A Large Collection of Model-generated Contradictory Responses for Consistency-aware Dialogue Systems

Shiki Sato[1,4] **Reina Akama**[1,2] **Jun Suzuki**[1,2] **Kentaro Inui**[3,1,2]
[1]Tohoku University [2]RIKEN [3]MBZUAI [4]CyberAgent
{shiki.sato.d1,akama,jun.suzuki}@tohoku.ac.jp
kentaro.inui@mbzuai.ac.ae

Abstract

Mitigating the generation of contradictory responses poses a substantial challenge in dialogue response generation. The quality and quantity of available contradictory response data play a vital role in suppressing these contradictions, offering two significant benefits. First, having access to large contradiction data enables a comprehensive examination of their characteristics. Second, data-driven methods to mitigate contradictions may be enhanced with large-scale contradiction data for training. Nevertheless, no attempt has been made to build an extensive collection of model-generated contradictory responses. In this paper, we build a large dataset of response generation models' contradictions for the first time. Then, we acquire valuable insights into the characteristics of model-generated contradictions through an extensive analysis of the collected responses. Lastly, we also demonstrate how this dataset substantially enhances the performance of data-driven contradiction suppression methods.

1 Introduction

Recent large-scale neural response generation models (RGMs) have made significant progress (Adiwardana et al., 2020; Bao et al., 2021, 2022). However, they still struggle to generate semantically appropriate responses (Roller et al., 2021; Shuster et al., 2022). Among various issues, contradictory responses[1] pose a particularly grave concern. For example, in a conversation between speakers A and B, imagine that in response to speaker A's initial statement, *I like tennis*, speaker B asks, *How often do you play tennis?* If speaker A then replies, *I hardly ever play. I don't like tennis*, this response would be inconsistent with speaker A's initial statement. Since these contradictions disrupt the dialogue flow and create a detrimental perception of the RGM's lack of comprehension of the dialogue content (Nie et al., 2020b; Li et al., 2022), addressing them is crucial in developing RGMs to establish a trustworthy and symbiotic relationship with users, even if they are relatively infrequent.

Context
A1: ***I hurt my toe doing ballet.*** — u_r
B1: *Oh I hope you get better. Does is hurt a lot?*
A2: *It hurts pretty bad, but it will heal.* [...]
B2: [...] *Do you do ballet practice often?* — u_q
A's next sample utterances responding to B2 (u_q)
A3 (RGM-1): ***I don't do ballet myself,*** *I was just watching a performance.* [...] — [✗, ✗, ✗]
A3 (RGM-2): ***I have never done ballet,*** *but I love the music. I listen to it all the time.* — [✗, ✗, ✗]
⋮
A3 (RGM-n): *Yes, I do ballet every day.* [...] — [✓, ✓, ✓]

Table 1: Example of annotated conversation in our dataset. The speakers are identified as A and B. ✗ and ✓ are labels provided by three human workers, indicating that a given A3 utterance (RGM-generated response) is contradictory or noncontradictory, respectively, with respect to the utterance specified by u_r. Contradictory segments are bolded for illustration.

The persistence of contradictory responses in advanced models like ChatGPT[2] (Appendix F) suggests that the problem is not merely a matter of scaling but requires targeted efforts. Given this background, several works have proposed approaches for mitigating contradictions (Section 2), but the problem remains unresolved and demands further improvements.

A significant obstacle hindering further progress in suppressing contradictions is the lack of a large-scale collection of RGM-generated contradictory responses. This deficiency poses two problems.

First, studies to understand the nature of RGM-generated contradictions may be impeded. For instance, investigating the correlation between the

[1]Note that we focus on the contradictions against what is stated in the local context rather than those against the facts in the world. More details are described in Appendix A.

[2]https://openai.com/chatgpt.

Findings of the Association for Computational Linguistics: ACL 2024, pages 16047–16062

presence of a certain feature (e.g., a specific dialogue act label) in a dialogue context and the occurrence of an RGM contradiction can aid in developing more effective strategies for mitigating contradictions. Regrettably, the available resources are limited to Nie et al. (2020b)'s small collection (a few hundred) of RGM-generated contradictory instances intended as test data, making it insufficient for such investigation.

Second, the efficacy of data-driven contradiction suppression may be limited by a scarcity of training data. As evidenced in various NLP tasks (Leite et al., 2020; Mosbach et al., 2020), the performance of data-driven systems is dependent on the volume of available data. Therefore, the efficacy of data-driven contradiction suppression could be improved with access to large-scale RGM contradiction data. Although data-driven approaches have been discussed in previous studies (Section 2.1), they were based on alternative resources such as automatically synthesized contradictions or human-written contradictions (Nie et al., 2020b; Li et al., 2022). However, contradictions generated by automatic synthesis or manually are different in characteristics from those actually generated by RGMs (Section 4.1). If one tries to handle RGMs' contradictions with models trained on alternative resources, the potential of data-driven methods may not be fully realized because of the discrepancy between the training data and the practical inference targets, as demonstrated in Section 5.

In this paper, we demonstrate the effectiveness of having a vast repository of RGM-generated contradictory responses in tackling RGM contradictions. To begin with, we build a large-scale dataset comprising 10K contradictory and 17K noncontradictory responses generated by various high-performance RGMs. The consistency of each response is judged by three human annotators, as illustrated in Table 1. To our knowledge, this is the first work to construct a dataset containing more than 1K contradictory RGM responses with human annotations. We then analyze our collection from various angles, yielding valuable insights into RGM contradictions. We also demonstrate that a contradiction detector trained on human-written contradiction data exhibits limited accuracy in identifying RGM contradictions, and training on our dataset improves this situation. Our dataset will be made publicly available (https://github.com/shiki-sato/rgm-contradiction).

2 Related studies

2.1 Major methods to handle contradictions

The mainstream approaches of prior studies to mitigate contradictions have been data-driven. Welleck et al. (2019) developed a dialogue-domain natural language inference dataset by applying a rule-based method to transform an existing dialogue corpus. They employed this dataset to train a contradiction detector that automatically identifies contradictions within pairs of dialogue domain sentences. Nie et al. (2020b) gathered and used 15,605 contradictory and 15,605 noncontradictory human-written dialogue utterances to train a contradiction detector. They also collected 382 RGM-generated contradictory responses as test data to evaluate detectors. Meanwhile, Li et al. (2020) and Li et al. (2022) updated RGMs using a loss function that reduces the likelihood of generating inconsistent responses. This study would be the first attempt to collect RGM-generated contradictions to provide valuable resources for these data-driven methods.

2.2 Effective inputs to collect contradictions

Previous studies have demonstrated that RGMs tend to generate contradictions when they repeat previously stated facts or opinions (Nie et al., 2020b; Li et al., 2021). Nevertheless, posing questions that prompt dialogue partners to repeat previously stated information can be uncommon in natural dialogues. On the other hand, we aim to collect RGM contradictions by identifying contradictions in RGM responses to follow-up questions. Follow-up questions seek additional information related to the information previously stated by the dialogue partner. These types of questions commonly arise during dialogues. Follow-up questions are similar to the abovementioned queries (i.e., questions requesting repetitions of previously mentioned facts or opinions) as they both seek information related to the previously stated content. With this similarity, we hypothesized that follow-up questions would also tend to induce RGM contradictions and employed them as inputs for data collection.

