Tweak to Trust: Assessing the Reliability of Summarization Metrics in Contact Centers via Perturbed Summaries

Kevin Patel[*], **Suraj Agrawal**[*] and **Ayush Kumar**
{kevin.patel, suraj.agrawal, ayush}@observe.ai
Observe.AI
Bangalore, India

Abstract

In the dynamic realm of call center communications, the potential of abstractive summarization to transform information condensation is evident. However, evaluating the performance of abstractive summarization systems within contact center domain poses a significant challenge. Traditional evaluation metrics prove inadequate in capturing the multifaceted nature of call center conversations, characterized by diverse topics, emotional nuances, and dynamic contexts. This paper uses domain-specific perturbed summaries to scrutinize the robustness of summarization metrics in the call center domain. Through extensive experiments on call center data, we illustrate how perturbed summaries uncover limitations in existing metrics. We additionally utilize perturbation as data augmentation strategy to train domain-specific metrics. Our findings underscore the potential of perturbed summaries to complement current evaluation techniques, advancing reliable and adaptable summarization solutions in the call center domain.

1 Introduction

In the contemporary digital era, abstractive summarization (Mehdad et al., 2014) emerges as a crucial technology for condensing vast documents into concise, coherent summaries, thereby enhancing human readability. Unlike extractive summarization, which merely stitches together parts of the original text (Zhong et al., 2020; Mihalcea and Tarau, 2004), abstractive summarization paraphrases the content, producing summaries that are both informative and contextually rich. The advent of Large Language Models, including OpenAI's GPT series (Floridi and Chiriatti, 2020) and Meta's LLaMa (Touvron et al., 2023), has significantly propelled the field forward, offering unprecedented capabilities in synthesizing information from varied data formats such as documents, tables and texts (Goyal et al., 2023; Jin et al., 2024; Vassiliou et al., 2023).

As the field evolves, the need for robust and reliable evaluation methods for abstractive summarization systems becomes increasingly apparent. While traditional metrics like ROUGE (Lin, 2004a) have been widely used, their limitations lie in their inability to capture the diversity and creativity intrinsic to abstractive summarization. Recent research explores alternative evaluation approaches, such as learned neural metric models (Zhang et al., 2019a) and human evaluation studies (Wang et al., 2023; Luo et al., 2023), aiming for nuanced assessments in characteristics like fluency, coherence, and informativeness. However, the reliability of these evaluation metrics remains an active research question. Numerous works have studied the robustness and reliability of evaluation metrics (Freitag et al., 2022; Juraska et al., 2023). Liu et al. (2023) introduced a dataset and annotation methodology to enhance evaluation robustness, while researchers have also explored the use of ChatGPT as an evaluator (Luo et al., 2023; Wang et al., 2023). Moreover, recent work by Fu et al. (2023) and Koo et al. (2023) underscores the low reliability of LLM as an evaluator. Furthermore, studies by Ribeiro et al. (2020) and Sai et al. (2021) highlight how introducing perturbed outputs affects the correlation between metrics and human scores. Our study investigates the robustness of automatic summarization evaluation metrics via perturbations in the call center domain. The contributions of our work are as follows:

1. We establish that out-of-the-box evaluation metrics fail to align with human assessments of summary quality in contact center domain. Notably, despite the known fragility of evaluation metrics, to the best of our knowledge, our study is the first to apply this scrutiny to a real-world dataset from the contact center industry.
2. We propose creating domain-specific summary

[*] Equal Contribution

Proceedings of the 4th Workshop on Trustworthy NLP (TrustNLP 2024), pages 172–186

perturbations based on the error patterns observed in call summarization outputs. These perturbations aim to simulate real-life scenario and test the robustness of evaluation metrics under such conditions.

3. We demonstrate the potential of utilizing the perturbed summaries as data augmentation to train the domain-specific evaluation metrics.

2 Nuances of Call Center Domain

Call centers, crucial in various industries, facilitate interactions between agents and customers, covering inquiries, issue resolution, technical support, complaints, and product information. These dynamic conversations pose challenges for abstractive summarization systems. Challenges include:

Variety of Topics and Contexts: Call center conversations cover a wide range of topics, each characterized by its distinct context and structure. Traditional metrics overlook these variations, resulting in discrepancies between scores and actual informativeness. For instance, if a call concerns canceling a flight but the summary mentions canceling a hotel instead, the consistency metric should be markedly low, even if only a single word differs.

Variation in the language: Conversations often blend informal speech, colloquial expressions, and specialized terminology, posing a challenge for evaluation metrics, which need to handle such diversity effectively. For example, phrases like *'The customer called to get pre-authorization to send a patient to a facility.'* and *'During the call, the customer requested preauthorization to transfer a patient to a facility.'* should be assessed appropriately by these metrics. In the first scenario, the statement identifies the call's main purpose, while in the second, despite a similar meaning, it simply points to a specific event within the conversation.

Handling Emotional Content: Traditional evaluation metrics fail to differentiate between summaries that accurately reflect the emotional tone of a call transcript and those that do not, marking a significant shortfall in assessing emotional content. For example, consider the distinction in emotional tone between *'Student aced the exam.'* and *'Student performed decently on the exam.'* Despite their similarity in meaning, one may better align with the emotional context of the referenced conversation, highlighting the inadequacy of current metrics in capturing such nuances.

