Evaluating Robustness of Open Dialogue Summarization Models in the Presence of Naturally Occurring Variations

Anonymous ACL submission

Abstract

Dialogue summarization involves summarizing long conversations while preserving the most salient information. Real-life dialogues often involve naturally occurring variations (e.g., repetitions, hesitations), and in this study, we systematically investigate the impact of such variations on state-of-the-art open dialogue summarization models whose details are publicly known (e.g., architectures, weights, and training corpora). To simulate real-life variations, we introduce two types of perturbations: *utterance-level* perturbations that modify individual utterances with errors and language variations, and *dialogue-level* perturbations that add non-informative exchanges (e.g., repetitions, greetings). We perform our analysis along three dimensions of robustness: *consistency*, *saliency*, and *faithfulness*, which aim to capture different aspects of performance of a summarization model. We find that both fine-tuned and instruction-tuned models are affected by input variations, with the latter being more susceptible, particularly to dialogue-level perturbations. We also validate our findings via human evaluation. Finally, we investigate whether the robustness of fine-tuned models can be improved by training them with a fraction of perturbed data and find that this approach does not yield consistent performance gains, warranting further research. Overall, our work highlights robustness challenges in current open models and provides insights for future research.

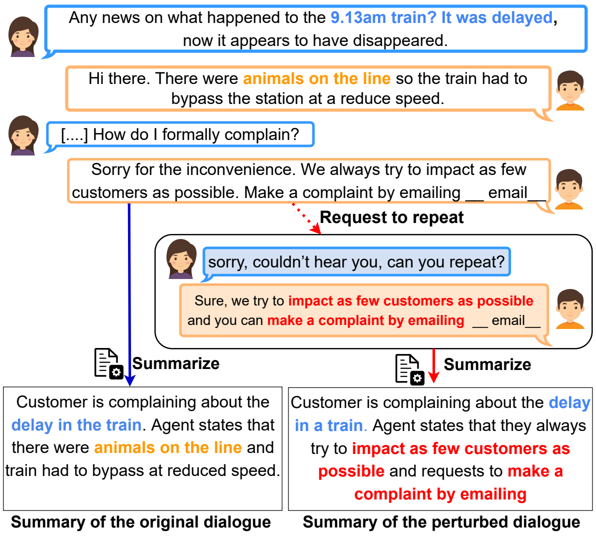


Figure 1: An example dialogue drawn from the TweetSum dataset, with a repeated utterance introduced as a perturbation. While the reference summary for the original dialogue includes the agent's explanation about the train delay, the summary of the perturbed dialogue includes information from the repeated utterance.

1 Introduction

Real-life conversations often exhibit a wide range of language variations, including typographical errors, grammatical mistakes, and certain exchanges such as repetitions and speaker interruptions, which are unrelated to the primary purpose of the conversation (Sacks et al., 1974). However, existing dialogue summarization datasets, which are used to train current summarization models, do not adequately capture these variations, as they are typically constructed by annotators simulating specific scenarios (Yuan and Yu, 2019) or extracted from English-speaking practice websites (Gliwa et al., 2019). Even the datasets consisting of real-life conversations (Feigenblat et al., 2021) can exhibit only a limited range of variations owing to practical limitations posed by the data collection process (e.g., high or low prevalence of conversations from different social demographics). Consequently, dialogue summarization models deployed in business scenarios encounter diverse variations not observed during training. This raises a crucial question: Can current dialogue summarization models effectively handle conversations with naturally occurring variations that are legitimate inputs but not observed in the training data?

In this work, we study the impact of naturally occurring variations on the performance of the state-of-the-art open dialogue summarization models (with publicly known architecture, weights, and training corpus) using three publicly avail-

Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024), pages 56–72
August 16, 2024 ©2024 Association for Computational Linguistics

able datasets. We examine the performance of encoder-decoder Transformer models in two setups a) fine-tuned on specific dialogue summarization datasets (Lewis et al., 2020; Zhang et al., 2019; Raffel et al., 2020b), and b) instruction-tuned models which have shown impressive zero-shot performance more recently (Gupta et al., 2022; Chung et al., 2022). Such models are often preferred in high-stakes business settings (e.g., medical, legal, and customer support) over proprietary models (e.g., ChatGPT), owing to user privacy concerns.