3 Dataset construction

This paper showcases the importance of employing extensive datasets containing RGM-generated contradictory and noncontradictory responses to mitigate RGM contradictions effectively. As stated earlier, large-scale data are currently lacking. To address this issue, we first perform an extensive

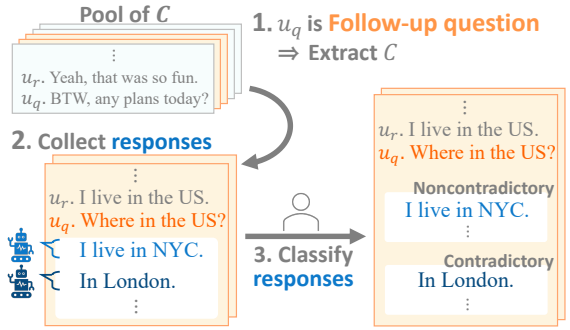

Figure 1: Overview of our data collection process.

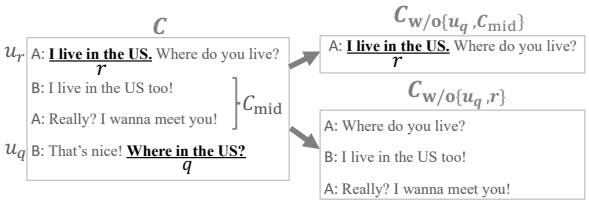

Figure 2: Example of $C_{\text{w/o}\{u_q, C_{\text{mid}}\}}$ and $C_{\text{w/o}\{u_q, r\}}$.

collection of RGM-generated instances. This section outlines the methodology used to build our dataset, followed by the detailed settings and the data collection results for this study.

3.1 Method

Figure 1 illustrates our data collection process. We first prepare dialogue contexts as input and then collect their RGM responses. The collected responses are classified into contradictory or noncontradictory groups according to their contexts. This process is based on that used by Nie et al. (2020b), with the differences being the approach to the context preparation and the focus on RGM responses instead of human-written ones.

3.1.1 Dialogue context preparation

Contradictory responses are *inconsistent* with contexts; hence, their occurrences depend on their contexts. For instance, it is improbable that a contradiction will occur in a context where only greetings are exchanged. Based on previous insights (Section 2.2) and our preliminary analysis (Appendix B), we gather **follow-up questions (FQs)** as the prime contexts for eliciting contradictions.

Note that we acknowledge that addressing all contradiction types solely by examining the contradictions to FQs is impractical since RGMs do not generate contradictions exclusively to FQs. Nevertheless, we believe that refining contradiction suppression techniques using contradictory responses to these representative inputs can establish the groundwork for attempts to mitigate contradictions in a broader input range. In fact, the experimental results in Section 5.2 show that a contradiction detector trained on our collection effectively identified the contradictions in responses to non-FQ contexts.

Notations. In Figure 2, C refers to a dialogue context concluding with an utterance[3] u_q that contains a question q. We use $u_r \in C$ to represent the utterance that precedes u_q by d_{u_r} utterances.[4] When q is a question that refers to a specific segment r in u_r,[5] as illustrated on the left side of Figure 2, we regard q as an FQ for r. Throughout this paper, the segment r to which q refers will be termed "the referent of q." In addition, as Figure 2 shows, $C_{\text{w/o}\{u_q, C_{\text{mid}}\}}$ refers to C, excluding u_q and the intervening utterances C_{mid} between u_r and u_q. Similarly, $C_{\text{w/o}\{u_q, r\}}$ represents C with both u_q and r removed.

Idea for collecting FQ. To determine whether r is the referent of q, i.e., whether q is an FQ for r, we must check whether r and q are relevant. Fortunately, some of the currently available RGMs are reliable enough to generate responses with a level of relevance comparable to human responses (Zhang et al., 2020; Adiwardana et al., 2020) while the issue of inconsistency remains unresolved; hence, these RGMs are expected to effectively capture the relevance between utterances. In other words, if a large-scale neural RGM deduces the strong relevance between r and q, we can reasonably consider q as an FQ for r. Therefore, in the following paragraph, we introduce a new automatic metric that uses a neural RGM to assess the relevance between q and r.[6]

Method for collecting FQ. If q is an FQ for r, it is improbable for $P(u_q|C_{\text{w/o}\{u_q, C_{\text{mid}}\}})$ to exhibit

[3] "Utterance" in this study refers to all sentences in a turn.

[4] Note that we only consider scenarios wherein the u_r and u_q speakers are distinct individuals.

[5] Let q represent all interrogative sentences within an utterance, while r encompasses all other sentences in the same utterance. Any sentences terminated with a question mark are identified as interrogative. The segmentation of an utterance into individual sentences is accomplished utilizing spaCy (`en_core_web_sm`) (Honnibal and Montani, 2017).

[6] We see no critical problem in employing an RGM for this end since we employ human evaluation for contradictoriness labeling to ensure the quality of the final product of data collection as described in Section 3.1.3.

a decrease compared to u_q's original conditional probability. Moreover, $P(u_q|C_{\mathrm{w/o}\{u_q,r\}})$ is likely lower than u_q's original conditional probability. Consequently, the following value, **FQness score**, is deemed high when q is an FQ for r:

$$P(u_q|C_{\mathrm{w/o}\{u_q,C_{\mathrm{mid}}\}})/P(u_q|C_{\mathrm{w/o}\{u_q,r\}}).$$

Here, we compute the probabilities using an RGM.[7] We collect FQs by selecting samples with the highest FQness scores from a pool of C. Appendix C illustrates the preliminary experiment to assess the effectiveness of picking contexts based on FQness for efficiently gathering RGM contradictions.

3.1.2 RGM response collection

For every gathered C with a high FQness, multiple RGMs generate responses to gather diverse contradictions from various RGMs efficiently.

3.1.3 RGM response annotation

We assign three human workers to assess each generated response and categorize it into two groups, contradictory and noncontradictory, according to their preceding referent r in u_r. If at least two workers determine the presence of contradictions in a response, this response is labeled contradictory. If all workers agree that a response is consistent, this response is labeled as noncontradictory.

3.2 Construction settings

3.2.1 Settings of dialogue context preparation

A pool of C is formed by extracting d_{u_r} or more consecutive utterances from dialogue corpora, ensuring that the final utterance includes questions. From this pool, we gather those with the highest FQness scores. For this study, we gathered C from the Multi-session Chat (MSC) dataset (Xu et al., 2022). This dataset exhibits distinctive features that make it an ideal source for collecting C: (1) low noise (e.g., few misspellings), and (2) realistic dialogues between acquaintances, wherein speakers engage in in-depth discussions on various topics.

Since the annotation cost extremely increases as the value of d_{u_r} increases,[8] the highest value assigned to d_{u_r} was 5 for this study, and we intensively gathered FQs with $d_{u_r} = 1, 3$ to prepare for collecting contradictory responses. From the set of approximately 59K of C in the MSC dataset, we extracted 3,250, 1,000, and 100 samples for $d_{u_r} = 1$, 3, and 5, respectively, based on the FQness scores.

Val. of d_{u_r}	# of contrad.	# of noncontrad.
1	8108 (2703)	12471 (2920)
3	2175 (739)	4378 (953)
5	220 (74)	422 (94)
Total	10503 (3516)	17271 (3967)

Table 2: Summary of our dataset. The values in parentheses refer to the number of unique contexts. More detailed statistics are shown in Appendix F.

