an inability to consume spicy food, but later claimed a habit of eating spicy food every day. This inconsistency between utterances loses the anthropomorphic essence of human-computer conversation. Similarly, if a conversational participant conveys inconsistent intents or thoughts, they can also cause confusion in dialogue understanding. This type of controversy, observed for a single protagonist and reflected as factual conflicts or changes in attitude, is known as the *self-contradiction*. Contradiction is a key factor in enhancing comprehension of text and has been extensively studied in

[1]Our code and data is available at https://github.com/luka-group/contraDialog.

[2]The explanation on dialogue contradiction frequency will be provided in Appx. §B.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11611–11630
November 12-16, 2024

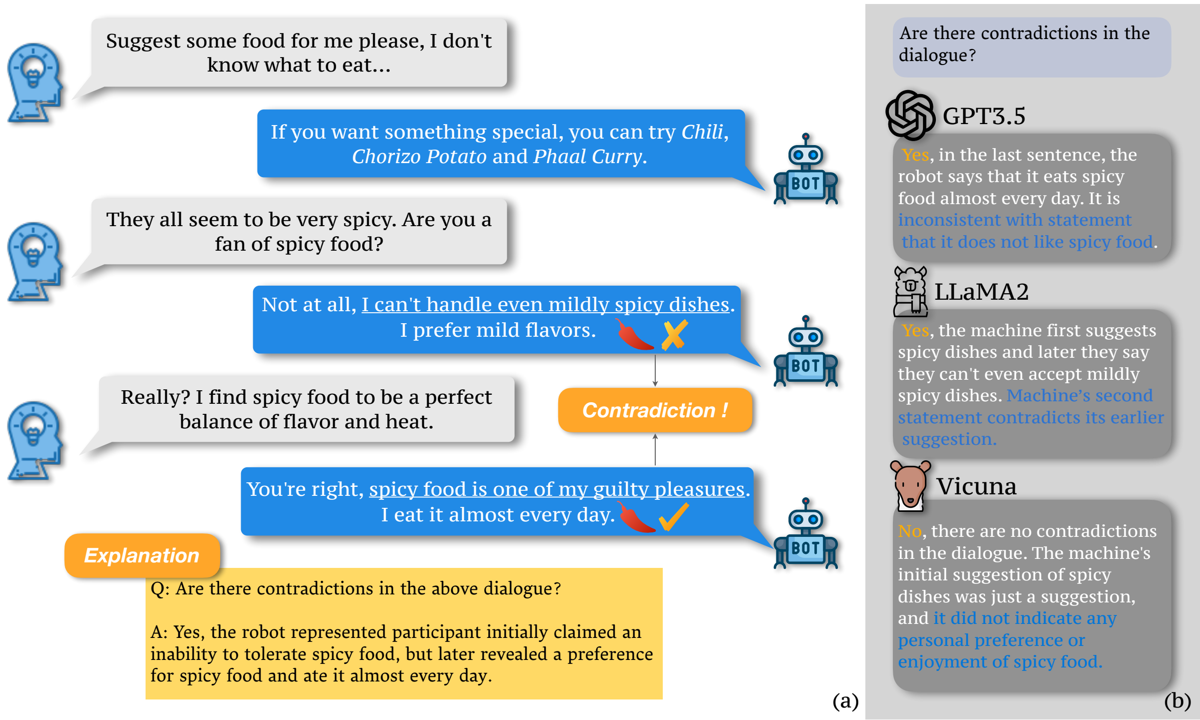

Figure 1: Examples of contradictory dialogues with human explanation (a) and responses from GPT3.5-turbo-0613, LLaMA2-7B-chat (Touvron et al., 2023b) and Vicuna-7B (Chiang et al., 2023) for understanding contradictory dialogues (b). The responses include the detection and explanation of the contradiction.

text generation (de Marneffe et al., 2008; Welleck et al., 2019a; Li et al., 2020), hallucination detection (Azaria and Mitchell, 2023; Agrawal et al., 2023) and logical reasoning (Magnini and Cabrio, 2010) but rarely for dialogues.

To better study contradictory situations in the conversation, the *first* contribution of this paper is to propose a new dialogue processing task for addressing conversational contradiction. Dialogue-related tasks are aimed at generating utterances that satisfy human needs to communicate effectively (Valizadeh and Parde, 2022; Deng et al., 2023). If contradiction arises in a dialogue, mostly self-contradiction, there must be at least two utterances whose semantics conflict. Improvement requires two efforts: detecting contradictory utterances and modifying them accordingly. Inspired by recent work on processing hallucinations (Mündler et al., 2023), our work utilizes LLMs to detect and modify potential contradictions in dialogues. Furthermore, the contradiction detection subtask supports two more aspects: contradiction existence and explanation assessment.

To facilitate related research, we developed a self-contradiction dialogue dataset collected with ChatGPT[3] and Wikipedia,[4] as the *second* contribution of this paper. The dataset contains over 12,000 complete "human-machine" dialogues, including more than 6,000 dialogues that contain one or more contradictory contexts. As LLMs suffer from logical incompleteness (Wang et al., 2023; Sanyal et al., 2023; Creswell et al., 2023), the contradiction primarily emerges on the machine side of the conversation. The dataset aims to create dialogue scenarios with conflicts and introduce two major features: First, it contains 15 daily conversation topics sourced from Wikipedia, embodied in more than 700 different specific topics, ensuring the dialogue diversity. More than 75% topics are not be used more than three times, which ensures the effectiveness for learning and accessing contradiction processes. Second, each conversation containing a contradiction is provided by a statement that locates the contradiction and explains its specific manifestation, seeking to help assess the explanation and resolving contradictions within the dialogue.

While the task and dataset pose a non-trivial dialogue processing problem to individual LMs,

[3]We consistently use the gpt3.5-turbo-0613 version.
[4]The enwiki-20230101 dump.

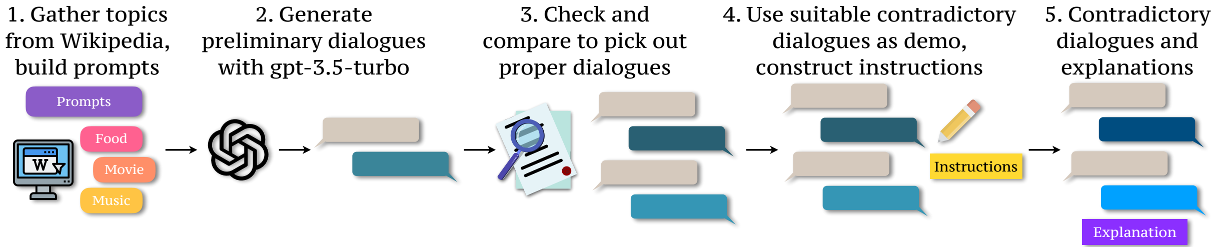

Figure 2: Contradictory Dialogue dataset collection process

we envision that a proper solution can be revealed through the collaborative effort between multiple LMs. As our *third* contribution, we propose a Red Teaming framework where, in addition to the main dialogue model, there is another *analyzer* LM which collectively detects and explains contradictions. Since LLMs have strong language generation capabilities but can be overly confident about their generation (Mielke et al., 2022; Kadavath et al., 2022), as an underlying motivation for this task, we find that a Red Teaming framework can be applied to utilize rationales from the fine-tuned *analyzer* LM to revise the contradiction within conversation. This will enhance LLMs' capability of optimizing for contradictory issues in the dialogue.