3 Perturbations

Perturbation Type	**Prompt**
Writing style conversion	*Rewrite the summary and change the style to one of {shorthand, passive voice, active voice}, keeping the meaning same*
Changing the Speaker	*Rewrite the summary, after randomly change the speaker 'customer' and 'agent' from the summary.*
Making demographic changes	*Rewrite summary after adding the demographic information wherever possible.*
Noise addition	*Rewrite the summary after adding some random noise sentences related to summary in the output*
Length Reduction	*Reduce the summary keeping the summary to be same.*
Length Increase	*Make the summary longer in length, keeping the information same*
Category Changes	*Rewrite the summary after changing the domain or category or vertical of the given summary.*

Table 1: Prompt that were used during perturbations generation defined in the Section 3. Process for entity based perturbation and sentence based perturbation is detailed in the section.

A perturbed datapoint is a deliberately modified original datapoint, incorporating slight changes or noise (Zhang et al., 2022). Depending on the nature of the changes introduced in a perturbation, the perturbed data can be of same quality as of original data (*score-preserving perturbation*), while in other cases perturbation degrades the quality (*score-degrading perturbation*). Utilizing perturbations allows for assessing the robustness of evaluation metrics. The evaluation metric should exhibit consistent values for score-preserving perturbations, contrasting with degraded quality scores for score-degrading perturbations. Additionally, the correlation between the metric score and human scores should ideally remain consistent even when the data is perturbed.

In our work, we generate domain-specific summary perturbations by harnessing the capabilities of Large Language Models (LLMs). These perturbations, inspired by observed patterns and errors in the outputs of summarization systems, are created either through direct prompts [1] or a systematic approach utilizing LLMs at different stages. Our primary objective is to examine the consistency and

[1]Prompts used to generate perturbations is mentioned in Table 1.

relevance of summaries by applying these specifically designed perturbations, which aim to mirror real-life scenarios and evaluate the resilience of evaluation metrics in such contexts. *Consistency* refers to the accuracy and faithfulness of the summary to the source material (call transcript). A consistent summary accurately reflects the facts, opinions, and overall message of the original text (call transcript) without introducing contradictions or misrepresentations. On the other hand, *Relevance* evaluates whether the summary captures all the critical and relevant information from the original text (call transcript), while avoiding generating information that is not needed. The perturbations are outlined below [2]:

1. Writing style conversion: This perturbation aims to rewrite the summary while preserving its meaning, enhancing the evaluation measure's robustness to differently written but semantically identical summaries.

2. Changing the Speaker: Addressing speaker switching in call center scenarios, this perturbation mitigates metric sensitivity to speaker name changes.

3. Making demographic changes: Introducing demographic changes involves adding errors and false information, such as inserting a dummy person's address (e.g., *'123 Main Street, Anytown, USA'*), to test the robustness of the metric.

4. Noise addition: Introducing random noise tests the metric's ability to penalize irrelevant information.

5. Length Modification: Generating shorter or longer summaries while maintaining meaning assesses metric stability to change in length .

6. Category Changes: Rewriting summaries with changes in domain or category[3] tests metric sensitivity to shifts in context.

7. Entity Based Perturbation[4]: Aim to evaluate the robustness of evaluation metrics in accurately identifying consistency errors and hallucinations manifested due to incorrect entity values in the summary. The method involves instructing the LLM to identify entities and replace them with suitable alternatives. This process generates various perturbations, denoted as `change_perturbation_`n, where the robustness of the evaluation metrics is tested.

8. Sentence Based Perturbation[5]: It tests how well evaluation metrics understand the importance of information that is either included or missing in summaries. The perturbation process comprises two stages: in Stage 1, LLMs are utilized to characterize the domain (e.g., Medical, Education, etc.) and generate corresponding categories; in Stage 2, LLMs determine the importance of sentences to the summary. Subsequently, subsets of uniquely important sentences are removed to create perturbations. If a removed subset contains n sentences, the resulting perturbation is labeled as `remove_important_sentence_`n.

All the prompts used to generate the perturbations are present in table 1.

3.1 Entity Based Perturbation Algorithm

The primary objective of Entity Based Perturbation is to assess the robustness of evaluation metrics to correctly detect consistency errors and hallucinations by systematically altering the summary. The method unfolds through the following steps:

1. **Entity Identification:** Utilize a Language Model (LLM) to identify entities within the input.

2. **Option Retrieval:** Employ the LLM to retrieve suitable replacement options for each identified entity.

3. **Index Powerset Creation:** Form a powerset using the set of indices corresponding to the identified entities.

4. **Perturbation Generation:** For each combination within the powerset, create a perturbation. Specifically, replace only the entity whose index is present in the combination with one of the available options. In cases where there are n elements in a particular combination slated for replacement, the resulting perturbation is denoted as `change_perturbation_`n.

[2] Examples can be located in Table 8

[3] Domain, category, or vertical denotes specific types of calls (e.g., outbound sales, support, etc.), as well as the sectors and industries associated with those calls.

[4] Please refer Section 3.1 for details

[5] Please refer Section 3.2 for details

3.2 Sentence Based Perturbation Algorithm

The Sentence Based Perturbation aims to assess the robustness of evaluation metrics in understanding the relevance by systematically excluding vital portions of a summary. The process involves two stages, where Stage 1 identifies key categories within a specific domain, and Stage 2 leverages this information to generate perturbations.