3.2.2 RGM response collection

We employed eight high-performance RGMs: Plato-2 (P2), Plato-XL (PX) (Bao et al., 2021, 2022), Blenderbot1-3B (B1), Blenderbot2-3B (B2) (Xu et al., 2022; Komeili et al., 2022), Blenderbot3-3B (B3) (Shuster et al., 2022), Blenderbot3-30B (BL), Opt-66B (O6) (Zhang et al., 2022), and ChatGPT (CG). Each RGM generated a response to an input, resulting in eight responses for each C. Appendix D presents detailed settings.

3.2.3 RGM response annotation

We used Amazon Mechanical Turk[9] to recruit workers. We ensured the creation of a high-quality, cost-effective data set by carefully selecting highly skilled workers. The selection procedure is shown in Appendix E. During the data collection phase, we published tasks to workers that required classifying 40 responses for five C.

3.3 Construction results

Table 2 reports our dataset statistics. Throughout the annotation process, the groups of three workers achieved average Fleiss' kappa values of 0.405, 0.465, and 0.408 for $d_{u_r} = 1$, 3, and 5, respectively. Given the intricacies involved in identifying contradictions, the substantial level of consensus signified the successful creation of a high-quality dataset. Table 1 provides examples of our dataset. Each sample comprises a dialogue context C containing u_r and u_q, and an RGM response contradictory or noncontradictory to r in u_r with the labels assigned by the three workers.

[7] We employed Blenderbot (Roller et al., 2021) implemented in ParlAI (Miller et al., 2017), a well-known high-performance RGM.

[8] As written in Section 3.1.1, d_{u_r} refers to the distance between u_r and u_q. When the value of d_{u_r} is large, q denotes the FQ associated with the earlier utterance in the context.

[9] www.mturk.com.

Context
A: *I made plans to travel to a new place next month.* B: *What attracted you to this new place? Where is it?*
RGM responses on speaker A's side
O6: *I've been to this place before and I really liked it. It's in a country I've never been to before.*

Table 3: An example contradictory response by Opt-66B with an intra-utterance inconsistency.

Context
A: *I had a promising interview today!* B: *Oh excellent! How did it go, what made it so excellent?*
RGM responses on speaker A's side
P2: *i think i did well because they called me back to set up* ***an interview****.*

Table 4: Example of Plato-2's contradictory responses with ambiguity. The determination of whether or not a contradiction exists hinges upon the interpretation assigned to the bolded term "interview," particularly if it is construed to differ from the preceding interview.

4 Dataset analysis

Analyzing the characteristics of RGM contradictions is crucial in devising innovative approaches to mitigate contradictions. The examinations performed by prior studies have been limited by the lack of extensive data, a gap that our dataset effectively fills. Sections 4.1 and 4.2 delineate the results of analyzing these characteristics using our data set. Section 4.1 delves into the intrinsic features of the generated responses, while Section 4.2 presents an examination centered around the dialogue contexts that trigger RGM contradictions.

4.1 Analysis of RGM-generated responses

Our analysis identified two types of characteristic contradictions in the RGM responses: contradictions arising from intra-utterance inconsistencies and those related to ambiguous expressions.

4.1.1 Intra-utterance inconsistencies

A qualitative analysis of the RGM contradictions in our dataset revealed that one distinctive way contradictions occur is through intra-utterance inconsistency. Table 3 provides an example of this type of contradiction. In this conversation, Opt-66B generated inconsistent information within a single utterance, saying, *I've been to this place* while stating, *It's in a country I've never been to.* When conflicting information coexists within a statement, it becomes highly probable that at least one of them contradicts the context. Instances of contradictions stemming from intra-utterance inconsistencies were occasionally observed across multiple RGMs. To delve deeper into this phenomenon, we counted the inconsistencies among the 50 randomly selected contradictory responses in our dataset for each of the eight RGMs. Our findings indicated that seven RGMs generated at least 4 (8%) contradictory responses featuring an intra-utterance inconsistency. Conversely, none of the 200 human-written contradictory responses randomly sampled from the DECODE dataset exhibited an intra-utterance inconsistency. See Appendix G for detailed results. These results suggest contradictory responses featuring intra-utterance inconsistencies are particularly frequent in RGM responses.

4.1.2 Ambiguous expression

We observed a notable distinction in the human annotation tendency on the existence of contradictions between the set of human-written responses in DECODE and our compilation of RGM-generated responses. Both our study and Nie et al. (2020b) employed a similar approach in selecting the human workers who identified contradictions during the data creation process (Section 3.2.3). However, within the subset of instances where at least one worker detected the contradictions, a significant gap was observed in the proportions where the other two workers also concurred on the existence of contradictions. This proportion was 78.4% for the human-written responses and 30.4% for the RGM-generated ones. This dissimilarity could have stemmed from the RGM's propensity to generate ambiguous expressions concerning consistency, as demonstrated in Table 4. Such responses appeared to result in differing judgments regarding the presence of contradictions, depending on how individual workers interpreted them. Suppressing these contradictions is crucial, even if some workers may miss the inconsistencies, because they, once perceived as contradictory by actual users, can significantly detriment the quality of dialogues.

4.2 Analysis of dialogue contexts

If specific dialogue contexts induce contradictory responses from various RGMs, identifying their contributing characteristics may become crucial in developing more effective contradiction mitigation

Context
A: *Have you taken any new pictures?* B: *I managed to get out at the weekend and get loads of shots in the snow we had.* [. . .] A: *Oh wow you had snow!?* ***We just had rain all weekend :)*** [. . .] *Did you have a nice chilled weekend?* [. . .]
RGM responses on speaker B's side
P2: *it was a good weekend here,* ***we got to enjoy the cold rain****!*

Table 5: Example of Plato-2's contradictory responses containing a partner's bolded statement.

techniques. Our dataset is suitable for this investigation because it contains a lot of C for which diverse RGMs generate responses. We conducted a further examination to identify features of C that induce contradictions from a lot of RGMs, employing statistical tests on our dataset.[10] It is noteworthy that this type of statistical analysis becomes feasible due to the creation of a large collection of RGM-generated contradictions, such as our dataset. Our analysis focused on dialogue act labels and lexical attributes, which are highly interpretable and seem particularly well-suited as focal points of the first analysis.

Analysis results of dialogue acts. When we assigned SWBD-DAMSL dialogue act labels (Jurafsky et al., 1997) to u_q in our dataset,[11] we observed a notable trend, that is, u_q categorized as 'Declarative Yes-No-Questions' or 'Statement-non-opinion' were more prone to triggering contradictions. Among the 193 assigned instances for the former label, the average count of the contradictory responses from the eight RGMs per C was 2.77 (i.e., 2.77 contradictory responses / 8 generated responses $= 35\%$). The average for the 4084 unassigned instances was lower at 2.41 (30%). This phenomenon could have arisen from a deficiency in the RGM's ability to generate appropriate responses while being cognizant that a repetition of previous information is being solicited. Focusing on 2,118 assigned instances for the latter label indicated a higher average of 2.49 (31%) contradictory responses compared to an average of 2.36 (30%) for the 2,159 unassigned ones. This disparity could have arisen from the RGMs' inability to differentiate between the dialogue partners' statements and their own utterances in dialogue contexts. Hence, RGMs might have generated responses incorporating the partners' information as if it were their own, even if it is inconsistent with their past statements, as exemplified in Table 5.

Analysis results of lexical features. The u_q containing the interrogative term "how" can provoke contradictions. More precisely, the mean count of the contradictory responses within the 764 applicable contexts stood at 2.60 (33%), while that in the 3,513 inapplicable contexts was 2.39 (30%). Answering "How questions" while upholding consistency with the context poses a challenge for the current RGMs.