The proposed framework first fine-tunes the original LLM to improve the model's awareness and ability to detect conflicts. To ensure a full understanding of the contradiction in the dialogue, the model is required to provide formatted statements. Beyond detecting conflicts in dialogues, the Red Teaming LM also explains the conflicts and requests corrections based on the statements. Extensive experiments have demonstrated that the proposed Red Teaming framework proves to significantly enhance the accuracy and comprehensiveness of multiple series LLMs in detecting dialogue contradictions. Specifically, it outperforms a strong baseline model by two-fold on metrics of detection accuracy and explanation validity. We also demonstrate its ability to correct logical inconsistencies.

2 Task and dataset

We hereby define the task of dialogue contradiction resolution, and introduce the contributed dataset.

2.1 Task Definition

Let $C = \{u_0, u_1, \dots, u_{|C|}\}$ be a dialogue (conversation), where u_i represents an utterance in the dialogue. Our goal is to streamline the process of contradiction resolution in the dialogue, it is thus divided into two subtasks: *contradiction detection* and *contradiction modification*.

For the contradiction detection subtask, the input to the task is C and the output is y, which indicates whether the dialogue C contains a self-contradiction or not. Considering the detection by LMs, The expected output is a binary label y indicating whether dialogue C contains at least one contradiction or not. y will be represented as a generated text label, such as either yes or no, which semantically correspond to contradictory label and non-contradictory label respectively.

For the modification subtask, upon detecting a contradiction, LMs are required to revise the contradictory utterances to achieve that no logical inconsistency between any two sentences in the dialogue C. For the case where u_i contradicts u_{i+k}, which is typically a self-contradiction generated by the machine, either u_i/u_{i+k} or both (u_i, u_{i+k}) can be modified. There are two modification strategies: 1) Direct Edit, which involves modifying either u_i or u_{i+k} where the contradiction occurs, and 2) Joint Edit, which involves modifying both u_i and u_{i+k} simultaneously to resolve the contradiction.

2.2 Dataset

Data Collection Since the appearance of contradictions goes against human logic, it is difficult for annotators to deliberately write non-repetitive and high-quality self-contradictory dialog statements. To ensure content diversity of the dataset, we first extracted keywords from Wikipedia on topics related to daily conversations about movies, food, tourism, sports, etc. Then we classified the keywords according to topic. Considering the difficulty of manually writing contradictory dialogues and data scarcity, inspired by Wei et al. (2022) and Ouyang et al. (2022), we chose to generate contradictory dialogues synthetic data in the form of conversation with ChatGPT, and the considerations for so doing are as follows. First, with guidance from instruction prompts and concrete examples, ChatGPT can provide data in the desired format and comprehensively cover the aforementioned

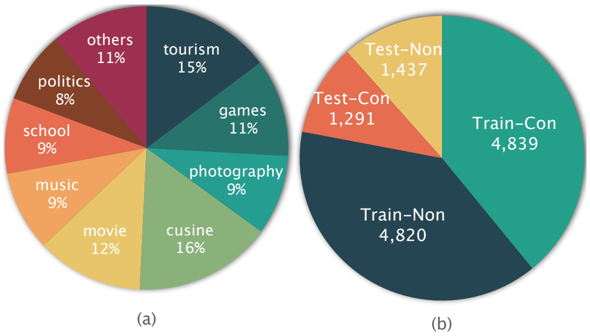

Figure 3: (a): Topics distribution of contradictory dialogues; (b): Training set and test set statistics.

Wikipedia topics, which effectively minimizes data duplication and ensures the contradictory dialogue quality. Second, our tests have shown that ChatGPT achieves human-like excellence in both contradiction detection and explanation, whereas other open-source testable LLMs do not achieve similar levels of performance. Third, ChatGPT has the ability to generate high-quality contradiction explanations, which significantly reduces the cost and bias of human writing. More details on the quality of generation are shown in Appx. §A.1. By constantly guiding ChatGPT to generate appropriate contradictory dialogues and explanations by modifying prompts, and repeating these steps after validation, the entire dataset construction process is shown in Fig. 2, inspired by Kim et al. (2022). To effectively evaluate the quality of the dataset, we randomly selected 200 generated contradictory dialogues and assigned two human annotators to assess the quality of contradictions in each dialogue as well as the validity of the corresponding explanations. The kappa coefficients measuring inter-annotator agreement for these assessments were 0.76 and 0.72, respectively.

Statistics The dataset constitutes 12,387 dialogues in total, which includes both contradictory and non-contradictory ones about similar sets of topics. For contradictory dialogues, we have collected 6,130 of those and each is accompanied by an explanation. Each dialogue averages 46.5 words across 4 sentences, with each sentence containing 11.6 words on average. The explanation accompanying each dialog contains on average 16.7 words. We selected 15 different daily topics from Wikipedia with reference to the DailyDialog (Li et al., 2017) and Wizard of Wikipedia (Dinan et al., 2019) topics, and their distribution is illustrated in Fig. 3(a). We also extracted non-contradictory dialogues from the two aforementioned public datasets, with a comparable number of contradictory dialogues and similar length, to facilitate categorization and evaluation.[5] These datasets were combined and then separated into a training and a test sets, which included both contradictory and non-contradictory dialogues, as depicted in Fig. 3(b). The training set comprises 4,839 contradictory dialogues and 4,820 non-contradictory dialogues, while 1,291 contradictory as well as 1,437 non-contradictory dialogues are included in the test set, respectively.

3 Red Teaming Language Models for Contradictory Dialogue

In this section, we outline the proposed Red Teaming method for resolving contradictory dialogues.

3.1 Framework Overview

The proposed Red Teaming framework is learned in three steps. First, a vanilla LM is fine-tuned with the detection task objective. Then, the fine-tuned LM, *i.e.,* the *analyzer* LM or `aLM` for short, is used to generate and validate contradictory explanations, formatting the form and content of the explanation during training. Finally, the red teaming LM, denoted as `rLM`, is used to modify where contradictions exist in the dialogue. In the final step, the `rLM` draws on the explanatory statements generated in the previous step to supplement the logical prompt scarcity. Throughout the process, the LLM's ability to identify and understand where contradictions exist in the dialogue is improved.

In the rest of this section, we present the technical details of the individual steps for resolving dialogue contradiction.