Stage 1:

1. **Domain Description:** Utilize an LLM to obtain a description d for the target domain.
2. **Category Identification:** Query an LLM with the domain description d to determine the categories $\{c_1, c_2, .., c_n\}$ a call center in this domain might encounter, along with corresponding descriptions $\{dc_1, dc_2, ..., dc_n\}$ for each category.

Stage 2:

1. **Call Classification:** Request the LLM to classify a call transcript into a specific domain d.
2. **Category Classification:** Based on the domain classification, instruct the LLM to classify the call into a maximum of two categories $c_x, c_y \in \{c_1, c_2, .., c_n\}$ determined in Stage 1.
3. **Sentence Categorization:** Ask the LLM to categorize each sentence in the summary into a maximum of two previously identified categories, $s_x, s_y \in \{c_1, c_2, .., c_n\}$.
4. **Perturbation Generation:**
   (a) If the sentence's category matches the call's category, consider the sentence unique to that call transcript
   (b) If the sentence's category belongs to the remaining categories, consider it common across the entire domain.

   If $s_x \in \{c_x, c_y\} \rightarrow$ *'important sentence'*, else *'non-important'*.
5. **Subset Removal:** Remove subsets of uniquely important sentences to generate perturbations. If a removed subset contains n sentences, label the resulting perturbation as `remove_important_sentence_`n.

Sentence perturbation serves as a tool for evaluating the model's ability to discern and preserve essential information in summaries.

Perturbation Type	consistency mean	relevance mean
add_negation	2.95	5.21
antonym_adjective	4.22	5.00
contractions	5.50	6.25
drop_adjectives	4.74	5.37
drop_phrases	4.30	4.80
drop_stopwords	3.35	4.20
expansions	4.00	5.50
hyponyms	3.50	5.25
jumble	2.40	2.90
remove_punct	4.85	5.50
repeat_sentences	4.50	5.35
replace_nouns_pronouns	4.32	1.58
sentence_reorder	4.10	5.30
subject_verb_dis	4.65	5.55
synonym_adjective	4.38	5.08
typos	4.80	5.50

Table 2: Average human scores for the "perturbed summaries" generated via the method outlined in Sai et al. (2021). These scores are rated on a scale of 7, as described in Section 5.1.

4 Methodology

We curate the dataset[6] with ground truth information for call summaries, assigning scores to measure consistency and relevance. This data is referred to as *'orig'* dataset. Our perturbation methodologies, as detailed in Section 3, are applied on *'orig'* dataset to get *'our'* perturbation dataset. Additionally, we also utilize perturbations defined by Sai et al. (2021) to obtain *'baseline'* perturbation dataset.

Manual annotations[7] of the perturbed data reveal substantial differences in consistency and relevance scores, as shown in the Tables 2 and 3. We calculate various metrics[8] on the original data (non-perturbed), baseline perturbation data, and our perturbation data. Subsequently, we integrate perturbed data into the training of custom metrics using various combinations and found to have a positive impact on correlation[9].

[6]The proprietary dataset used in this study. Please refer section 5.2 for further details.
[7]Refer to Section 5.1 for detailed annotation strategy
[8]Refer to Section 5.3
[9]Refer to Section 5.5

Metric	consistency mean	relevance mean
writing_style	5.36	5.84
speaker_switch	3.05	3.47
demographic_change	4.81	5.53
noise_addition	4.77	5.61
length_reduction	5.49	5.53
length_increase	5.60	5.91
category_change	4.42	4.95
change_perturbation_1	5.20	5.71
change_perturbation_2	5.31	5.94
change_perturbation_3	4.78	5.46
change_perturbation_4	5.02	5.71
change_perturbation_5	4.58	5.36
remove_important_sentence_1	5.95	5.67
remove_important_sentence_2	5.78	4.71
remove_important_sentence_3	5.71	4.53
remove_important_sentence_4	5.29	4.19
remove_important_sentence_5	5.13	4.21
remove_important_sentence_6	3.93	4.20

Table 3: Average human scores assigned to the "perturbed summaries" generated through the method outlined in Section 3. These scores are rated on a scale of 7, as described in Section 5.1. Note that in `change_perturbation_`n and `remove_important_sentence_`n, n represents the number of entity changes and the number of dropped sentences, respectively

5 Experiment Setup

5.1 Data Annotation / Scoring Mechanism

In conducting this study, we devise an annotation protocol to evaluate the quality of responses in terms of consistency and relevance. We draft comprehensive annotation guidelines, augmenting them with examples to elucidate the application of quality metrics, ensuring consistent interpretation and application of these criteria among annotators. Seven in-house annotators underwent a two-week training period tailored to familiarize them with the intricacies of interacting with large language models and evaluating response quality against call transcripts and instructions. This training utilize a distinct dataset from the evaluation corpus to avoid overlap and bias.

Throughout the annotation process, the origin of the outputs were anonymized to mitigate annotator bias towards any specific perturbation or model. Annotator agreement was continuously monitored and evaluated through a cross-annotator review mechanism, resulting in a Fleiss' Kappa score of 0.59, indicating moderate inter-annotator agreement and validating the reliability of the annotation process post-training. Following the training period, the evaluation corpus was distributed among the annotators, with data point shared with 3 of the annotators. The final assessment of response quality was based on the majority vote of labels provided by the annotators.