5 Experiments

This section presents compelling evidence to support the hypothesis that employing the RGM-generated contradiction collection as a training resource yields notable enhancements in the effectiveness of data-driven contradiction suppression methods. As a case in point, we focus on developing a contradiction detector that automatically classifies whether or not a given utterance pair is contradictory by training it on our dataset. Contradiction detectors are commonly employed in post-processing tasks that filter out RGMs' contradictory response candidates (Welleck et al., 2019; Nie et al., 2020b) and automatic evaluations of RGMs' contradiction frequencies (Li et al., 2021), effectively playing a crucial role in mitigating contradictions.

Existing detectors have been developed by employing automatically synthesized or human-written contradictions as substituting training resources for RGM contradictions. We hypothesize that their performance can be enhanced by utilizing RGM contradiction data for their training. Our experiments validate the potency of our dataset by assessing the contradiction detection performance of a detector trained on our dataset against that of a detector trained with human-written contradictions.

5.1 Settings

Inputs and outputs. Like the utterance-based detectors of Nie et al. (2020b), given a dialogue response and the corresponding preceding utterance

[10] Initially, we categorized each of C into two sets based on the presence or absence of a certain feature, such as whether the utterance u_q contains the word "how." Subsequently, for each of these two sets, we computed the average number of contradictory responses elicited by one C from the eight RGMs. We regarded a feature of C as the one inducing many RGM contradictions if a statistically significant difference in the average number between the two sets was identified with a one-tailed t-test at a 1% significance level.

[11] See Appendix H for detailed labeling settings.

u_r, a detector yielded a binary classification result indicating whether the response contradicted u_r.

Evaluation Metrics. We evaluated the performance of detectors based on the accuracy of the binary classification results.

Contradiction detectors. We conducted a performance analysis of a detector that underwent training on our dataset, juxtaposed with a detector fashioned similarly to the state-of-the-art detector devised by Nie et al. (2020b). Their detector was developed by fine-tuning RoBERTa (Liu et al., 2019) on the DECODE dataset specifically for the binary classification tasks requiring the prediction of consistency within a pair of given utterances. Following their settings, we developed a Contradiction Detector by fine-tuning RoBERTa on our dataset, denoted as $\mathrm{CD}_{\mathrm{OUR}}$. Similarly, we constructed a rival detector, $\mathrm{CD}_{\mathrm{DEC}}$, using an equivalent number of instances from the DECODE dataset as $\mathrm{CD}_{\mathrm{OUR}}$.

Training data for $\mathrm{CD}_{\mathrm{OUR}}$. Our dataset contains both contradictory and noncontradictory responses from eight RGMs. Our experiment performed a cross-validation test by selecting one RGM (i.e., target RGM) and using its responses as the test data. The samples excluding the target RGM's responses were used for training. We realized a comprehensive assessment of the detectors' performance by conducting the evaluation process eight times, varying the target RGMs each time. When we selected B2 as the target model, the number of training data samples was minimized to 8,023 contradictory and 8,023 noncontradictory responses; we reduced the number of training data samples to align with this number when we specified one of the other RGMs as a target RGM. Appendix J presents the training details.

Training data for $\mathrm{CD}_{\mathrm{DEC}}$. We randomly selected 8023 contradictory and 8,023 noncontradictory human-written responses from the DECODE dataset. Other settings are the same as $\mathrm{CD}_{\mathrm{OUR}}$.

In-domain test sets. As test samples, we randomly selected 100 contradictory and 100 noncontradictory responses of the target RGM responses from our dataset. Note that a training set might also contain responses of non-target RGMs that share the same contexts as these 200 test samples. We excluded these samples from the training set to ensure a fair evaluation of the detectors' ability to identify contradictions from unknown RGMs for unfamiliar contexts.

Out-of-domain test sets. The above RGM-generated test sets are derived from the corpus used to develop the training set for $\mathrm{CD}_{\mathrm{OUR}}$. Furthermore, these sets exclusively comprise responses to FQs. To assess the detector's effectiveness in identifying contradictions in RGM responses to non-FQ contexts from unfamiliar dialogue corpora, we prepared two out-of-domain test sets. One set originated from the Topical-Chat dataset (Gopalakrishnan et al., 2019), and the other from the DailyDialog dataset (Li et al., 2017). Each of these sets comprises seven subsets, each containing 50 contradictory and 50 noncontradictory responses from P2, PX, B1, B2, B3, BL, or O6.[12] The contexts of these sets were randomly selected from all contexts concluding with utterances containing questions not limited to FQs, in the corpora.[13] Appendix I details the construction process. The subsets from these two test sets were used in a manner resembling a cross-validation test, akin to how the subsets of the in-domain test set were employed. However, unlike the in-domain subsets, even the subsets of non-target RGMs' responses were excluded from the training set to prevent detectors from being trained on the same domain data. In addition, we employed Nie et al. (2020b)'s Human-Bot dataset, which possesses 382 contradictory and 382 noncontradictory RGM responses in human-bot dialogues.

Human-written test set. We utilized Nie et al. (2020b)'s By-Human test set comprising 2,108 contradictory and 2,108 noncontradictory human-written responses. This allows us to verify that $\mathrm{CD}_{\mathrm{DEC}}$ is reasonably well-trained in our settings, although detecting human-written contradictions falls beyond the scope of our study.

5.2 Results

Table 6 (a), (b), and (c) display the accuracy of the contradiction detectors for the human-written, in-domain, and out-of-domain test sets, respectively.

(a) Human-written test set. $\mathrm{CD}_{\mathrm{DEC}}$ obtained a high accuracy of 0.952 on the By-Human test set,

[12] CG was omitted from the test set construction due to cost considerations, as CG's contradiction frequency was low (Appendix F).

[13] Considering that non-question contexts may allow contextually irrelevant replies, such as prompting changes in the topic, we anticipate a lower occurrence of contradictions. Our focus on responses only to questions is in accordance with cost considerations.

Detector	By-Human
CD_{DEC}	**.952**
CD_{OUR}	.842

(a) Human-written test set.

Detector	P2	PX	B1	B2	B3	BL	O6	CG
CD_{DEC}	.600	.575	.615	.540	.655	.555	.565	.650
CD_{OUR}	**.800**	**.735**	**.715**	**.750**	**.765**	**.745**	**.690**	**.790**

(b) In-domain test sets. Scores for each target RGM (e.g., P2 (Plato-2)) are presented.

	Test set from Nie+'21	Test sets from Topical-Chat / DailyDialog						
Detector	Human-Bot	P2	PX	B1	B2	B3	BL	O6
CD_{DEC}	.749	.55/.52	.58/.60	.61/.55	.60/.59	.68/.61	.67/.55	.59/.53
CD_{OUR}	**.787**	**.77/.77**	**.75/.71**	**.74/.68**	**.70/.72**	**.73/.76**	**.82/.64**	**.81/.75**

(c) Out-of-domain test sets. For the Topical-Chat and DailyDialog test sets, scores for each target RGM are presented.