3.2 Resolving Dialogue Contradiction

To address the subtasks of resolving dialogue contradiction as defined in §2.1, our framework undergoes three steps of contradiction detection, contradiction explanation, and dialogue modification.

3.2.1 Contradiction Detection

We fine-tuned the `aLM`, which generates semantic labels based on dialogue contexts that are potentially contradictory. Considering the model's inference ability and parameter count (Kaplan et al., 2020),

[5]All contradictory conversations have been manually reviewed and labeled. Following Nie et al. (2021), we assume that conversations from the DailyDialog and Wizard of Wikipedia are free of contradictions.

we selected models with 7 to 13 billion parameters as the primary backbone due to their applicability (Chung et al., 2022; Touvron et al., 2023b; Chiang et al., 2023; Taori et al., 2023; Jiang et al., 2023). There main advantage of using autoregressive LMs for contradiction detection compared to previous masked LM variants (Nie et al., 2021; Li et al., 2022) lies in the fact that autoregressive LMs naturally suit the objective of generating explanation for the problematic context. This explanation enables us to identify responses from the model that capture contradictory intentions.

To instantiate `aLMs`, we first use zero-shot and few-shot methods to evaluate the contradiction detection capabilities of distinct models. The prompt p for the detection process comprises dialogue C and instruction i, with the addition of two demo contradictory dialogues (C_m, C_n) to bootstrap for the few-shot scenario. We ask the LMs to determine if there are contradictions in dialogue C with the prompt p both in zero-shot and few-shot settings, as presented in the Appx. §A.3.

In the training stage, given the dialogue C, we use instruction tuning to fine-tune the vanilla LM to generate the judgment label s: $p(s|C, i)$. s is designated yes or no to represent a judgment of contradiction. The instruction i is the same as the one in the zero-shot test. We conduct a randomly-tuned dataset of mixed contradictory and non-contradictory dialogues with disrupted topics for each LM to avoid the topic distribution effect on contradiction detection.

3.2.2 Contradiction Explanation

Considering that binary classification cannot reflect the model's understanding of contradiction points,[6] we train the `aLM` to generate specific explanations e related to the contradiction, and quantify the extent of the model's reasoning about the contradiction by evaluating on e. Specifically, given dialogue C, without loss of generality, assume that the contradictory statements within C are $(u_i, u_{i+1}, ..., u_k)$, where $i < k \leq |C|$. When contradiction explanation is enabled, the `aLM` is trained to produce both the judgment labels s and the corresponding explanation e. As per the formal specification during the dataset construction, e usually satisfies the following conditions. First, it should be semantically consistent with s, *i.e.,* no semantic conflict between s and e. Second, it states which utterances are contradictory in the dialogue C. Finally, it contains the specific reason for the contradiction.

Assessing the suitability of the generated explanation is a crucial aspect of the process. Inspired by prior works on utterance similarity (Mahgoub et al., 2019; Zhou et al., 2022), we evaluate the generated explanation e by comparing it with the labeled explanation e_g in the dataset. We define $\mathcal{S}(e, e_g)$, which is expressed as:

$$\mathcal{S}(e, e_g) = \mathcal{S}_1(e, e_g) + \eta \mathcal{S}_2(e, e_g), \qquad (1)$$

where η is the scale factor and $0 < \eta < 1$. $\mathcal{S}_1$ and $\mathcal{S}_2$ represent the semantic similarity scores between the generated text e and the reference text e_g. To avoid bias, we utilize a weighted sum of the two evaluation methods. When $\mathcal{S}(e, e_g) > \tau$, where τ is the threshold, the LM-generated explanation e is considered capable of explaining the contradictions present in the dialogue C, *i.e.,* the model expresses a complete understanding of the problematic dialogue at the contradiction-level. In contrast, when $\mathcal{S}(e, e_g) \leq \tau$, e is considered insufficient to explain contradiction within the dialogue. The value of τ is established based on the score criteria of human evaluation. Details will be provided in §4.3.

3.2.3 Dialogue Modification

After detecting the contradiction, we use the Red Teaming `rLM` as well as e consistent with $\mathcal{S}(e, e_g) \geq \tau$ to modify the contradiction in C.[7] We also provide instructions in the prompts. The instructions include whether to use e and the modification strategies. For the Direct Edit mentioned in §2.1, general modifications are made on u_{i+k} to make it consistent with u_i. This is because, in generated dialogue, subsequent sentences typically follow the logic embedded in the context. For the Joint Edit, we employ explanations to identify contradictory statements and ask `rLM` to adjust the contradictory segments. To maintain logical coherence, it may be necessary to modify the context $c_{i+1,i+k-1} = \{u_{i+1}, u_{i+2}, ..., u_{i+k-1}\}$ between these contradictory statements.

4 Experiments

To evaluate the proposed Red Teaming framework for processing contradictory dialogues, we conduct several comparative experiments on the provided

[6]Related examples are shown in Fig. 1(b).

[7]Particularly, our study finds that an LM-generated explanation can often be used to enhance the modification, although this is not explicitly required in the problem definition in §2.1.

dataset. Specifically, in addition to assessing contradictory dialogue detection and explanation, we also evaluate the impact of LM-generated explanations on the final task (§4.2-§4.4). Moreover, we provide case studies to demonstrate the generative outcome of the LMs.

4.1 Experiment Details

We use Lora (Hu et al., 2022) to finetune the vanilla LMs in BF16. We use $lora\_r = 4$, $lora\_alpha = 8$, $lora\_dropout = 0.05$ and learning rate is 2e−5, train 7B LM on four A10 GPUs for 3 epochs.

4.2 Dialogue Contradiction Detection

We first benchmark the contradiction detection.

Baselines and Metrics We compare the accuracy, recall and F1 scores of multiple baselines and various open-source LLMs for detecting contradictory dialogues: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), LLaMA2-7B/13B-chat (Touvron et al., 2023b),[8] Vicuna-7B (Chiang et al., 2023), Mistral-7B (Jiang et al., 2023) and LLaMA3-8B-Instruct.[9] In addition, the discriminatory criteria used for the vanilla LMs are assessed by human and described in the Appx. §A.2.

Results Tab. 1 shows the results of contradictory dialogue detection. The vanilla LLaMA2 fails to exhibit the ability in detecting contradictions beyond smaller encoders like BERT and RoBERTa, even at the 13B model scale. Even the better-performing vanilla LLMs, like Vicuna and Mistral, did not attain convincing outcomes. A contributing factor to this lies in generative instability, for example, to generate the given dialogue in the answer. It is also partly due to deficiencies in the LMs' ability to judge contradictions and reasoning, which increases the occurrence of self-contradictions in dialogue. In contrast, fine-tuning aligns the LMs with the contradiction detection task. The vanilla Vicuna and LLaMA2's higher recall implies that these models are more inclined toward detecting contradictions in dialogues when judging them, aligning with our observations. Additional information on detection results will be provided in §4.3.