To simulate variations we design two kinds of perturbations: (a) utterance-level perturbations, and (b) dialogue-level perturbations (defined in Section 3), which are inspired by common real-life interaction patterns from the Natural Conversation Framework (Moore and Arar, 2019). We evaluate the performance of summarization models along three conceptually distinct robustness dimensions—*consistency*, *saliency*, and *faithfulness*—and elaborate on their empirical relationship.

Our analysis reveals that both fine-tuned and instruction-tuned models are impacted by utterance and dialogue-level perturbations. Instruction-tuned models are impacted more than fine-tuned models and are also more susceptible to dialogue-level perturbations than utterance-level perturbations. Both types of models show a preference for information from repeated, long, and leading utterances in the dialogue. Figure 1 shows an example where the model includes repeated utterances in the summary, whereas the non-repeated original utterance wasn't included in the summary before perturbation. We also validate our findings via human evaluation.

Finally, we investigate whether fine-tuned models improve by training with perturbed data. We find that this approach does not consistently enhance performance, and different perturbations require varying amounts of training examples for gains. Thus, further research is needed to address these robustness challenges.

2 Related Work

Prior work has investigated the robustness of language understanding models mainly focusing on classification tasks (Moradi and Samwald, 2021). Some dialogue-related classification tasks have also been explored, including dialogue act prediction (Liu et al., 2021), intent detection and slot tagging (Einolghozati et al., 2019; Sengupta et al., 2021), state tracking and dialogue modeling (Cho et al., 2022; Tian et al., 2021; Zhu et al., 2020; Kim et al., 2021; Peng et al., 2020).

Some studies have also investigated the robustness of neural language generation models, including neural machine translation (Niu et al., 2020; Karpukhin et al., 2019; Vaibhav et al., 2019), question answering (Peskov et al., 2019), and open domain multi-document summarization (Giorgi et al., 2022). However, some of these studies consider perturbations that are of extreme nature (e.g., random shuffling and deletion of words) and may occur rarely in the real world. Ganhotra et al. (2020) investigated the impact of natural variations on response prediction tasks in goal-oriented dialogues.

For summarization task in particular, previous studies focused on summarizing news articles and documents (Jing et al., 2003; Meechan-Maddon, 2019; Krishna et al., 2022). However, the nature of noise in a multi-party dialogue differs significantly from noise in documents. While some types of noise (e.g., spelling mistakes, grammatical errors) could occur in both, the patterns such as repetitions, reconfirmations, hesitations, and speaker interruptions (Sacks et al., 1974; Feng et al., 2021; Chen and Yang, 2021) are peculiar to dialogues, posing unique challenges for accurate and robust summarization. The focus of this work is to assess the robustness of *dialogue summarization models* in the presence of *naturally occurring variations*, which has been understudied in the prior literature.

3 Simulating Naturally Occurring Variations

To introduce naturally occurring variations in conversations, we consider two kinds of simulated perturbations, utterance-level and dialogue-level. We apply each perturbation individually to a dialogue to study its impact systematically. Our perturbations are inspired by the Natural Conversation Framework (Moore and Arar, 2019), created after analyzing real-world conversations across various use cases and provides common interactive patterns that occur in real life.[1] Appendix A.1 lists examples for each perturbation.