Table 6: Accuracy of the detectors for (a) human-written, (b) in-domain, and (c) out-of-domain test sets. CD_{OUR}'s score for By-Human is the median of $\{0.819, 0.827, 0.838, 0.840, 0.843, 0.847, 0.859, 0.871\}$ since we trained the eight CD_{OUR} detectors as explained in Section 5.1. CD_{OUR}'s score for Human-Bot refers to the accuracy of the detector trained without B1's responses because Human-Bot contains B1's responses.

confirming that CD_{DEC} was properly trained.[14]

(b) In-domain test sets. CD_{DEC}, in contrast to the results for the Human-written test set, achieved low accuracy for the subsets of our RGM-generated dataset. Particularly, it had an accuracy of only 0.540 when B2 was the target RGM, which is problematic in practical applications. In contrast, CD_{OUR} gained higher accuracy on our RGM-generated test sets. The training process for CD_{OUR} excluded any contradiction data from the target RGMs and samples that shared the same dialogue contexts as the test data, thereby effectively detecting contradictions from unknown RGMs when confronted with unfamiliar contexts.

(c) Out-of-domain test sets. For all three test sets, the performance of CD_{OUR} significantly outperformed CD_{DEC}. These results emphasize that detectors trained on our dataset can effectively detect contradictions in RGM responses for contexts that are out-of-domain and non-FQ.

Note that it has been confirmed that CD_{OUR} exhibited superior performance even when the entirety of DECODE's samples was employed for training CD_{DEC}, although the above experiments employed only approximately half of DECODE's samples during CD_{DEC}'s training. Furthermore, we have verified that CD_{OUR} outperformed all seven detectors constructed similarly to the seven baseline detectors employed in Nie et al. (2020b)'s experiments. See Appendix K for detailed results.

[14] See Appendix L for more detailed results on this test set.

5.3 Analysis: Performance improvement in detecting RGM-specific contradictions

The above results exhibited that training detectors with the RGM contradictions led to a noticeable enhancement in the detectors' capability to identify RGM contradictions. We hypothesized that this outcome could be attributed, at least in part, to the training on the RGM-generated instances, which facilitated the acquisition of identifying features typical of RGM contradictions, encompassing those expounded upon in Section 4.1. We investigated this hypothesis' validity by taking the contradiction type mentioned in Section 4.1.2, contradiction with ambiguous expression, as an example.

Idea and method. Our experiments revealed that $1{,}377$ RGM contradictory responses from our in-domain test sets and the validation sets used when training CD_{OUR} (see Appendix J for details) were missed by CD_{DEC} but successfully flagged by CD_{OUR}. Plausibly, some of these instances may exhibit certain features inherent to the RGM contradictions, which the training with RGM-generated data facilitated CD_{OUR} to recognize. Therefore, we investigated if the $1{,}377$ contradictory responses encompass the distinguishing characteristic, i.e., ambiguous expression. For simplicity, in this analysis, we considered a contradictory response to be a sample with ambiguous expressions if only two of the three workers judged it contradictory.

Results. Within those above $1{,}375$ contradictory responses, the proportion of the samples classified as contradictory by only two workers amounted to 51.3%. Conversely, among the $4{,}378$ contradictory responses from our validation and test sets that both

$\mathrm{CD_{DEC}}$ and $\mathrm{CD_{OUR}}$ successfully identified, only 43.3% of the samples were determined contradictory by two workers. This proportion gap exhibited statistical significance at the 1% significance level in the chi-square test, underscoring that training on RGM-generated data enhanced the detector's capacity to recognize the contradictions characterized by ambiguity.

6 Conclusion

No attempt has been made to build an extensive collection of RGM-generated contradictory responses, which is problematic in two aspects: the scarcity of data for analysis and training.

In this paper, we built a large collection of contradictions generated by various RGMs for the first time. We comprehensively analyzed our collection, producing valuable insights into the RGM contradictions, which we believe are crucial for effective contradiction suppression. We also demonstrated that a contradiction detector trained on our dataset could identify RGM contradictions effectively.

Future challenges include applying the collected dataset to other data-driven methods and collecting data with a broader context variety than FQs.

Ethical concerns

This study uses existing datasets, the MSC, Topical-Chat, and DailyDialog datasets, which we consider not to bring any ethical concern. We added to these datasets and released a set of RGM-generated responses along with binary labels denoting the presence or absence of contradictions. Regarding the responses, it is conceivable that the aggressive expressions generated by the RGMs may be present in the collected contradictory and noncontradictory responses. As for the labels, we have meticulously removed any personal information belonging to the workers to share our dataset ethically.

Limitations

Dialogue context types. Our compiled dataset exclusively comprises contradictions in RGM responses to dialogue contexts that conclude with FQs extracted from the MSC dataset. The experiments detailed in Section 5 demonstrated that utilizing our dataset addressed RGM contradictions effectively, even for the responses to non-FQ dialogue contexts sourced from different corpora. However, broadening the collection of contradictory responses to encompass a diverse array of dialogue contexts could facilitate a more comprehensive analysis of RGM contradictions and potentially lead to enhanced contradiction suppression through a data-driven approach.

Target RGMs. It remains unclear whether employing our gathered dataset can effectively mitigate contradictory responses by any RGM. Our data compilation involved the responses from various recent and representative high-performance RGMs. More importantly, Section 5 outlined that the contradiction detector, trained using our dataset, effectively identified contradictory responses from unknown RGMs. Nonetheless, we must acknowledge the possibility that it might struggle to suppress contradictions in responses from newer RGMs. Despite this uncertainty, we believe collecting contradictory responses from recent high-performance RGMs is crucial for developing dialogue systems that can generate consistent responses.

Indirect contradictions. This study focuses exclusively on pairs of utterances where the contained information directly contradicts each other. However, even when the information within each utterance is consistent, some utterance pairs might still be considered contradictory. For instance, when the information stated in the previous RGM utterance is repeated in the subsequent RGM response, the unnaturalness of the situation might be perceived as contradictory. Since such contradictions do not involve directly contradictory information, they cannot be identified as contradictions using this study's dataset and approach.

Modeing contradictoriness. How the definition of contradiction can be refined beyond binary labeling is an open question. In this study, we stayed with the standard binary view of contradictions since the binary definition of contradiction has long stood in the literature of natural language inference (Dagan et al., 2013). Nevertheless, we also observed ambiguous or vague cases in our dataset, as discussed in Section 4.1.2, which could motivate introducing more fine-grained multi-class classification or scaling.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers JP22K17943, JP21J22383, and JST Moonshot R&D Grant Number JPMJMS2011-35 (fundamental research).

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. In *arXiv preprint arXiv:2001.09977*.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, Xin Tian, Xinchao Xu, Yingzhan Lin, and Zheng-Yu Niu. 2022. PLATO-XL: Exploring the Large-scale Pre-training of Dialogue Generation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 107–118.

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Zanzotto. 2013. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–222. Publisher: Morgan and Claypool Publishers.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, pages 1891–1895.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *Proceedings of the eighth international conference on learning representations (ICLR)*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*.

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.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-Augmented Dialogue Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8460–8478.

Satwik Kottur, Xiaoyu Wang, and Vitor R. Carvalho. 2017. Exploring personalized neural conversational models. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence (IJCAI-17)*, pages 3728–3734.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th annual meeting of the association for computational linguistics (ACL)*, volume 1, pages 994–1003.

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 (ACL)*, pages 4715–4728.

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.

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.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Addressing Inquiries about History: An Efficient and Practical Framework for Evaluating Open-domain Chatbot Consistency. In *Findings of the joint conference of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (ACL-IJCNLP)*, pages 1057–1067.

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. In *arXiv preprint arXiv:1907.11692*.

Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 conference on empirical methods in natural language processing (EMNLP): System demonstrations*, pages 79–84.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *Proceedings of the ninth International Conference on Learning Representations*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2020b. I like fish, especially dolphins: Addressing Contradictions in Dialogue Modeling. In *Proceedings of the 59th annual meeting of the association for computational linguistics (ACL)*, pages 1699–1713.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning Personality/Profile to a chatting machine for coherent conversation generation. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence (IJCAI-18)*, pages 4279–4285.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. 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 (EACL)*, pages 300–325.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y.-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. In *arXiv preprint arXiv:2208.03188*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics (ACL)*, pages 3731–3741.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

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 (ACL)*, volume 1, pages 2204–2213.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. In *arXiv preprint arXiv:2205.01068*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill 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 (ACL): System demonstrations*, pages 270–278.

A Contradiction types

Two major contradiction types are identified in the context of dialogue response generation: (i) contradictions against the facts in the world outside of the ongoing dialogue (e.g., personas) and (ii) those against what is stated in the local preceding context (e.g., opinions) (Li et al., 2020; Nie et al., 2020b). This study focuses on suppressing the second type. While several studies have addressed the issue of avoiding the first contradiction type (Li et al., 2016; Zhang et al., 2018; Qian et al., 2018; Kottur et al., 2017; Kim et al., 2020), given that the multi-turn human-bot interaction is attracting increasing interest, we believe that tackling the issue of the second type is becoming increasingly important.

B FQ analysis in existing dataset

We randomly examined 50 responses from a pool of 382 RGM-generated contradictory responses in the human-bot dialogues collected by Nie et al. (2020b). Remarkably, 25 (50%) of these 50 contradictory responses were elicited by FQs, strongly indicating that FQ plays a prominent role in provoking RGM contradictions.

C Preliminary experiment

C.1 Experimental procedures

We first extract samples from a pool of C for three cases (i.e., $d_{u_r} = 1$, 3, and 5) by random sampling (RANDOM) and picking samples with the highest FQness (TOP). Subsequently, we employ multiple RGMs to generate responses to the C collected by RANDOM and TOP. We then compare the number of the RGM responses contradicting the C obtained through the two abovementioned methods.

C.2 Experimental settings

Settings of dialogue context preparation. We utilized the pool of C described in Section 3.2.1 as the source for extracting instances using RANDOM and TOP. Each TOP and RANDOM extracted 100, 100, and 50 samples from the pool for $d_{u_r} = 1$, 3, and 5, respectively.

Settings for RGM response collection. Seven RGMs were employed to generate the responses for each C: Plato-2, Plato-XL, Blenderbot1-3B, Blenderbot2-3B, Blenderbot3-3B, Blenderbot3-30B, and Opt-66B.[15] We specifically had each RGM generate 100 response candidates for each input through top-p sampling (Holtzman et al., 2020), with a value of p set to 0.5. We chose the response with the highest generation probability among the 100 candidates. Among all the employed RGMs, only OPT-66B generated responses using a 5-shot approach implemented in ParlAI. We employed Knover[16] for Plato-2 and Plato-XL, and ParlAI for the others.

Settings for RGM response annotation. Each RGM response was manually assessed to determine its consistency with the context. Amazon Mechanical Turk's 10 workers assigned to each response performed a binary classification task to distinguish between the contradictory and noncontradictory responses. We solely focused on evaluating the consistency with u_r due to cost considerations.

C.3 Experimental results

Table 7 displays the comparison results, which confirmed that more contradiction labels were assigned to the responses for C with a high FQness. This observation underscored the tendency of C with a higher FQness to provoke more RGM contradictions.

	$T=1$	$T=2$	$T=3$	$T=4$
RANDOM	194 / 700 (27.7%)	67 / 700 (9.6%)	33 / 700 (4.7%)	11 / 700 (1.6%)
TOP	**238** / 700 (**34.0%**)	**101** / 700 (**14.4%**)	**50** / 700 (**7.1%**)	**23** / 700 (**3.3%**)

(a) $d_{u_r} = 1$

	$T=1$	$T=2$	$T=3$	$T=4$
RANDOM	246 / 700 (35.1%)	77 / 700 (11.0%)	31 / 700 (4.4%)	10 / 700 (1.4%)
TOP	**270** / 700 (**38.6%**)	**141** / 700 (**20.1%**)	**81** / 700 (**11.6%**)	**43** / 700 (**6.1%**)

(b) $d_{u_r} = 3$

	$T=1$	$T=2$	$T=3$	$T=4$
RANDOM	98 / 350 (28.0%)	20 / 350 (5.7%)	6 / 350 (1.7%)	3 / 350 (0.9%)
TOP	**126** / 350 (**36.0%**)	**50** / 350 (**14.3%**)	**25** / 350 (**7.1%**)	**17** / 350 (**4.9%**)

(c) $d_{u_r} = 5$

Table 7: The number of contradictory responses to C extracted by RANDOM and TOP. Each value denotes the count of responses judged contradictory to u_r by at least T workers out of 10.

[15] Note that the ChatGPT API service was not yet available when the preliminary experiment was conducted.

[16] www.github.com/PaddlePaddle/Knover.

		P2	PX	B1	B2	B3	BL	O6	CG	Total
$d_{u_r} = 1$	# of contrad.	840	845	1263	1526	1472	908	1177	77	8108 (2703)
	# of noncontrad.	1759	1726	1172	967	1028	1628	1420	2771	12471 (2920)
$d_{u_r} = 3$	# of contrad.	208	301	362	395	361	230	287	31	2175 (739)
	# of noncontrad.	629	505	451	402	433	601	524	833	4378 (953)
$d_{u_r} = 5$	# of contrad.	26	30	42	35	27	22	33	5	220 (74)
	# of noncontrad.	56	47	42	44	46	56	50	81	422 (94)
Total	# of contrad.	1074	1176	1667	1956	1860	1160	1497	113	10503 (3516)
	# of noncontrad.	2444	2278	1665	1413	1507	2285	1994	3685	17271 (3967)

Table 8: The number of responses in our dataset. "# of ✗" and "# of ✓" denote the numbers of contradictory and noncontradictory responses, respectively. Values in parentheses refer to the number of types of contexts.

D Settings for dataset construction

Each of the eight RGMs generated one response to an input, resulting in eight responses for each C. We enhanced the efficiency of gathering the contradictions by choosing the final response of an RGM to an input from the top 100 candidates with the highest contradiction probability predicted by the state-of-the-art contradiction detector (Nie et al., 2020b). We utilized top-p sampling to collect the 100 candidates. We set a value of p to 0.5, which was lower than the default value of 0.9 used in major platforms, such as ParlAI (Miller et al., 2017), to avoid sampling responses with low generation probabilities. This allows us to gather candidates with a high generation probability by the RGM and a high likelihood of being contradictory. As in Appendix C, only OPT-66B generated responses using a 5-shot approach. We employed OpenAI's API[17] for ChatGPT. For the other RGMs, we used the same platforms as described in Appendix C.

E Worker selection for dataset construction

We first presented a task with obviously correct answers. It contained 21 dialogue responses requiring classification into contradictory or noncontradictory according to their preceding referent r. We exclusively handpicked workers who scored fewer than two incorrect answers in this task.

F Details of collected dataset

Table 8 illustrates the number of contradictory and noncontradictory responses obtained from each of the eight RGMs outlined in Section 3.

[17] https://platform.openai.com.

G Frequency of intra-utterance inconsistencies

Table 9 illustrates the frequency of intra-utterance inconsistencies in 50 randomly sampled contradictory responses of each of the eight RGMs in our dataset and 200 randomly extracted human-written contradictory responses from the DECODE dataset. The counting of intra-utterance inconsistencies was carried out by the author.