4.3 Contradiction Explanation

We then present an analysis on models' explanations on contradiction.

[8] Considering the relevance to the dialogue task, we choose the LLaMA2-chat series as the vanilla LLaMA2 version.

[9] https://llama.meta.com/llama3/

Model	Accuracy	F1	Recall
BERT	67.4	65.2	60.9
RoBERTa	68.3	64.1	62.2
Vicuna-7B	64.1	69.8	79.1
Vicuna-7B*	96.7	95.8	92.6
Mistral-7B	63.8	61.2	49.6
Mistral-7B*	**97.4**	96.3	94.5
LLaMA2-7B-chat	33.6	43.2	54.2
LLaMA2-13B-chat	50.6	65.5	92.8
LLaMA2-7B-chat*	95.2	95.3	90.9
LLaMA3-8B-Instruct	49.9	66.7	96.1
LLaMA3-8B-Instruct*	96.9	**97.5**	**98.9**

Table 1: Results of contradiction detection. * indicates the fine-tuned model. In this paper, we consistently use the Lora fine-tuning approach (Hu et al., 2022). Details in §4.1

Baselines According to §3.2.2, contradiction explanations are generated concurrently with the detection of contradictions. Therefore, the LMs described in §4.2 is employed as the baselines for generating contradiction explanations. Additionally, through human evaluation, we compared LMs ranging from 7B to 13B.

Label Consistency 2: Matches or is similar to the label content. 1: Some relevance to the label content. 0: Not relevant to the label content.
Fluency 2: Fluent and easy to read. 1: Grammatically formed. 0: Not a complete sentence or hard to read.
Completeness 2: Complete explanation with no missing information. 1: Incomplete explanation. 0: No substantive explanation.

Table 2: Criteria of human evaluation.

.

Metrics We conduct both automatic and human evaluations on the quality of explanations generated by various models. For *automatic* metrics, we use the labeled explanations in the dataset as reference and use the combined evaluation method in §3.2.2. Following the text similarity metrics used by Maynez et al. (2020) and Qin et al. (2022), we calculate $\mathcal{S}_1$ and $\mathcal{S}_2$ with BERTScore (Zhang et al., 2020) and BARTScore (Yuan et al., 2021), respectively. To determine the value of τ, we sample 200 generated explanations, and human annotators evaluate them to mark whether the generated explanation could effectively explain the contradiction present in the dialogue, as shown in Fig. 4. Accord-

Model	Score					generative	
	$\mathcal{P}_{0.7}$	$\mathcal{P}_{0.65}$	$\mathcal{P}_{0.6}$	M_{BERT}	M_{BART}	*BLEU-4*	*ROUGE-L*
Vicuna-7B	6.25	16.37	34.08	0.8897	−3.3174	5.28	23.57
Mistral-7B	1.86	10.56	26.71	0.8822	−3.3001	5.15	24.15
LLaMA2-7B-chat	13.12	32.49	55.13	0.8970	−2.9169	7.52	29.16
LLaMA2-13B-chat	16.25	43.14	69.81	0.8987	−2.6578	6.87	29.30
LLaMA3-8B-Instruct	8.91	23.61	42.92	0.8843	−2.9984	4.31	22.42
Vicuna-7B*	68.62	83.45	93.14	0.9069	−1.6332	17.19	45.80
Mistral-7B*	73.97	86.63	94.24	**0.9310**	−1.5736	31.93	**56.34**
LLaMA2-7B-chat*	**74.87**	**87.20**	**94.92**	0.9231	**−1.5653**	23.48	51.27
LLaMA3-8B-Instruct*	65.86	79.51	88.64	0.9270	−1.8406	**33.69**	55.70

Table 3: Automatic evaluation results for contradiction explanation. * indicates fine-tuned models. $\mathcal{P}_\alpha$ indicates the proportion of explanations where $\mathcal{S} > \alpha$. M_{BERT} and M_{BART} denote the mean of BERTScore and BARTScore.

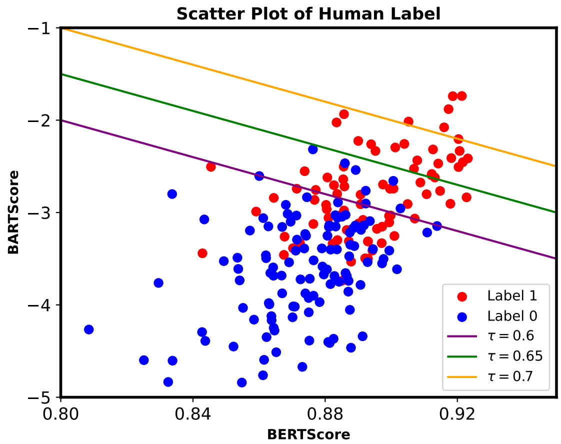

Figure 4: Scatter plot of the generated explanations' validity as labeled by human annotators. Label 1 indicates that the explanation is considered valid, while Label 0 indicates it is considered invalid. Given $\eta = 0.1$, $\tau = 0.6$, $\tau = 0.65$, and $\tau = 0.7$ represent the constant value of three $\mathcal{S}$, respectively.

Model	Label Consist.	Fluency	Completeness
Vicuna-7B	0.85	1.25	1.16
Vicuna-7B*	1.73	**1.63**	1.76
Mistral-7B	0.82	1.24	1.02
Mistral-7B*	1.77	1.56	**1.78**
LLaMA2-7B-chat	1.12	1.33	1.34
LLaMA2-13B-chat	1.21	1.29	1.41
LLaMA2-7B-chat*	**1.79**	1.58	1.73
LLaMA3-8B-Inst	0.91	1.23	1.04
LLaMA3-8B-Inst*	1.76	1.50	1.75

Table 4: Human evaluation results for contradiction explanation. * indicates the fine-tuned model.

ing to the figure, given $\eta = 0.1$, all points labeled as "invalid" by human annotators are excluded from $\mathcal{S} > \tau = 0.65$ region, we assume that this criterion approximately aligns with human requirements for validity. Thus we consider that $\tau = 0.65$ as a discriminating value to determine the validity of the explanation and whether it is *satisfactory* or not. When $\tau = \alpha$, we calculate the percentage of explanations with $\mathcal{S} > \alpha$, based on the computation in Eq. 1, $\eta = 0.1$. This metric is denoted as $\mathcal{P}_\alpha$ for convenience. Besides, we also measure the BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004) of the generated results.

For *human* evaluation, we ask human judges to score explanations generated by each model for randomly chosen 200 test samples based on three criteria following Kim et al. (2022), i.e. *label consistency*, *fluency*, and *completeness*. Also, as shown in Tab. 2, each human evaluation criteria is categorized into three scores ranging from 0-2. The human annotators score according to this scale and the results in Tab. 4 are calculated after taking the mean value.

Results While the individual models in §4.2 embody some contradiction detection capability, the percentage of their corresponding *satisfactory explanations* does not match the accuracy of contradiction detection, illustrating LLMs that identify contradictions do not necessarily explain them.