3.1 Utterance-level Perturbations

The utterance-level perturbations modify a single utterance and are adapted from (Liu et al., 2021). We perturb each utterance of the dialogue. For perturbations where multiple words in an utterance can be perturbed (e.g., spelling mistake, character casing), we consider only low-modification levels (i.e., perturb a word with 0.2 probability), which

[1] Some examples include patterns such as C1.0 (opening greeting agent), C4.6 (closing success check), B2.1.0 (repeat request), A2.8 (hold request).

also cause a considerable change in model performance.[2]

Typographical Errors Typographical errors occur when participants try to type quickly in chat-based interactions. We use simple regex-based perturbations, e.g., punctuation marks removal, whitespace removal or addition, changing letter casing, and substitutions of common expansions and contractions. We introduce spelling errors following the approach of Yorke as used in (Mille et al., 2021), replacing random letters with other letters closely co-located on the keyboard positions. We ensure that mistakes are not introduced in a proper-noun phrase (e.g., restaurant name) to avoid changes in important information.

Grammatical Errors We focus on two frequent grammatical errors: dropping determiners and subject-verb disagreements. To drop determiners, we drop all the words in a sentence with the DET tag. To introduce subject-verb disagreement, we identify auxiliary verbs (via AUX tag) and convert between plural and singular forms as appropriate, keeping the tense unchanged.

Language-use Variations Users can vary in their choices of dialect and vocabulary. We consider three language-use perturbations: substituting adjectives with synonyms, inflectional variations, and synthetic African American Vernacular English (AAVE) dialect. For synonym substitution, we substitute adjectives in an utterance with their WordNet (Miller, 1998) synonyms. To introduce inflectional variations, we follow the approach proposed in Dhole et al. (2021), where we lemmatize each content word in an utterance, randomly sample a valid POS category, and re-inflect the word according to the chosen category. To transform an utterance to synthetic AAVE dialect, we use the set of lexical and morphosyntactic transformation rules proposed by Ziems et al. (2022).

3.2 Dialogue-level Perturbations

We introduce new utterances that contribute no additional information, to test a model's ability to focus on the overall meaning of a conversation and identify salient information.

Repetitions Repeating and rephrasing occur commonly in real-life spoken conversations. In this perturbation, we randomly select an utterance to repeat.[3] We then inject a synthetic utterance requesting the other participant to repeat the information (e.g., 'Sorry, I couldn't hear you, can you repeat?').[4] Since humans tend to rephrase the original message slightly instead of repeating it verbatim, we paraphrase the original utterance before including it as a response to the request for repetition. We use Qian et al. (2019)'s paraphraser for this task. The rest of the dialogue remains unchanged. This perturbation allows us to examine repetition bias; i.e., does the model consider repeated utterances more significant, even when they do not contain important information?

Time delays A participant may ask the other party to wait while they gather information. To simulate this, we add three synthetic utterances consecutively: a request to wait (e.g., 'Just give me a few minutes.'), an acknowledgment from the other participant (e.g., 'Sure'), and an expression of gratitude from the first participant (e.g., 'Thanks for waiting.'). These utterances are inserted after a randomly selected utterance from the participant being asked to wait.

Greeting and closing remarks It is also common to begin a conversation with a friendly greeting and end with some closing remarks. For the greetings perturbation, we insert a greeting as the first utterance, such as 'Hi! I am your customer support assistant. How may I help you today?' in customer support dialogues and 'Hey there!' in open-domain chit-chat. For the closing remarks perturbation, we insert a final message: 'Thank you for contacting us.' in customer support dialogues and 'Cool, talk to you later!' in open domain chit-chat. Each perturbation is applied individually to a dialogue. Both of these perturbations help us investigate structural biases present in dialogue summarization models, also known to impact news summarization models (Xing et al., 2021; Jung et al., 2019). For instance, the greeting perturbation helps examine lead bias (preference for the first utterance), and closing remarks perturbation helps examine recency bias (preference for the last utterance).

[2]See Appendix A.5 for analysis with different perturbation rates.

[3]See Appendix A.4 for targeted perturbations, where we select an utterance to repeat based on its saliency.