Our findings indicated that seven RGMs generated at least 4 (8%) contradictory responses featuring an intra-utterance inconsistency. The sole exception was ChatGPT (CG), whose subset of contradictory responses did not encompass intra-utterance inconsistency. This suggests that a small number of large-scale RGMs, such as ChatGPT, are progressing toward eradicating inconsistencies within individual utterances. Nevertheless, even a sophisticated model like Opt-66B generates contradictions with intra-utterance inconsistency.

Conversely, none of the 200 responses randomly sampled from the DECODE dataset exhibited an intra-utterance inconsistency ("Human" in Table 9). These results suggest contradictory responses featuring intra-utterance inconsistencies are particularly frequent in RGM responses.

H Settings of dialogue act labeling

For the analysis in Section 4.2, we developed a labeler to assign dialogue act labels to the utterances u_q in our dataset.

H.1 Development of dialogue act labeler

Inputs and outputs. Suppose that the t-th utterance u_t in a dialogue consists of the n segments $(u_{t,1}, u_{t,2}, ..., u_{t,n})$ and that each segment is assigned one dialogue act label. In addition, let the utterance immediately preceding u_t be u_{t-1}.

RGM	P2	PX	B1	B2	B3	BL	O6	CG	Human
Frequency	4 / 50 (8%)	5 / 50 (10%)	6 / 50 (12%)	8 / 50 (16%)	12 / 50 (24%)	8 / 50 (16%)	5 / 50 (10%)	0 / 50 (0%)	0 / 200 (0%)

Table 9: The frequency of intra-utterance inconsistencies in contradictory responses of RGMs and humans. The column labeled "Human" represents the frequency of intra-utterance inconsistencies in human-written contradictory responses extracted from the DECODE dataset.

	# of randomly sampled contradictory / noncontradictory instances for training				
Detector	SNLI	MultiNLI	DialogueNLI	AdversarialNLI-R3	DECODE
$\text{CD}_{\text{SNLI+MNLI}}$	4012 / 4012	4011 / 4011	0 / 0	0 / 0	0 / 0
CD_{ALL}	1004 / 1004	1003 / 1003	2006 / 2006	2005 / 2005	2005 / 2005
$\text{CD}_{\text{ALL}-\text{DNLI}}$	1338 / 1338	1337 / 1337	0 / 0	2674 / 2674	2674 / 2674
$\text{CD}_{\text{ALL}-\text{ANLI}}$	1338 / 1338	1337 / 1337	2674 / 2674	0 / 0	2674 / 2674
$\text{CD}_{\text{ALL}-\text{DEC}}$	1338 / 1338	1337 / 1337	2674 / 2674	2674 / 2674	0 / 0
CD_{DNLI}	0 / 0	0 / 0	8023 / 8023	0 / 0	0 / 0
CD_{ANLI}	0 / 0	0 / 0	0 / 0	8023 / 8023	0 / 0

Table 10: The number of contradictory and noncontradictory instances gathered for the training of each baseline detector. Following the experiment by Nie et al. (2020b), training instances were randomly sampled from five datasets: the SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), DialogueNLI (Welleck et al., 2019), AdversarialNLI-R3 (Nie et al., 2020a), and DECODE dataset. In gathering negative (contradictory) instances from Natural Language Inference (NLI) datasets, the premise text of an NLI instance labeled as contradiction was designated as u_r, and the hypothesis text of the same NLI instance was considered the corresponding contradictory response. Similarly, in the extraction of positive (noncontradictory) instances from NLI datasets, entailment- or neutral-labeled NLI instances were utilized. In this case, the premise text of the NLI instance was treated as u_r, and the hypothesis text was regarded as the noncontradictory response.

We design a labeler to predict the dialogue act label of the i-th segment $u_{t,i}$ within u_t, using u_{t-1} and the segments of u_t up to the i-th segment $(u_{t,1}, u_{t,2}, ..., u_{t,i})$ as inputs.

Dataset for training. We developed a labeler by fine-tuning RoBERTa.[18] We employed the Switchboard Dialogue Act Corpus with the 42 clustered SWBD-DAMSL dialogue act labels (Jurafsky et al., 1997) for fine-tuning. The preprocessed dataset we used for this study[19] comprises training data with 192,390 segments, validation data with 3,272 segments, and evaluation data with 4,078 segments.

Hyperparameters. We fine-tuned RoBERTa by employing the implementation of Hugging Face (Wolf et al., 2020) with its default settings, excluding a few parameters.[20]

Performance of developed labeler. The test set accuracy on the aforementioned preprocessed dataset, used for fine-tuning RoBERTa, was 80.4%.

[18] https://huggingface.co/FacebookAI/roberta-large.

[19] https://github.com/shreyangshu12/Dialogue-act-classification.

[20] early_stopping_patience: 2, learning_rate: 1e-5, train_batch_size: 256, and weight_decay: 0.01

H.2 Dialogue act labeling for our dataset

The developed labeler was utilized to assign dialogue act labels to the utterance u_q in our dataset. To simplify the process, we treated each sentence in u_q as a segment,[21] assigning a dialogue act label to each sentence. We used the utterance immediately preceding u_q and all sentences up to the i-th in u_q as input in order to predict the dialogue act label for the i-th sentence in u_q. This process was applied to all sentences within u_q. In our analysis in Section 4.2, for the sake of simplicity, we regarded that a certain label had been assigned to u_q if one or more sentences within u_q were assigned that specific label.

I Settings for test set construction

We first constructed two pools of C by extracting $d_{u_r} = 1, 3$ or more consecutive utterances from the Topical-Chat dataset and the DailyDialog dataset, respectively, ensuring that the final utterance contains questions. From each of the two pools, we randomly sampled those consecutive utterances. Specifically, for the Topical-Chat dataset, we sampled 300 and 100 samples from the pool for

[21] We split an utterance into sentences using NLTK sentence tokenizer (Bird and Loper, 2004).

Detector	By-Human
$\text{CD}_{\text{SNLI+MNLI}}$	.777
CD_{ALL}	.935
$\text{CD}_{\text{ALL−DNLI}}$	.930
$\text{CD}_{\text{ALL−ANLI}}$	.934
$\text{CD}_{\text{ALL−DEC}}$	.856
CD_{DNLI}	.771
CD_{ANLI}	.790
CD_{DEC}	.952
CD_{OUR}	.842
$\text{CD}_{\text{DECfull}}$	**.958**

(a) Human-written test set.

Detector	P2	PX	B1	B2	B3	BL	O6	CG
$\text{CD}_{\text{SNLI+MNLI}}$	.675	.600	.675	.615	.620	.630	.575	.650
CD_{ALL}	.705	.595	.615	.550	.675	.645	.625	.730
$\text{CD}_{\text{ALL−DNLI}}$	.690	.615	.650	.640	.635	.615	.635	.690
$\text{CD}_{\text{ALL−ANLI}}$	.705	.645	.620	.550	.680	.605	.640	.650
$\text{CD}_{\text{ALL−DEC}}$	.690	.625	.635	.630	.625	.600	.655	.720
CD_{DNLI}	.595	.580	.575	.570	.600	.540	.550	.565
CD_{ANLI}	.680	.595	.635	.605	.595	.600	.610	.615
CD_{DEC}	.600	.575	.615	.540	.655	.555	.565	.650
CD_{OUR}	**.800**	**.735**	**.715**	**.750**	**.765**	**.745**	**.690**	**.790**
$\text{CD}_{\text{DECfull}}$	.625	.620	.565	.585	.595	.625	.575	.715

(b) In-domain test sets. Scores for each target RGM are presented.