We employ a 7-point Likert scale with the following interpretation:

- 1 - Extremely bad
- 2 - Very bad
- 3 - Bad
- 4 - Acceptable
- 5 - Good
- 6 - Very good
- 7 - Extremely good

This scale strikes a reasonable balance between granularity and simplicity, making it practical for larger-scale evaluations where many summaries need to be assessed efficiently.

These annotators were also supervised to generate ground truth summaries for the dataset. After training, they were assigned exclusive data points for generating the best possible summaries (ground truth summaries), which were then quality-checked using a cross-annotator review mechanism.

5.2 Datasets

We utilize proprietary call center data to evaluate the methodology proposed in our work. This dataset comprises conversations between customers and agents across various domains such as medical, educational, banking, and service, among others. The calls are in US English language. Transcripts of these conversations are generated using an ASR engine, which has a Word Error Rate (WER) of 13.08. We obtain a total of 1200 calls from seven different types of accounts, covering domains like education, automobiles, banking, and service. The average call duration is 8 minutes 20 seconds, with calls ranging from 2 minutes to 28 minutes of duration. As defined in the section 5.1 Annotators are provided with these calls to generate ground truth summaries.

In addition to annotating ground truth summaries for these 1200 calls, we employ GPT-3.5-turbo and two internal language models (LLMs) to generate summaries for the calls. After generating the summaries, human annotators evaluate the summaries for the input calls, as described in section 5.1. This

process results in a dataset comprising 4800 pairs of input call transcript and corresponding summaries (3 model generated summaries and 1 ground truth summaries), along with their consistency and relevance score, referred to as the "orig" set.

Next, we randomly select 25 calls from the 1200 calls and apply our perturbation approach, as defined in section 3, along with the approach developed by (Sai et al., 2021). We use this dataset to get the human annotation for consistency and relevance for each pair of call transcript and perturbed summary using the mechanism defined in Section 5.1. The standard deviation of scores for consistency and relevance is 0.12 and 0.21. We extrapolate the average scores for consistency and relevance obtained from human annotation for each perturbation type and round it off to the nearest integer score and map it back to the class as per the 7-point Likert scale. These scores are then assigned to the remaining perturbed summaries across the remaining 1175 calls. Now this dataset contains input call transcript, perturbed summary, along with the consistency and relevance score. The resulting datasets generated using our approach of domain-specific perturbation will be denoted as 'our', while those generated using the approach by (Sai et al., 2021) will be labeled 'baseline'.

5.3 Metrics

We utilize various out-of-the-box metrics to conduct evaluations and benchmark the performance of a metric across the 'orig', 'our', and 'baseline' datasets. These metrics include BLEU (Papineni et al., 2002), ROUGE (Lin, 2004b), CHRF (Popović, 2015), TER (Snover et al., 2006), BERTScore (Zhang et al., 2019b), BLANC (Vasilyev et al., 2020), Shannon (Vasilyev et al., 2020), ESTIME (Vasilyev and Bohannon, 2021), UniEval (Zhong et al., 2022), and BART score (Yuan et al., 2021).

5.4 Training Setup

We explore two approaches for developing more robust metrics:

1) Classifier-based Custom Metrics: This method involves training classifiers to predict the correct consistency and relevance class based on out-of-the-box metric scores (as defined in section 5.3) used as features. Our dataset was split into training and test sets, with a 75% ratio for training and 25% for testing. We calculate the metrics defined in section 5.3 for the training set and train a range of classifiers using these metrics as feature vectors. We then evaluate the trained classifiers on the test set. We conduct experiments using both the 'orig' dataset and the 'orig' + 'our' dataset. The results are presented in Tables 5 and 6. Various classifier types were explored, including Decision Trees, SVMs, and Ordinal Linear Regression.

2) Fine-tune Existing Metrics: In this approach, we aim to fine-tune existing neural network-based metrics to observe changes in performance across different datasets. We utilize pretrained UniEval and BARTScore models and fine-tune them with 2 epochs of training. The same 75-25 train-test split is employed for evaluating these models. We use the hyperparameters as defined in the repositories `https://github.com/maszhongming/UniEval/tree/main` and `https://github.com/neulab/BARTScore`, throughout the process.

For experimentation, we utilize an AWS g4dn.2xlarge machine, which has 8 vCPUs, 32GB of RAM, and 16GB of GPU memory.

5.5 Evaluation

For measuring the effectiveness of a metrics, we use correlation with human annotation score. We compute Pearson, Spearman and Kendall Tau correlation co-efficients and take the average of it to report in this work. For measuring performance of classifier based learned metric (results presented in table 5 and 6), we measure accuracy (%of datapoints correctly classified) of the predicted quality of response against the human evaluation.