[4]We use this utterance to operationalize the repetition perturbation, inspired by spoken dialogues. However, repetitions can also appear in written dialogues (e.g., sending the same message multiple times to ensure communication, emphasizing points, or dealing with technical issues.). Furthermore, models trained on written dialogues are often deployed to summarize transcripts of spoken dialogues, where such utterances are more common.

Split and combined utterances In chat-based conversations, participants can have varying preferences for either conveying information over multiple consecutive utterances or sending one long message. To simulate split utterance perturbation, we divide a randomly sampled utterance into consecutive utterances by splitting it at every five words. Conversely, to simulate combined utterance perturbation, we identify sequences of consecutive utterances from a single participant in a dialogue and concatenate them. We combine consecutive utterances from only one participant at a time. Each perturbation is applied individually to a dialogue. Both these perturbations allow us to examine long bias (the model's preference to include a long utterance over shorter utterances, even when multiple short utterances include salient information).

3.3 Quality evaluation of perturbed dialogues

We conduct a human validation of the perturbed dialogues. The goal of this evaluation is to ensure that our perturbations do not alter the dialogue's meaning or introduce new information, thereby validating the quality of our perturbed test set. We sample 20 dialogues and their summaries from each of the three datasets (§5.1) and perturb each dialogue with all of the utterance and dialogue-level perturbations, resulting in a total of `480` dialogues. Two annotators are asked to determine whether the reference summary for the original dialogue remains valid for all the perturbed dialogues (see Appendix A.2 for details on annotation guidelines). In cases of disagreement, a third annotator breaks the tie. The annotators marked 97.5% of the perturbed dialogues as being reasonably summarized by the summary of the original dialogue, thus validating the use of proposed perturbations to investigate the robustness of dialogue summarization models. Our human evaluation also suggests that our perturbations do not drastically alter the dialogue and the dialogues remain readable and semantically consistent. Otherwise, for an altered dialogue, the original summary would have been marked invalid.

4 Quantifying Robustness

For tasks involving text generation, such as dialogue summarization, measuring robustness involves determining the relationship between different pairs of natural language texts. As a result, the robustness of generative tasks is less well-defined, compared to a classification task (Liu et al., 2021) and can manifest in several ways. We consider three dimensions for measuring robustness issues that can arise in dialogue summarization.

Let x denote the original dialogue, y_r be the reference summary of the original dialogue, f be the summarization model trained on $(x, y_r) \sim D$, and $f(x)$ be its prediction over x. Let $x' = x + \delta$ denote the perturbed dialogue and $f(x')$ be its predicted summary.

`Consistency` A model is consistent (and hence robust) under a perturbation (δ) if the two summaries, $f(x)$ and $f(x' = x + \delta)$, are *semantically similar*, resulting in minimal change. We quantify the change in model-generated output as follows,

$$\Delta z_c = \frac{|\texttt{SCORE}(f(x), f(x)) - \texttt{SCORE}(f(x), f(x'))|}{\texttt{SCORE}(f(x), f(x))} \quad (1)$$

further simplified as,

$$\Delta z_c = 1 - \texttt{SCORE}(f(x), f(x')) \quad (2)$$

where `SCORE` is any text similarity metric (e.g., BERTScore) that assigns a value of 1 for identical inputs and 0 for dissimilar inputs. By definition, $\Delta z_c \in [0, 1]$. Note that consistency is sufficient but not necessary for robustness: a good summary can be expressed in diverse ways, which leads to high robustness but low consistency.

`Saliency` Assuming that the reference summary includes the most salient information conveyed in the input dialogue, we compute the change in salient information captured by the model-generated summaries (before and after perturbation) w.r.t the reference summary as follows:

$$\Delta z_s = \frac{|\texttt{SCORE}(y_r, f(x)) - \texttt{SCORE}(y_r, f(x'))|}{\texttt{SCORE}(y_r, f(x))} \quad (3)$$

where `SCORE` is any text similarity metric (e.g., BERTScore). Since Δz_s measures the normalized change in similarity scores, $\Delta z_s \in [0, 1]$.