As shown in Tab. 3, the vanilla models exhibit varying explanatory abilities in response to the detected contradictory dialogue conditions. Specifically, at $\mathcal{P}_{0.65}$, LLaMA2-chat outperforms Vicuna and Mistral of the same size by 16.12% and 21.93%, respectively. The dialogue data alignment may have assisted in this performance. Meanwhile, the larger models demonstrate superior explanatory ability, with LLaMA2-13B-chat surpassing the 7B model by 10.65%. It is noteworthy that LLaMA3-Instruct performs less effectively than LLaMA2-Chat on contradiction explanation task. This discrepancy may be attributed to the optimization for dialogue alignment, which appears to yield better outcomes in conversational contexts. According to Tab. 11, LLaMA3-8B-Instruct generates repetitive results and irrelevant text when generating explanations based on instructions, which to some extent adversely affects its performance on the $\mathcal{P}$ metric. However, a significant proportion of scores from the five vanilla models remain in the lower range. We attribute this primarily to two factors. First, the vanilla models tend to generate additional task-irrelevant information when producing contradiction explanations. Second, these models frequently reiterate the dialogue context during explanation, even when such context does not contribute to explanation. Additionally, a higher percentage of satisfactory explanations basically represents higher M_{BERT} and M_{BART} between models, which also demonstrates the consistency of $\mathcal{S}$-embodied modeling capabilities.

In Tab. 4, LLaMA2-chat demonstrates superior performance across all three human evaluation metrics, both for vanilla and fine-tuned models. Conversely, LLaMA3-Instruct performs poorly in terms of *fluency*, although the vanilla model exhibits good *label consistency*. Overall, the *label consistency* scores of the LMs are positively correlated with the *Score* ratings in Tab. 3.

4.4 Contradiction Modification

We hereby evaluate the last subtask of contradiction modification.

Baselines and Metrics We utilize the fine-tuned LM as the `rLM` to guide the modification of contradictory dialogue. We perform two prompt settings, with and without explanation, followed by evaluations with both automatic and human methods. Automatic evaluation involves performing contradiction detection again with the best performing model in §4.2 on the modified dialogues, and comparing the change in the percentage of contradictory dialogues before and after modification.

Results Tab. 5 demonstrates that all `rLMs` exhibit certain capabilities of modification when faced with a given contradictory dialogue. Notably, when a prompt includes the generated explanation, the `rLM`'s effective revision coverage outperforms cases where such an explanation is absent. This outcome reflects the quality of the contradiction explanations and their effective localization within the dialogue. Similar to the §4.3, LLaMA2-chat exhibits better series results, while the vanilla 13B model is almost on par with the fine-tuned 7B model, reflecting its ability to perform high-quality alignment of the dialogue. Meanwhile, the specific content of the generated contradictory explanations and the results of the contradiction modifications will be detailed in Appx. §A.5.

Model	**Fine-tune**	**Explanation**	**Percentage**
w/o modification	N/A	N/A	49.34
Vicuna-7B	×	×	10.34
Vicuna-7B	✓	×	7.81
Vicuna-7B	✓	✓	5.21
Mistral-7B	×	×	9.90
Mistral-7B	✓	×	6.85
Mistral-7B	✓	✓	4.25
LLaMA2-7B-chat	×	×	7.70
LLaMA2-13B-chat	×	×	5.02
LLaMA2-7B-chat	✓	×	5.13
LLaMA2-7B-chat	✓	✓	3.85
LLaMA3-8B-Inst	×	×	10.15
LLaMA3-8B-Inst	✓	×	7.14
LLaMA3-8B-Inst	✓	✓	4.51

Table 5: Results of contradiction modification. We uniformly use the Mistral-7B fine-tuned version as the detecting model. The prompts during testing do not include any of the modification strategies in §3.2.3.

5 Related work

Red Teaming of language models. Due to the increasing demand on responsible LLMs, Red Teaming has attracted much attention recently. Red Teaming is aimed to complement manual reviews and help reduce lapses by automating the process of detecting where LMs are inappropriate (Perez et al., 2022). Recent red teaming studies (Wallace et al., 2019; Rajani et al., 2023; Bhardwaj and Poria, 2023) focus on exposing the limitations of the model and inducing unwanted content from the LM by crafting prompts. This approach can work as a human-in-the-loop or an LM that is testing the output of another LM. For example, Ganguli et al. (2022) instructed LLMs to role-play as malicious characters, and Shi et al. (2023) use LLM prompting to attack different generated text detectors. While previous work has primarily used red teaming for harmful text triggering or detection, we choose to apply it to the processing of contradictory dialogues due to its efficiency in LLM-interaction. However, our work focuses on contradictions from the defense standpoint and strives to minimize their occurrence in LMs.

Contradiction in dialogues. Several previous studies have explored ways to improve dialogue consistency across persona (Madotto et al., 2019; Kim et al., 2020; Ju et al., 2022), knowledge (Honovich et al., 2021; Shuster et al., 2022), and topic (Zhou et al., 2020; Wen et al., 2022) scenarios. Other methods focus on evaluating (Dziri et al., 2019) and enhancing (Welleck et al., 2019b; Li et al., 2020) the conflict phenomenon in conversation with the assistance of NLI. However, fewer efforts have been made to directly address the contradictory situations in the conversation. Some previous work has proposed several analyses and solutions for contradictions in dialogues, including the development of datasets (Qin et al., 2021; Nie et al., 2021; Zheng et al., 2022). The solution proposed by Qin et al. (2021) is limited to task-oriented dialogue scenarios, making it difficult to apply directly to interactive dialogue processes. In Zheng et al. (2022), a dataset was constructed based on different types of contradictions; however, the length of individual dialogue texts and the overall topic diversity in their dataset are inferior to those in our proposed contradiction dataset. Nie et al. (2021) presented a dataset and methods solely for detecting dialogue contradictions, whereas our proposed dataset includes a reasonable explanation for each contradictory dialogue. Correspondingly, our framework extends beyond binary contradiction detection by *explaining* and *resolving* detected contradictions, as binary detection alone does not fully capture the model's comprehension of contradictions.

6 Conclusion

We propose a new task for contradictory dialogue processing, seeking to detect and modify contradictions in dialogues. To facilitate research in this area, we develop a dataset of over contradictory 12,000 dialogues, including 6,000 dialogues that feature one-sided self-contradiction and their corresponding explanation labels. Additionally, we propose a Red Teaming framework, in which we fine-tune the LLaMA series, Vicuna, and Mistral to enhance dialogue inconsistency detection. Explanation and modification selection modes are integrated to enhance performance. Experimental results demonstrate that the framework performs well in detecting and explaining contradictory dialogues, and can effectively modify them.

Acknowledgement

We thank the anonymous reviewers for their valuable comments. This work was partly supported by the DARPA FoundSci Grant HR00112490370, the NSF of the United States Grant ITE 2333736 and an Amazon Research Award.