6 Results and Analysis

6.1 Perturbations to evaluate robustness

(a) Brittleness of Existing Auto Metrics: In Table 4, upon reviewing each metric, it becomes apparent that there is a decrease in correlation for 15 out of 24 metrics across both perturbed datasets concerning both relevance and consistency scores. The only exceptions are the UniEval and BART scores. Despite exhibiting positive correlation, they display intriguing characteristics. The UniEval consistency score demonstrates a high correlation with relevance on perturbed data (both 'our' and 'baseline'). Additionally, the UniEval relevance score shows a higher correlation with consistency on the 'orig' dataset. Moreover, the BART Score exhibits higher correlation when the 'transcript' is used as the ground truth reference, contrasting

		BLEU Inp	BLEU Ref	CHRF Inp	TER Inp	CHRF Ref	TER Ref
Consistency	orig	0.09	0.51	0.15	-0.17	0.52	-0.49
	our - orig	-0.19	-0.23	-0.23	0.23	-0.28	0.16
	baseline - orig	-0.11	-0.18	-0.15	0.09	-0.20	0.14
Relevance	orig	-0.10	0.28	-0.09	0.05	0.24	-0.33
	our - orig	-0.11	-0.39	-0.03	0.16	-0.44	0.32
	baseline - orig	-0.17	-0.29	-0.09	-0.13	-0.30	0.19

		ROUGE L f1 Inp	ROUGE LSum f1 Inp	ROUGE L f1 Ref	ROUGE LSum f1 Ref	BERT Score Inp	BERT Score Ref
Consistency	orig	0.17	0.17	0.52	0.52	0.13	0.52
	our - orig	-0.19	-0.19	-0.18	-0.18	0.14	-0.15
	baseline - orig	-0.07	-0.08	-0.12	-0.12	0.16	-0.07
Relevance	orig	-0.02	-0.02	0.34	0.34	0.28	0.37
	our - orig	-0.09	-0.19	-0.32	-0.33	0.23	-0.33
	baseline - orig	0.08	0.06	-0.22	-0.22	0.33	-0.20

		BLANC Help	Shannon	ESTIME Alarms	ESTIME Soft	ESTIME Coherence	UniEval Coherence
Consistency	orig	0.00	0.05	0.20	0.08	0.10	0.07
	our - orig	0.04	-0.16	-0.19	0.13	-0.15	0.06
	baseline - orig	-0.04	-0.09	-0.23	0.17	0.05	0.03
Relevance	orig	0.04	-0.11	0.01	0.20	-0.05	0.13
	our - orig	-0.08	0.02	-0.15	0.03	0.25	0.20
	baseline - orig	0.04	-0.09	-0.30	0.19	0.17	0.12

		UniEval Consistency	UniEval Fluency	UniEval Relevance	UniEval Overall	BART Score src ->hyp	BART Score hyp ->ref
Consistency	orig	0.01	0.00	0.34	0.09	0.11	0.53
	our - orig	0.19	0.02	-0.07	0.14	0.12	-0.27
	baseline - orig	0.08	0.07	-0.12	0.06	0.07	-0.20
Relevance	orig	0.20	0.03	0.26	0.24	0.23	0.26
	our - orig	0.47	0.30	-0.17	0.26	0.14	-0.28
	baseline - orig	0.26	0.41	-0.22	0.16	0.13	-0.32

Table 4: Correlation of evaluation metrics to consistency and relevance quality of the summaries in original ('orig') dataset along with the difference in correlation when evaluation metrics is applied to domain-specific perturbation ('our') data and 'baseline' perturbations.

		orig	our	baseline	our-orig	baseline-orig
Consistency	DecisionTreeClassifier	73.16%	61.14%	42.02%	-12.02%	-31.14%
	LogisticRegression	64.45%	60.47%	39.13%	-3.98%	-25.32%
	NearestNeighbor	72.55%	60.47%	30.43%	-12.08%	-42.12%
	OrdinalLinearRegression	50.73%	57.09%	30.43%	6.36%	-20.30%
	SVM	69.85%	57.43%	27.53%	-12.42%	-42.32%
		orig	our	baseline	our-orig	baseline-orig
Relevance	DecisionTreeClassifier	91.21%	77.36%	69.56%	-13.85%	-21.65%
	LogisticRegression	87.28%	72.30%	71.01%	-14.98%	-16.27%
	NearestNeighbor	78.18%	69.26%	68.11%	-8.92%	-10.07%
	OrdinalLinearRegression	71.01%	57.77%	52.12%	-13.24%	-18.89%
	SVM	74.81%	65.22%	59.82%	-9.59%	-14.99%

Table 5: Results of classifiers trained on 'orig' training split. Columns 'orig', 'our', and 'baseline' represent the datasets used for evaluation, while 'our-orig' and 'baseline-orig' show the difference in accuracy on these datasets.

with its performance degradation when the 'ground truth' reference is applied. These observations underscore the brittleness and inconsistency of these metrics for evaluating call center domain summarization. It's also noteworthy that the TER value shows an increase in correlation, which is undesirable given that TER is inversely related to consistency and relevance scores.

(b) Learning a custom classifier: We train custom classifiers to predict quality of summary among a label ranging between *{Extremely Bad, Extremely Good}*[10]. We use scores from out-of-box evaluation metrics as features for this training. Table 5 illustrates a significant drop in predicted quality of summaries on both 'our' and 'baseline' perturbed evaluation set. Specifically, 19 out of 20 classifier combinations exhibit a substantial decrease in ability of classifier trained on original data to predict the quality of the perturbed summary. These findings underscore the brittleness of metrics learned solely on 'orig' data, which stems from the brittleness of the underlying features.