`Faithfulness` Faithfulness refers to the extent to which the generated summary is supported by the content of the input dialogue, thus accurately reflecting the information without introducing spurious or fabricated details, commonly termed as hallucinations. We compute the change in faithfulness as follows:

$$\Delta z_f = \frac{|\texttt{SCORE}(x, f(x)) - \texttt{SCORE}(x, f(x'))|}{\texttt{SCORE}(x, f(x))} \quad (4)$$

where `SCORE` is any text-based precision metric measuring the fraction of information in the summary ($f(x)$) supported by the input dialogue

(x) (e.g., BERTScore-Precision). Since Δz_f measures the normalized change in precision scores, $\Delta z_f \in [0, 1]$. Note that, the second term in the numerator compares x with $f(x')$ since we are interested in measuring the fraction of summary information supported by the 'original dialogue.' Furthermore, since our added perturbations do not add any new information to the dialogue, x and x' would essentially contain the same information. Clearly, for all three dimensions, the higher the Δz, the lower the robustness of the model.

5 Evaluating Robustness

We present our key observations on how various perturbations impact the model performance.

5.1 Implementation Details

Datasets We consider two task-oriented dialogues, TWEETSUMM (Feigenblat et al., 2021) and TODSum (Zhao et al., 2021), both consisting of conversations between an agent and a customer. TODSum comprises dialogues from multiple subdomains (restaurants, movies, etc), collected via crowdsourcing where annotators are tasked to generate dialogues based on a given scenario. In contrast, TWEETSUMM focuses solely on customer support conversations occurred at Twitter. We also include SAMSUM (Gliwa et al., 2019), a corpus of chit-chat dialogues between two or more friends.

Models We analyze the robustness of three Transformer based encoder-decoder models for dialogue summarization, `Pegasus-large` (568M parameters) (Zhang et al., 2019), `BART-large` (400M parameters) (Lewis et al., 2020) and `T5-base` (220M parameters) (Raffel et al., 2020a), whose details are publicly available. All models have a comparable number of parameters. We fine-tune each model on the train split of the respective dataset. We use beam search[5] with size 5 to generate summaries. We also investigate the robustness of instruction-tuned versions of two of these models, `DIAL-BART0` (406M parameters) (Gupta et al., 2022) and `FLAN-T5-large` (783M parameters) (Chung et al., 2022), used as zero-shot summarizers, without fine-tuning on the three dialogue summarization datasets considered in this work.

Metrics We evaluate summaries using `BERTScore` (Zhang et al., 2020), which has been shown to better correlate with human judgment (Fischer et al., 2022). `BERTScore` calculates precision, recall, and F1 scores by comparing a model-generated summary to a reference summary. We use `F1` to compute *consistency* and *saliency*, and `precision` to compute *faithfulness*. To validate observed trends, we additionally evaluate summaries using ROUGE-L metric (Lin, 2004), which measures lexical overlap, and SummaC metric (Laban et al., 2022), which measures factual consistency. For all the reported results, we observe similar trends via ROUGE-L and SummaC (Tables 11,12,13 in Appendix A.8). While we report results using these metrics, the three robustness dimensions can be computed using any evaluation metric. For each reported result, we use a non-parametric bootstrap (Wasserman, 2004, ch. 8) to infer confidence intervals (CIs). We utilize 10^4 bootstrap samples of the dialogues to report 95% bootstrap CIs via the normal interval method (Wasserman, 2004, ch. 8.3).

5.2 How robust are fine-tuned models?

Fine-tuned dialogue summarization models are affected by both utterance and dialogue level perturbations Table 1 shows the change in *consistency*, *saliency*, and *faithfulness* owing to utterance and dialogue level perturbations on all three datasets. All three models are equally impacted by various perturbations. Models trained on `TweetSum` and `SAMSum` are impacted equally by both utterance-level and dialogue-level perturbations. `TODSum` is the least impacted, since this dataset contains template-based summaries where only entities from the dialogue are required to be filled. We see a major impact on faithfulness, with the highest impact on the model trained on the `TODSum` dataset.