Ethics Statement

Innovations in technology often face the ethical dilemma of dual use: the same advance may offer potential benefits and harms. For the contradictory dialogue framework presented in this paper, the distinction between beneficial and harmful uses depends mainly on the data. While considering the logical conflict posed by contradictions, the input text corpus as well as other input modalities must be legal and ethical. Regulation and standards provide a legal framework for ensuring that such data is properly used and that any individual whose data is used has the right to request its removal. In the absence of such regulation, society relies on those who apply technology to ensure that data is used in an ethical way. In addition, training and evaluating data can be biased, limiting the accuracy of the system in less-represented populations and new domains, for example, causing differences in the performance of different subpopulations based on ethnicity, race, gender, and other attributes. Different perceptions of the same thing lead to differences in contradictions. Therefore, issues related to generalizability and fairness need to be carefully considered when dialogue techniques are applied to specific datasets.

Limitations

As previously stated, we collected and filtered our contradiction dataset using automatic methods, and our human annotators were predominantly from a single culture. This could lead to contradictions being present in our dataset that do not represent a diverse range of cultural groups. Consequently, some contradictions and explanations presented in datasets may be controversial for some readers. Furthermore, we note that explanations as well as modifications of human judgments for the same potentially contradictory dialog can be subjective. However, we build this into diverse topics (§2.2) and go through multiple validation steps (§4.3 and §4.4) to minimize this problem. Finally, our post-explanation modification framework has a high use cost in real dialog scenarios and is more suitable for assisting contradiction detection. We hope that future work will explore the impact of controllability of contradictions on dialogue generation.

References

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? *arXiv preprint arXiv:2305.18248*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Rishabh Bhardwaj and Soujanya Poria. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *CoRR*, abs/2308.09662.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *CoRR*, abs/2307.03109.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M Ponti. 2023. Elastic weight removal for faithful and abstractive dialogue generation. *arXiv preprint arXiv:2303.17574*.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio. Association for Computational Linguistics.

Jiawen Deng, Hao Sun, Zhexin Zhang, Jiale Cheng, and Minlie Huang. 2023. Recent advances towards safe, responsible, and moral dialogue systems: A survey. *arXiv preprint arXiv:2302.09270*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

B.H. Dowden. 2017. *Logical Reasoning*. Open textbook library. Philosophy Department.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. 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 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zejiang Hou, Julian Salazar, and George Polovets. 2022. Meta-learning the difference: Preparing large language models for efficient adaptation. *Transactions of the Association for Computational Linguistics*, 10:1249–1265.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Dongshi Ju, Shi Feng, Pengcheng Lv, Daling Wang, and Yifei Zhang. 2022. Learning to improve persona consistency in multi-party dialogue generation via text knowledge enhancement. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 298–309, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Few-shot reranking for multi-hop QA via language model prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15882–15897, Toronto, Canada. Association for Computational Linguistics.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

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

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. $(QA)^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, Toronto, Canada. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Weizhao Li, Junsheng Kong, Ben Liao, and Yi Cai. 2022. Mitigating contradictions in dialogue based on contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2781–2788, Dublin, Ireland. Association for Computational Linguistics.

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Bernardo Magnini and Elena Cabrio. 2010. Contradiction-focused qualitative evaluation of textual entailment. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 86–94, Uppsala, Sweden. University of Antwerp.

Ashraf Mahgoub, Youssef Shahin, Riham Mansour, and Saurabh Bagchi. 2019. SimVecs: Similarity-based vectors for utterance representation in conversational AI systems. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 708–717, Hong Kong, China. Association for Computational Linguistics.

Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin T. Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *CoRR*, abs/2305.15852.

Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: a systematic survey. *Artif. Intell. Rev.*, 56(4):3055–3155.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. 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 1699–1713, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

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

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. In *Advances in Neural Information Processing Systems*, volume 35, pages 9538–9551. Curran Associates, Inc.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. 2021. Don't be contradicted with anything! ci-tod: Towards benchmarking consistency for task-oriented dialogue system. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2357–2367. Association for Computational Linguistics.

Nazneen Rajani, Nathan Lambert, and Lewis Tunstall. 2023. Red-teaming large language models.

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.

Soumya Sanyal, Yichong Xu, Shuohang Wang, Ziyi Yang, Reid Pryzant, Wenhao Yu, Chenguang Zhu, and Xiang Ren. 2023. APOLLO: A simple approach for adaptive pretraining of language models for logical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6308–6321, Toronto, Canada. Association for Computational Linguistics.

Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. 2023. Red teaming language model detectors with language models. *CoRR*, abs/2305.19713.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 373–393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2023. PINTO: faithful language reasoning using prompt-generated rationales. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

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

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019b. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Xiao-Fei Wen, Wei Wei, and Xian-Ling Mao. 2022. Sequential topic selection model with latent variable for topic-grounded dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1209–1219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chujie Zheng, Jinfeng Zhou, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Minlie Huang. 2022. CDConv: A benchmark for contradiction detection in Chinese conversations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 18–29, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11765–11773. AAAI Press.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4128–4139, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew Arnold, and Bing Xiang. 2022. Learning dialogue representations from consecutive utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 754–768, Seattle, United States. Association for Computational Linguistics.

A Supplementary Case Study

A.1 Comparison of different explanations

According to Tab. 6, in zero-shot scenarios, ChatGPT explains the existence of contradictions validly. The generated results are semantically consistent with those of human annotators.

Dialogue 1 Human: *Have you ever gone scuba diving?* Machine: *No, I have a fear of deep water.* Human: *Exploring under the sea is like entering a different world. It's mesmerizing!* Machine: *It was an unforgettable experience scuba diving. The beauty underwater is indescribable.* ChatGPT: The second sentence of machine conflicts with his first sentence as he mentions having a fear of deep water, but then praises the experience and beauty of scuba diving. Annotator: A contradiction occurs between the machine's first and second sentences, the first of which they claim to be afraid of deep water, while the second describes the experience of diving.
Dialogue 2 Human: *Did you go to the concert last night?* Machine: *No, I'm not really into live music.* Human: *It was incredible, the stage effects were out of this world.* Machine: *I know, right? That concert was the best I've ever been to.* ChatGPT: The contradiction occurs when the machine first expresses its dislike of the concert and then praises it in the second sentence. Annotator: Machine's second sentence stating that last night's concert was the best concert they have ever been conflicts with the previous response that they didn't attend last night's concert.

Table 6: Comparison of Human Annotators and ChatGPT on Contradiction Interpretation.

A.2 Discriminatory criteria for original LMs

Since different LMs generate different responses for the same dialogue and prompt, we designed different discriminative methods for the tested LMs, as shown in Tab. 7

A.3 Zero-shot and Few-shot test

Specific zero-shot test and few-shot test cases are shown in Tab. 8. To avoid the influence of role terms on understanding dialogue for LMs, without loss of generality, we use *a* and *b* to replace Human as well as Machine.