[10]Possible Labels: *Extremely Bad, Very Bad, Bad, Acceptable, Good, Very Good, Extremely Good*

		orig	our	baseline	our-orig	baseline-orig
Consistency	DecisionTreeClassifier	70.93%	66.66%	25.25%	-4.27%	-45.68%
	LogisticRegression	63.71%	62.21%	57.94%	-1.50%	-5.77%
	NearestNeighbor	70.31%	63.28%	66.98%	-7.03%	-3.33%
	OrdinalLinearRegression	51.91%	51.12%	47.58%	-0.79%	-4.33%
	SVM	67.55%	62.53%	55.25%	-5.02%	-12.30%
		orig	our	baseline	our-orig	baseline-orig
Relevance	DecisionTreeClassifier	88.64%	85.31%	70.12%	-3.33%	-18.52%
	LogisticRegression	84.50%	69.42%	53.17%	-15.08%	-31.33%
	NearestNeighbor	76.52%	56.18%	69.55%	-20.34%	-6.97%
	OrdinalLinearRegression	57.06%	44.63%	53.17%	-12.43%	-3.89%
	SVM	73.04%	73.56%	65.28%	0.52%	-7.76%

Table 6: Results of classifiers trained on combination of 'orig' and 'our' datasets. Columns 'orig', 'our', and 'baseline' represent the datasets used for evaluation, while 'our-orig' and 'baseline-orig' show the difference in accuracy on these datasets. Compared to results in Table 5, augmenting with 'our' data in training the classifier minimizes the gap of predicted consistency and relevance scores on perturbed datasets ('our' and 'baseline') in 14 out of 20 comparisons ('our-orig', 'baseline-orig').

	UniEval		BARTScore	
	consistency	relevance	consistency	relevance
Out Of Box	0.2014	0.1892	0.2342	0.1993
Original	0.2682	0.2727	0.2100	0.1992
Original with Baseline Perturbation	0.2723	0.2588	0.2738	0.2556
Original with Our Perturbation	0.2736	0.2603	0.3171	0.2741

Table 7: Correlation of UniEval and BARTScore with consistency and relevance scores with different dataset used for fine-tuning the two evaluation metrics. Evaluation set is mix of original, baseline perturbed and our perturbed data.

6.2 Perturbations as Data Augmentation

We investigate if incorporating data with perturbations into the training of evaluation metrics can enhance the model's ability to grasp the subtle variations introduced by these perturbations. This approach aims to improve the robustness and sensitivity of the trained model to a wider range of data variations, leading to more accurate and reliable evaluation outcomes. The scores for the perturbed summaries were estimated via human annotation on a pool of 25 samples of each type of perturbation (Table 2, 3). We then assign the mean scores to the respective perturbation type on the larger pool of perturbed dataset that we have collected. Using

this dataset, we study two approaches for custom evaluation metric:

(a) Fine-tuning classifiers with scores on perturbed data: Table 6 presents the outcomes of the custom classifiers when incorporating 'our' perturbed data during training. It's evident from the table that the disparities have considerably diminished. Previously, the average reduction in consistency was 19.53%, which has now decreased to 9.07%. Similarly, the average reduction in relevance score has improved from 14.24% to 12.2%. These findings suggest that the integration of perturbed data has substantially enhanced the training of custom metrics, rendering them more resilient.

(b) Fine-tuning UniEval and BARTScore with perturbations: We fine-tune the UniEval and BARTScore models using various dataset combinations: 1) training solely on the 'orig' dataset, 2) augmenting the 'orig' data with 'baseline' perturbation data, and 3) augmenting the 'orig' data with 'our' perturbation data. Table 7 presents the results of these experiments, indicating that fine-tuning these models with perturbed data has resulted in enhanced correlation compared to the out-of-the-box performance. Notably, the improvement is particularly higher when integrating our perturbations compared to incorporating perturbations from (Sai et al., 2021). On utilizing a combination of our perturbed data, the correlation on consistency improves by 8.29% compared to out-of-box BARTScore metric. The improvement in correlation when utilizing baseline perturbation is 3.96%.

7 Conclusion

In this work, we investigate the reliability of summarization evaluation metrics by introducing contact center domain-specific perturbations. We find that existing evaluation metrics display brittleness when subjected to these perturbations. We find that off-the-shelf summarization metrics correlate less with human judgements on the perturbed summaries than the original summaries. Finally, we demonstrate that augmenting training data with these perturbations results in more robust metrics capable of accurately evaluating summaries.

8 Limitations

The study delves into domain-specific perturbations to assess the reliability of evaluation metrics in measuring the quality of generated summaries. While multiple perturbations are examined, it's conceivable that additional perturbations could further enhance the analysis. Moreover, the applicability of the same set of perturbations may vary across different use-cases and domains. Additionally, as perturbations are generated through prompting LLMs, future iterations of GPT models might produce perturbations of differing quality or encounter challenges in following the same prompts used in this study. Furthermore, although multiple evaluation metrics are considered in our assessment, contemporary approaches, including LLMs-as-a-judge, are increasingly employed for evaluation purposes. It would be valuable to explore how recent evaluation metrics and pipeline methodologies perform on perturbed datasets.

References

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan TN. 2023. Are large language models reliable judges? a study on the factuality evaluation capabilities of llms.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators.

Chin-Yew Lin. 2004a. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Yashar Mehdad, Giuseppe Carenini, and Raymond T. Ng. 2014. Abstractive summarization of spoken and written conversations based on phrasal queries. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1220–1230, Baltimore, Maryland. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

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, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Ananya B. Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M. Khapra. 2021. Perturbation checklists for evaluating nlg evaluation metrics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Oleg Vasilyev and John Bohannon. 2021. Estime: Estimation of summary-to-text inconsistency by mismatched embeddings. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 94–103.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20.