Impact of utterance perturbations Table 2 shows that these perturbations have a comparable impact (shown averaged over all three models). Models trained on `TODSum` exhibit little change in consistency and saliency, but a significant change in faithfulness. This is expected since the `TODSum` summaries are extractive, following a pre-defined template, and only require substituting entity information extracted from the dialogue. Since the template is fixed and the summaries can only change in entity information before and after perturbation and w.r.t reference summary, we see a small change in consistency and saliency. However, we observe a large change in faithfulness, as this dimension focuses on the factual correctness of the summary.

Impact of dialogue perturbations: Table 3 reports the impact of dialogue-level perturbations (averaged over all models) and shows significant changes for repetition, time delays, greetings, and

[5] Nucleus sampling omitted to avoid sampling variance.

Dataset	Model	Utterance Perturbations			Dialogue Perturbations		
		$\Delta z_c\%$	$\Delta z_s\%$	$\Delta z_f\%$	$\Delta z_c\%$	$\Delta z_s\%$	$\Delta z_f\%$
TweetSum	BART	17.48±0.32	13.37±0.68	24.68±1.98	16.77±0.40	10.25±2.04	14.48±1.98
	Pegasus	16.73±0.42	17.18±1.04	29.51±5.20	16.67±0.42	11.33±1.97	21.03±5.20
	T5	17.89±0.37	14.44±0.82	16.67±2.94	17.02±0.38	11.78±1.35	9.81±2.94
TODSum	BART	7.26±0.24	3.87±0.16	51.71±17.09	5.85±0.24	2.70±0.42	19.07±15.06
	Pegasus	5.20±0.21	3.50±0.17	37.85±10.74	3.26±0.17	1.74±0.32	22.92±19.33
	T5	7.19±0.26	3.86±0.17	35.25±11.46	5.12±0.23	2.11±0.34	28.13±29.91
SAMSum	BART	13.06±0.36	6.57±0.25	11.39±0.73	22.05±0.52	5.11±0.65	6.62±1.28
	Pegasus	14.21±0.39	6.59±0.26	8.21±2.05	20.59±0.54	4.35±0.5	6.74±5.52
	T5	13.58±0.36	6.72±0.28	4.08±2.77	21.18±0.49	4.5±0.48	4.78±2.22

Table 1: Robustness scores of fine-tuned models using BERTScore. Higher the score, the lower the robustness.

Dimension	Dataset	Typographical	Grammar	Language Use
$\Delta z_c\%$	TweetSum	24.65±0.54	23.32±0.87	20.43±0.69
	TODSum	9.97±0.30	5.82±0.38	5.73±0.28
	SAMSum	16.27±0.36	16.93±0.71	17.78±0.48
$\Delta z_s\%$	TweetSum	16.27±1.93	16.93±2.7	17.78±1.96
	TODSum	5.59±1.32	3.12±1.04	2.96±0.89
	SAMSum	7.38±2.23	7.44±1.54	7.38±1.13
$\Delta z_f\%$	TweetSum	28.01±6.43	26.13±9.42	19.55±8.14
	TODSum	36.73±6.76	25.30±9.81	30.31±8.82
	SAMSum	11.17±1.75	9.98±1.83	8.97±1.57

Table 2: Impact of utterance perturbations. Models are equally impacted by different perturbations.

split utterances. For instance, when subjected to repetitions, the models tend to include repeated utterances in the summary, even if they were previously deemed unimportant (repetition bias; Figure 1). Additionally, the models demonstrate a preference for the first utterance in a dialogue (lead bias), rendering them susceptible to greetings perturbation. This observation aligns with prior findings for news summarization, where sentences at the beginning of an article are more likely to contain summary-worthy information. Similarly, in customer-support conversations, the first utterance frequently addresses the primary issue faced by the customer. Consequently, models trained on such datasets exhibit lead bias. Finally, the models prefer lengthy utterances in the summary (long bias), by being more affected by split perturbations, and less affected by short utterances combined.