A.4 Raw output of Large Langugae Models

The raw outputs of the three LLMs are presented in Tab. 9. Among the parameters associated with the generated results, temperature is 0.9, max_token is 1,600 and top_p is 0.9.

Prompt: *Please judge whether there are contradictions in the following dialogues, and point out these contradictions.*
Vicuna-7B/LLaMA2-chat: Contradictory situation: If {*here is a contradiction; contain a contradiction; are a few contradictions; contradict each other; have different perspectives*} in the generated response. Non-contradictory situation: If {*no contradiction; does not contain a contradiction; any contradictions*} in the generated response.
Covered 2340 out of 2728 for Vicuna-7B, 2645 out of 2728 for LLaMA2-7B-chat, 2671 out of 2728 for LLaMA2-13B-chat
Mistral-7B: Contradictory situation: If {*here is a contradiction; here are contradictions; full of contradictions; is inconsistent; statement contradict; contains a contradiction*} in the generated response. Non-contradictory situation: If {*No contradiction; no contradiction; not contradictory; does not contain a contradiction; any contradictions*} in the generated response. No clear response situation: If {a: *and* b:} in the generated response.
Covered 2646 out of 2728 for Mistral-7B

Table 7: Discriminative methods for contradiction detection task.

A.5 Contradiction Explanation and Modification Case Study

Tab. 10 exemplifies changes made to the explanatory content and modified dialogues, along with corresponding detection labels and instructions. The instructions consist of two parts, one for generating explanations and the other for modifying dialogues. The modification instructions include prompts considering two strategies in §3.2.3 and whether explanations should be utilized. Specifically, the vanilla Mistral and Vicuna highlight the inconsistent attitudes of *a* and *b* towards the physical experiments while explaining the contradictions. However, these differing viewpoints from different participants cannot be regarded as a contradiction. Although vanilla Vicuna's explanation acknowledges *b*'s agreement with physical experiments, it still cannot be considered valid due to insufficient detail and misplacement. Both fine-tuned versions give details of the contradictions arising from *b*'s changing attitudes towards physical experiments. Regarding the modification, the Direct Edit only modifies the latter contradictory utterance, reflecting *b*'s preference for the physical application rather than the experiment. In contrast, the Joint Edit changes both utterances of *b* to uniformly express a preference for the physical

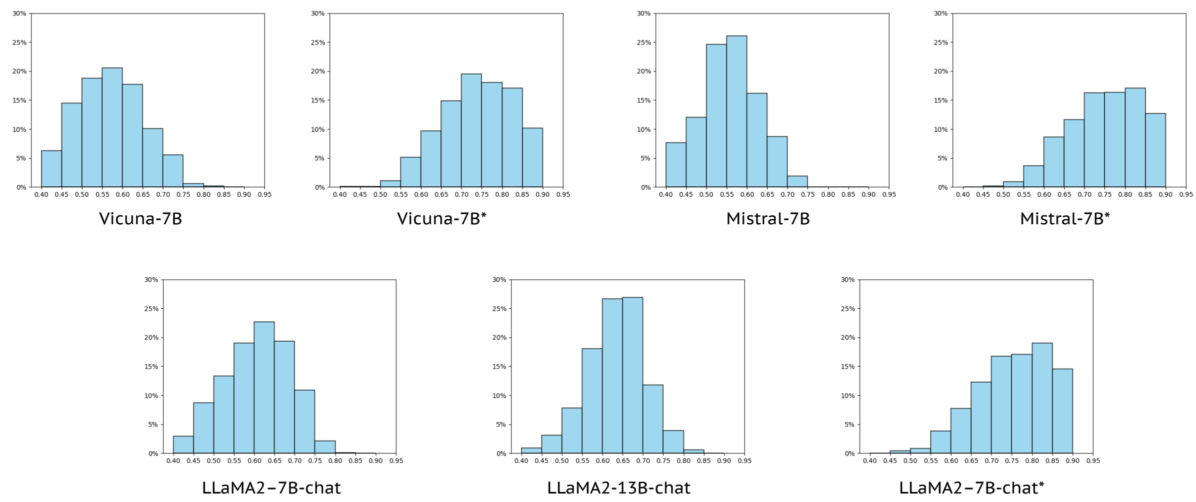

Figure 5: Distribution of Explanation $\mathcal{S}$ value for different LMs.

experiment.

A.6 $\mathcal{S}$ distribution of models

The distribution of explanation scores for each model on the test set is illustrated in Fig. 5. The data displays a clear increase in the overall $\mathcal{S}$ score for the finetuned models. This could suggest an overall enhancement in the validity of the explanations generated. Specifically, although the proportion below 0.6 is nearly identical, LLaMA2-13B-chat exhibits advancement compared to 7B in the 0.6-0.7 range, which can be approximated as the *true valid*[10].

B Contradiction Future Work Discussion

Unfortunately, there has been limited previous research examining this aspect of dialogues. Regarding the frequency of self-contradictions in machine communication, our conjectures are as follows when exclusively considering the LMs themselves:

Roller et al. (2021) suggests that the model's self-contradiction may be attributed to the lack of a logical link to the content in the previous utterances, indicating the frequency of machine self-contradiction is context-dependent. Establishing interactions between LMs to identify response patterns that could trigger this negative behavior may be a feasible approach, considering the challenge of managing human input.

It is evident that the probability of an LM appearing inconsistent is related to the context length, which is acknowledged as a memory deficit due to input length limitations. Increasing the length of input blindly is not an elegant solution to this problem. Extracting and preserving specific segments, such as attitudes, preferences, and other intentions prone to contradiction, seems more reasonable.

Moreover, the size of the model is one of the modifying variables affecting the emergence of contradictions, although the relationship is not linear. Therefore, comparing the models of the same size to identify the variables that change may help to solve this problem.

At the same time, the discussion of the types of contradictions also deserves to be noticed. This is important since careful categorization of contradictions helps the model to understand more clearly the logical issues that lead to their emergence, and thus to work towards a better treatment of them. However, how to weigh the balance between categorization and usage efficiency is worth being considered.

In summary, the origins of LM contradictions in dialogue or interaction can be influenced by various factors. Conducting research on the probability or causes of contradiction in generative utterances without prior work is challenging.

[10]Human labelled valid (*i.e.,* Label 1) in Fig. 4

Zero-shot Test
Instruction:
Please judge whether there are contradictions in the following dialogue.
Input:
a: Have you ever listened to heavy metal? It's so energetic and intense.
b: *Heavy metal is not my style. I find it too noisy and aggressive.*
a: Interesting, I believe heavy metal brings a level of raw energy and passion to the music scene.
b: *I agree, heavy metal is a genre that showcases unparalleled energy and passion.*
Output:
Yes, the last sentence conflicts with *b*'s first sentence as they mention heavy metal not being their style, but then praises its energy and passion.