	Test set from Nie+'21	Test sets from Topical-Chat / DailyDialog						
Detector	Human-Bot	P2	PX	B1	B2	B3	BL	O6
$\text{CD}_{\text{SNLI+MNLI}}$	.729	.58/.66	.63/.60	.59/.54	.53/.64	.58/.64	.60/.58	.61/.67
CD_{ALL}	.783	.63/.66	.63/.64	.67/.54	.61/.53	.63/.62	.72/.56	.61/.59
$\text{CD}_{\text{ALL−DNLI}}$	.783	.65/.69	.65/.62	.69/.57	.61/.56	.71/.69	.66/.56	.55/.65
$\text{CD}_{\text{ALL−ANLI}}$	.775	.68/.66	.61/.64	.61/.61	.60/.63	.70/.66	.70/.56	.60/.57
$\text{CD}_{\text{ALL−DEC}}$	.771	.60/.67	.57/.63	.56/.55	.57/.65	.64/.66	.72/.62	.62/.62
CD_{DNLI}	.736	.60/.60	.62/.56	.57/.60	.60/.62	.63/.53	.62/.54	.71/.55
CD_{ANLI}	.688	.63/.64	.70/.69	.65/.57	.61/.69	.68/.62	.65/.61	.59/.66
CD_{DEC}	.749	.55/.52	.58/.60	.61/.55	.60/.59	.68/.61	.67/.55	.59/.53
CD_{OUR}	.787	**.77/.77**	**.75/.71**	**.74/.68**	**.70/.72**	**.73/.76**	**.82/.64**	**.81/.75**
$\text{CD}_{\text{DECfull}}$	**.829**	.57/.52	.64/.59	.67/.56	.64/.59	.69/.66	.71/.54	.59/.54

(c) Out-of-domain test sets. For the Topical-Chat and DailyDialog test sets, scores for each target RGM are presented.

Table 11: Accuracy of CD_{OUR} and all rival detectors for (a) human-written, (b) in-domain, and (c) out-of-domain test sets.

$d_{u_r} = 1$ and 3, respectively. Similarly, for the DailyDialog dataset, we sampled 200 and 100 samples from the pool for $d_{u_r} = 1$ and 3, respectively. The other settings were the same as in our large-scale dataset construction described in Section 3, except for the method of collecting C described above and that responses of ChatGPT were not collected.

J Settings of detector training

Samples for training. A negative pair comprised a contradictory response in our dataset and the preceding utterance u_r. In contrast, a positive pair comprised a noncontradictory response from our dataset and one randomly selected from its preceding utterances by the same speaker. This was because the responses annotated as noncontradictory with u_r are also likely to be noncontradictory with the other preceding statements. By introducing randomness into the selection of preceding utterances for pairing with a noncontradictory RGM response, we aimed to create positive pairs comprising unrelated utterances. These pairs could be valuable for training detectors to recognize that unrelated pairs should be categorized as noncontradictory.

Hyperparameters. We fine-tuned RoBERTa[22] by employing the implementation of Hugging Face (Wolf et al., 2020) with its default settings, excluding a few parameters.[23] We updated the model parameters until we reached a point where early stopping was triggered. Early stopping was determined by assessing the accuracy of validation data, a distinct subset comprising 10% of the training data and withheld from the training process. We saved the model parameters with the highest accuracy on the validation data at each learning rate and ultimately selected that with the highest validation accuracy among all the saved parameters.

K Comparison to diverse rival detectors

Table 11 presents the outcomes of evaluating the contradiction detection capabilities of CD_{OUR} in comparison to diverse rival detectors.

Comparison to baselines employed in Nie et al. (2020b). In addition to CD_{DEC}, we developed seven detectors named $\text{CD}_{\text{SNLI+MNLI}}$, CD_{ALL},

[22] https://huggingface.co/FacebookAI/roberta-large.

[23] `train_batch_size`: 128, `weight_decay`: 0.01, `eval_steps`: 200, `early_stopping_patience`: 1, and `learning_rate`: {1e-6, 5e-6, 1e-5, 5e-5}.

d_{u_r}	Contradictory	Noncontradictory
1	.762 (772 / 1013)	.956 (930 / 973)
3	.762 (425 / 558)	.968 (430 / 444)
5	.770 (191 / 248)	.987 (376 / 381)
7	.833 (160 / 192)	.973 (182 / 187)
9	.887 (63 / 71)	.976 (80 / 82)
11	.714 (10 / 14)	1.00 (33 / 33)
13	1.00 (8 / 8)	1.00 (5 / 5)
15	1.00 (1 / 1)	-

Table 12: $\mathrm{CD_{OUR}}$'s distance-wise accuracy on the By-Human test set. The values are listed separately for contradictory and noncontradictory responses.

$\mathrm{CD_{ALL-DNLI}}$, $\mathrm{CD_{ALL-ANLI}}$, $\mathrm{CD_{ALL-DEC}}$, $\mathrm{CD_{DNLI}}$, and $\mathrm{CD_{ANLI}}$, corresponding to the seven RoBERTa-based baseline detectors utilized in Nie et al. (2020b)'s experiments. The training settings for all detectors were consistent with those of $\mathrm{CD_{DEC}}$, with the only variation being the source of the training data (Table 10). We subsequently juxtaposed these detectors with $\mathrm{CD_{OUR}}$. Table 11 illustrates that $\mathrm{CD_{OUR}}$, formed using our dataset, exhibited a higher accuracy in identifying RGM contradictory responses compared to those rival detectors.

Comparison to a detector trained with more human-written data. Additionally, we developed the detector $\mathrm{CD_{DECfull}}$ through training on all instances within the DECODE dataset, encompassing 15,605 contradictory and 15,605 noncontradictory responses. This training process followed the methodology employed for $\mathrm{CD_{DEC}}$. Noteworthy is the observation that, despite $\mathrm{CD_{OUR}}$'s training dataset being approximately half the size of $\mathrm{CD_{DECfull}}$, it demonstrated superior performance across all RGM-generated test sets, with the exception of the Human-Bot set. The Human-Bot test data comprises the utterances in first-meeting dialogues. Given that $\mathrm{CD_{DECfull}}$'s training dataset also encompasses human-written contradictory responses following first-meeting dialogues, it is conceivable that the overlap in domains enabled $\mathrm{CD_{DECfull}}$ to recognize contradictions in the Human-Bot test data.

L Distance-wise accuracy of our detector

This study collected only contradictory responses with $d_{u_r} \leq 5$ because the annotation cost prohibitively increases as the value of d_{u_r} increases. To examine whether our dataset also contributes to suppressing contradictions with $d_{u_r} > 5$, we computed the distance-wise accuracy of some of the results reported in Section 5.

Settings. In our experiments in Section 5, we trained a contradiction detector ($\mathrm{CD_{OUR}}$) with our dataset and showed its performance on several test sets. Among those test sets, only the By-Human test set contained a large number of contradictory instances with $d_{u_r} > 5$. Therefore, we calculated $\mathrm{CD_{OUR}}$'s distance-wise accuracy on the By-Human test set.

Results. Table 12 shows the distance-wise decomposition of the performance of $\mathrm{CD_{OUR}}$ on By-Human, where we computed the percentage of the test instances correctly classified by $\mathrm{CD_{OUR}}$ for each d_{u_r}. The results demonstrate that the detector's performance did not degrade for instances with a longer distance ($d_{u_r} > 5$).