5.3 Effect of model size on robustness

Table 4 shows the change in consistency for models with different number of parameters: `BART-base`, `BART-large`, `T5-base`, and `T5-small`. The models are almost equally affected by perturbations, irrespective of size, suggesting that robustness issues cannot be mitigated by scaling the model size.

5.4 How robust are instruction-tuned models when used as zero-shot summarizers?

`DIAL-BART0` and `FLAN-T5-large` are instruction-tuned on multiple tasks, with `DIAL-BART0`, in particular, is instruction-tuned on dialog-specific tasks. However, neither model was trained on the `TweetSum` dataset, providing a zero-shot setting to evaluate their dialogue summarization capabilities. As depicted in Table 5, both `DIAL-BART0` (Δz_c=30.37% for utterance and 34.30% for dialogue) and `FLAN-T5` (Δz_c=38.23% for utterance and 44.12% for dialogue) are much more sensitive to perturbations compared to their fine-tuned counterparts (Δz_c=17.36% for utterance and 16.82% for dialogue, averaged over three models).

In contrast to fine-tuned models, the zero-shot models are affected more by the dialogue-level perturbations (Δz_c=34.30% for `DIAL-BART0` and Δz_c=44.12% for `FLAN-T5`) than utterance-level perturbations (Δz_c=30.37% for `DIAL-BART0` and Δz_c=38.23% for `FLAN-T5`). Among utterance-level perturbations, similar to the fine-tuned models, zero-shot models are also impacted equally by all perturbations. Among dialogue-level perturbations as well, similar to the fine-tuned models, zero-shot models are most impacted by repetitions, greetings and split utterances (Appendix A.6).

We additionally consider a recent instruction-tuned large language model, `Llama-2-70B`, with only publicly available weights. This model is also significantly larger (70B) than the other models (<0.9B). Our results show high sensitivity to perturbations for this model (Δz_c=47.10% for utterance and Δz_c=54.53% for dialogue perturbations), though we leave detailed human evaluation of the outputs of this model for future work.

5.5 Validity of findings with human evaluation

We conduct another human evaluation to confirm the trends observed with automatic similarity metrics. Specifically, we collect similarity scores between summary pairs using human annotations instead of automated similarity metrics (e.g., BERTScore). The goal is to ensure that robustness trends observed with automated metrics are similar to those from human evaluation.

We use the consistency dimension for this evaluation for two main reasons: 1) Empirically, the three robustness dimensions exhibit a strong correlation (Table 10). Thus, using any of the three

Dimension	Dataset	Repetitions	Time Delays	Greetings	Closing Remarks	Split	Combine
	TweetSum	18.04±0.59	14.15±0.85	20.01 ±1.34	9.80±1.0	16.71±0.83	6.77±0.36
$\Delta z_c\%$	TODSum	5.96±0.39	4.31±0.4	6.61±0.59	2.02±0.4	4.38±0.36	-
	SAMSum	27.32±0.46	22.19±0.67	32.89±0.99	16.29±0.89	11.63±0.59	7.80±0.52
	TweetSum	12.49±3.45	10.53±1.47	15.23±5.98	6.03±2.23	11.13±1.45	5.40±1.34
$\Delta z_s\%$	TODSum	3.31±0.98	2.20±0.67	3.48±0.88	1.10±0.66	2.19±1.11	-
	SAMSum	10.87±0.23	8.38±0.98	12.63±0.95	6.04±1.14	14.65± 0.96	7.05±1.26
	TweetSum	19.34±5.91	15.81±1.2	18.31±9.23	6.99±8.28	15.11±7.47	8.65±1.42
$\Delta z_f\%$	TODSum	64.74±6.67	22.74±1.66	50.98±9.51	10.52±9.89	23.37±8.23	-
	SAMSum	17.99±8.91	12.76±2.44	21.25±0.91	10.28±0.95	16.05±5.91	10.21±1.91

Table 3: Robustness to dialogue perturbations. Models are most susceptible to repetitions and time delays (repetition bias), greetings (lead bias), and split utterances (long bias). `TODSum` dataset has no consecutive utterances from the same speaker, thus we do not perform combine utterance perturbation on this dataset.