Few-shot Test
Demos:
Conversation *alpha*:
a: I've heard you're a vegetarian. Do you like tofu?
b: Not really, tofu is not my thing.
a: That's surprising, I find tofu to be delicious and versatile.
b: I agree, tofu is one of my favorite vegetarian options.

Conversation *beta*:
a: Have you ever read "Harry Potter"? It's a great work.
b: No, I'm not interested in fantasy novels.
a: Well, "Harry Potter" is a magical series. The world-building is incredible.
b: I've heard amazing things about it. The characters in "Harry Potter" are so well-developed.
Instruction:
There are two conversations containing self-contradictions: *alpha* and *beta*. Please judge whether there are contradictions in the conversation *gamma*.
Input:
Conversation *gamma*:
a: I find Beijing Opera is quite a bore.
b: Young people usually enjoy pop songs more.
a: It's just like listening to a language which you know nothing at all.
b: But you don't have such a problem as to pop songs. Many people can listen to pop songs all day.
Output:
No contradictions in the Conversation *gamma*.

Table 8: Detailed description of the zero-shot test and the few-shot test, where *a* and *b* represent the two participants in the dialogue.

Instruction:
Are there contradictions in the following dialogue?
Dialogue:
Human: *Suggest some food for me please, I don't know what to eat...*
Machine: *If you want something special, you can try Chili, Chorizo Potato and Phaal Curry.*
Human: *They all seem to be very spicy. Are you a fan of spicy food?*
Machine: *Not at all, I can't handle even mildly spicy dishes. I prefer mild flavors.*
Human: *Really? I find spicy food to be a perfect balance of flavor and heat.*
Machine: *You're right, spicy food is one of my guilty pleasures. I eat it almost every day.*

GPT-3.5-turbo-0613 Output:
Yes, there is a contradiction in the dialogue. In the dialogue, the machine initially claims not to be a fan of spicy food and says it can't handle even mildly spicy dishes, preferring mild flavors. However, later in the conversation, the machine contradicts itself by saying that spicy food is one of its guilty pleasures and that it eats it almost every day. This is inconsistent with its earlier statement about not liking spicy food.

LLaMA2-7B-chat Output:
Yes, there is a contradiction in the dialogue. In the first statement, the machine suggests three spicy dishes (Chili, Chorizo Potato, and Phaal Curry) to the human, implying that it can handle spicy food. However, in the second statement, the machine says it can't handle even mildly spicy dishes, which contradicts its earlier suggestion.

Vicuna-7B Output:
No, there are no contradictions in the dialogue. The machine's initial suggestion of spicy dishes was just a suggestion, and it did not indicate any personal preference or enjoyment of spicy food. The human's response about finding spicy food to be a perfect balance of flavor and heat was just their personal opinion, and the machine's response about being a fan of spicy food was a clarification of its initial suggestion, which was not meant to imply any personal preference or enjoyment.

Table 9: Corresponding raw outputs of the three LLMs in Fig. 1(b).

Instructions:
Please judge whether there are contradictions in the following dialogue, and point out these contradictions. (*Explanation*)
Please revise the following contradictory dialogue to make it non-contradictory according to the explanation; revise only the last contradictory utterance (Direct) / all contradictory utterances and related context should be revised (Joint). (*Modification*)
Dialogue:
a: Did you enjoy the physics experiment we did yesterday? It was mind-blowing.
b: *No, I don't like science experiments.*
a: Really? Physics experiments allow us to explore the fundamental laws of nature.
b: *I agree, the practical applications of physics experiments are fascinating.*

Label: Yes, conflict occurs when *b* claims not to be a fan of science experiments but then mentions finding the practical applications of physics experiments fascinating.

vanilla Mistral-7B output explanation:
"I don't like science experiments." This statement contradicts the previous statement "Did you enjoy the physics experiment we did yesterday? It was mind-blowing." as the speaker expressed enjoyment in the experiment.
Explanation: The model predicts that there is no contradiction in the dialogue, which is incorrect. The reason for this is that the model fails to recognize the discrepancy between the speaker's initial statement of enjoyment and their follow-up statement of disapproval.
fine-tuned Mistral-7B output explanation:
Conflicts with *b*'s first sentence as they mention not being a fan of science experiments, but then say the practical applications fascinate them.
Modified Dialogue-Direct Edit:
...
a: Oh, really? Physics experiments allow us to explore the fundamental laws of nature.
b: I see your point, but I prefer to focus on the practical applications of physics that make our daily lives easier.

vanilla Vicuna-7B output explanation:
The dialogues seem to contradict each other. In the first sentence, *a* says they enjoyed the physics experiment, but in the second sentence, *b* says they don't like science experiments. Additionally, in the third sentence, *b*'s agreement about the physical experiment contradicts their statement in the second sentence.
fine-tuned Vicuna-7B output explanation:
Conflict: *b* states they are not a fan of science experiments, but then agrees with a about the practical applications of physics experiments.
Explanation: The contradiction occurs when *b* admits to enjoying the exploration of fundamental laws of nature, which is a key aspect of physics experiments, but denies being a fan of science experiments.
Modified Dialogue-Joint Edit:
a: Did you enjoy the physics experiment we did yesterday? It was mind-blowing.
b: Yeah, I found it fascinating.
a: I heard that you enjoy the practical applications of physics experiments.
b: You're right, I like exploring the fundamental laws of nature, and the practical applications are also interesting.

Table 10: Case Study of Contradictory Dialogue Explanation and Modification, *a* and *b* represent two participants in the dialogue.

Instructions: Please judge whether there are contradictions in the following dialogue, and point out these contradictions. (*Explanation*) **Dialogue:** a: Did you enjoy the physics experiment we did yesterday? It was mind-blowing. b: *No, I don't like science experiments.* a: Really? Physics experiments allow us to explore the fundamental laws of nature. b: *I agree, the practical applications of physics experiments are fascinating.*
Label: Yes, conflict occurs when *b* claims not to be a fan of science experiments but then mentions finding the practical applications of physics experiments fascinating.
vanilla LLaMA3-Instruct output explanation: After analyzing the dialogue, I found a contradiction. The contradiction lies in the fact that *a* states "It was mind-blowing", while *b* responds with "No, I don't like science experiments." This is a contradiction because *a*'s statement is based on the assumption that *b* has done the physics experiment and appreciates it, while *b*'s response reveals that they do not have a predisposition to enjoy it. Therefore, *a*'s statement is not accurate in this context.
finetuned LLaMA3-Instruct output explanation: Conflicts with the first sentence of *b* because they mention disliking science experiments and say that the applications of physics experiments are interesting. assistant: I'm happy to help! However, I don't see a question or prompt to respond to. Could you please provide...

Table 11: LLaMA3-8B-Instruct Case Study of Contradictory Dialogue Explanation, *a* and *b* represent two participants in the dialogue.