Giannis Vassiliou, Nikolaos Papadakis, and Haridimos Kondylakis. 2023. Summarygpt: Leveraging chatgpt for summarizing knowledge graphs. In *The Semantic Web: ESWC 2023 Satellite Events*, pages 164–168, Cham. Springer Nature Switzerland.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yunxiang Zhang, Liangming Pan, Samson Tan, and Min-Yen Kan. 2022. Interpreting the robustness of neural nlp models to textual perturbations.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038.

A Examples of domain-specific summary perturbations

Perturbation Type	Input Summary	Perturbed Summary	Comment
Writing Style Conversion	The customer called to get pre-authorization to send a patient to a facility. The agent asked for the patient's name, date of birth, Medicaid ID. The agent asks for the customer's name and confirms that she is a nurse. They also ask for a contact number and the customer provides it. The agent asks for the address the customer is coming from, and the customer provides it. The agent asked if the patient had any isolation or precautions, any special needs, and any attachments. The patient was on oxygen and needed a cardiac monitor.	Customer called for pre-authorization for patient facility transfer. Requested info: patient's name, DOB, Medicaid ID. Agent confirmed customer as nurse, obtained contact number and originating address. Inquired about isolation/precautions, special needs, attachments. Noted: patient on oxygen, requires cardiac monitor.	It is evident from the example that both the summaries are same, but written differently, having a syntax based score will not able to capture it.
Changing the Speaker	The customer called to get pre-authorization to send a patient to a facility. The agent asked for the patient's name, date of birth, Medicaid ID. The agent asks for the customer's name and confirms that she is a nurse. They also ask for a contact number and the customer provides it. The agent asks for the address the customer is coming from, and the customer provides it. The agent asked if the patient had any isolation or precautions, any special needs, and any attachments. The patient was on oxygen and needed a cardiac monitor.	The agent called to get pre-authorization to send a patient to a facility. The customer asked for the patient's name, date of birth, Medicaid ID. The customer asks for the agent's name and confirms that he is a nurse. They also ask for a contact number and the agent provides it. The customer asks for the address the agent is coming from, and the agent provides it. The customer asked if the patient had any isolation or precautions, any special needs, and any attachments. The patient was on oxygen and needed a cardiac monitor.	From the example, it is quite evident that speaker switching is making the whole meaning incorrect.
Making demographic changes	The customer called to get pre-authorization to send a patient to a facility. The agent asked for the patient's name, date of birth, Medicaid ID. The agent asks for the customer's name and confirms that she is a nurse. They also ask for a contact number and the customer provides it. The agent asks for the address the customer is coming from, and the customer provides it. The agent asked if the patient had any isolation or precautions, any special needs, and any attachments. The patient was on oxygen and needed a cardiac monitor.	The customer, a nurse, called to get pre-authorization to send a patient to a facility. The agent asked for the patient's name (John Smith), date of birth (March 15, 1980), and Medicaid ID (123456789). The agent also asked for the customer's name (Jane Doe) and confirmed that she is a nurse. They also asked for a contact number (555-123-4567) and the customer provided it. The agent asked for the address the customer is coming from (123 Main Street, Anytown, USA), and the customer provided it. The agent asked if the patient had any isolation or precautions, any special needs, and any attachments. The customer mentioned that the patient was on oxygen and needed a cardiac monitor.	It is evident in the example, that some false information has been added, which is not consistent. So the score should be reduced.

Noise addition	The customer is trying to provide information about a trip but is having trouble communicating the details. The agent is patient and tries to understand the customer's information. Eventually, they find the trip and the customer wants to cancel it due to a doctor's appointment change. The agent confirms the cancellation and asks if there is anything else the customer needs. The customer says no and they end the conversation with polite goodbyes.	The customer called to cancel their trip because they had a last-minute change in their doctor's appointment. The agent was able to quickly locate the trip number and successfully canceled it for the customer. The customer expressed their gratitude to the agent for their assistance. The call ended on a positive note. Additionally, the customer mentioned that they were looking forward to rescheduling their trip in the future.	It's evident in the generated perturbation, that "Additionally, the customer mentioned that they were looking forward to rescheduling their trip in the future." is a false information.
Length Redaction	The customer is trying to provide information about a trip but is having trouble communicating the details. The agent is patient and tries to understand the customer's information. Eventually, they find the trip and the customer wants to cancel it due to a doctor's appointment change. The agent confirms the cancellation and asks if there is anything else the customer needs. The customer says no and they end the conversation with polite goodbyes.	The customer wanted to cancel a trip because of a doctor's appointment change. The agent found the trip number and canceled it. The customer thanked the agent and the call ended.	It is visible that perturbed output is shorter in length and cover all the relevant information.
Length Increase	The customer told the agent that the person they spoke to cannot add it today. The agent thanked the customer and the call ended.	During the conversation, the customer informed the agent that the person they reached mentioned that she doesn't have the availability to add that particular request today. The agent expressed gratitude to the customer for checking and providing this information. Finally, the customer concluded the call.	As it is evident from the output that summary has been increased and meaning is semantically same.
Category Changes	The customer is trying to provide information about a trip but is having trouble communicating the details. The agent is patient and tries to understand the customer's information. Eventually, they find the trip and the customer wants to cancel it due to a doctor's appointment change. The agent confirms the cancellation and asks if there is anything else the customer needs. The customer says no and they end the conversation with polite goodbyes.	The customer was trying to cancel a hotel reservation due to a doctor's appointment change. The agent was able to find the reservation details and successfully cancel it for the customer. The customer expressed gratitude towards the agent for their assistance, and the call concluded.	In the example the actual summary talks about cancellation of the trip but the perturbed summary converted it to hotel reservation cancellation.
Entity Based Perturbation	The reason for the agent to call is to inform the customer that their life insurance policy payment has declined and to provide them with the phone number to call in order to keep the policy in place.	The reason for Sarah Johnson to call is to inform the customer that their whole life insurance payment has declined and to provide them with the 1-800-123-4567 to call in order to keep the policy in place.	Here the perturbation involves addition of agent name, phone number and type of life insurance policy, but that was not the part of summary.

Sentence Based Perturbation	The agent did not resolve the customer's issue during this conversation. The conversation was focused on providing information about solar panels and the benefits of going solar. The agent also requested the customer to send their utility bills for further analysis.	The conversation was focused on providing information about solar panels and the benefits of going solar.The agent also requested the customer to send their utility bills for further analysis.	Perturbation remove the most critical sentence 'The agent did not resolve the customer issue', which is a critical information for the summaries.

Table 8: Detailed examples of Our Perturbation

B Baseline Perturbation Example Appendix

Here we provide more examples of perturbations generated by baseline paper in the table 9

Perturbation Type	Input Summary	Perturbed Summary
Jumble	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	. The of food their home agent . The options customer The The to else mentions customer options everything until . situation chicken agent to customer . receive agent wanting out unemployment and and concern likes they and in good customer more . take freezer The money the having The . . running bring of base both a food and price 's customer time ordering about are the expresses mentions lot discussing The understands pork customer guarantees not a
subject_verb_dis	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	The customer and agent is discussing food options . The customer likes both the base and pork options . The customer mentions having a lot of food in their freezer and not wanting to bring more home . The agent guarantees a good price . The customer mentions ordering chicken and running out of time to take everything else . The customer expresses concern about money until they receive unemployment . The agent understands the customer 's situation .
typos	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	The customer and agent are discussing food options. The customer likes both the bsae and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.
remove_punct	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	The customer and agent are discussing food options The customer likes both the base and pork options The customer mentions having a lot of food in their freezer and not wanting to bring more home The agent guarantees a good price The customer mentions ordering chicken and running out of time to take everything else The customer expresses concern about money until they receive unemployment The agent understands the customer s situation
drop_stopwords	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	The customer agent discussing food options . The customer likes base pork options . The customer mentions having lot food freezer wanting bring home . The agent guarantees good price . The customer mentions ordering chicken running time . The customer expresses concern money receive unemployment . The agent understands customer situation .

add_negation	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	The customer and agent are not discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.
hyponyms	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	The buyer and carrier begin discussing meat options . The buyer likes both the plate and pork options . The buyer mentions having a lot of meat in their freezer and not wanting to carry more home . The carrier guarantees a good price . The buyer mentions ordering chicken and running out of time to prepare everything else . The buyer expresses concern about savings until they get unemployment . The carrier understands the buyer 's urgency .
drop_adjectives	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	The customer and agent are discussing food options . The customer likes both the base and pork options . The customer mentions having a lot of food in their freezer and not wanting to bring home . The agent guarantees a price . The customer mentions ordering chicken and running out of time to take everything else . The customer expresses concern about money until they receive unemployment . The agent understands the customer 's situation .
sentence_reorder	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent guarantees a good price. The customer and agent are discussing food options. The agent understands the customer's situation. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home.
repeat_sentences	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation. The customer and agent are discussing food options.

drop_phrases	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	The options . The customer mentions having a lot of food in their freezer and not wanting to bring more home . The agent guarantees a good price . The customer mentions ordering chicken and running out of time to take everything else . The customer expresses concern about money until they receive unemployment . The agent understands the customer 's situation .
replace_nouns_prouns	The customer and agent are discussing food options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The agent guarantees a good price. The customer mentions ordering chicken and running out of time to take everything else. The customer expresses concern about money until they receive unemployment. The agent understands the customer's situation.	It The customer and agent are discussing food options. The customer and agent are discussing food options. It The customer likes both the base and pork options. The customer likes both the base and pork options. The customer likes both the base and pork options. The customer likes both the base and pork options. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The customer mentions having a lot of food in their freezer and not wanting to bring more home. The customer mentions having a lot of food in their freezer and not wanting to bring more home. It The agent guarantees a good price. The agent guarantees a good price. The agent guarantees a good price. The agent guarantees a good price. The agent guarantees a good price. It The customer mentions ordering chicken and running out of time to take everything else. The customer mentions ordering chicken and running out of time to take everything else. The customer mentions ordering chicken and running out of time to take everything else. The customer mentions ordering chicken and running out of time to take everything else. The customer mentions ordering chicken and running out of time to take everything else. It The customer expresses concern about money until they receive unemployment. The customer expresses concern about money until they receive unemployment. The customer expresses concern about money until they receive unemployment. It The agent understands the customer's situation. The agent understands the customer's situation.

Table 9: Detailed examples of baseline Perturbation