Model	Parameters	Utterance Perturbations			Dialogue Perturbations		
		$\Delta z_c\%$	$\Delta z_s\%$	$\Delta z_f\%$	$\Delta z_c\%$	$\Delta z_s\%$	$\Delta z_f\%$
BART-large	440	17.48 ±0.33	13.37±0.68	24.68±0.85	16.77±0.40	10.25±2.01	14.48±1.98
BART-base	140	18.2 ±0.30	16.42±0.58	25.78±0.89	18.2±0.30	13.28±1.84	15.6±2.29
T5-base	220	17.89 ±0.37	14.44±0.82	16.67±2.94	17.02±0.38	11.78±1.35	9.81±2.94
T5-small	60	19.15 ±0.32	14.18±0.53	25.31±2.16	19.15±0.32	8.03±2.72	18.64±5.69

Table 4: Evaluating robustness of different sized fine-tuned models on the `TweetSum` dataset.

dimensions would suffice for human evaluation, and (2) Among the three dimensions, consistency is easiest to use for human evaluation since it only requires the comparison of two summaries.

We collected annotations via the Appen platform (`https://appen.com/`), asking annotators to compare summaries of the perturbed and unperturbed dialogue, ranking their similarity on a Likert scale of 1 (dissimilar) to 4 (identical or paraphrases). To collect annotations, we used the same set of 20 dialogues as in §3.3 from the `TweetSum` dataset. Each dialogue was perturbed with one of the eight categories (utterance- and dialogue-level), yielding 160 summary pairs to be annotated.

We collected 3 annotations per summary pair, totaling 480 annotations; after filtering out noisy annotations, we conducted our analysis on the remaining 314 examples (Appendix A.3 provides annotation procedure and guidelines). We aggregate annotations using majority voting to get similarity scores. To compute consistency scores (equation 1), we map the Likert scale to continuous numeric scores from 0 to 1. We compute mean scores across all pairs for a given dataset and perturbation.

As shown in Figure 2, we observe similar trends, with models exhibiting repetition, long, and lead biases, and that models are affected nearly equally by all utterance perturbations. While the absolute values of Δz_c differ between calculations using automatic metrics and human annotations, the relative impact of different perturbations on the model is similar. For instance, combined utterances and closing remarks have the least impact than repetition, greetings, and split utterance perturbations.[6]

[6]Except time delays, owing to noise in human annotations.

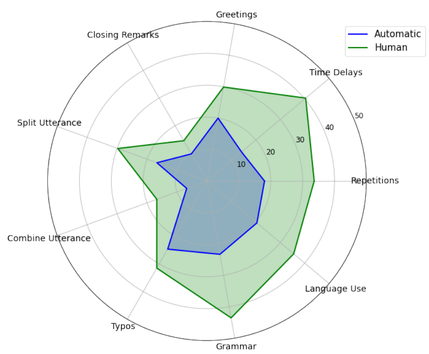


Figure 2: Comparison of consistency scores obtained via human annotations of similarity and the automatic metric on the TweetSum dataset. While the absolute values of Δz_c differ, the relative impact of different perturbations on the model is similar.

5.6 Relationship among dimensions

While theoretically, three dimensions (§4) measure different aspects of robustness, empirically they exhibit a strong correlation of $> 84\%$ across datasets and models (details in Table 10 in Appendix).

This observation can be conceptually explained to some extent. For instance, high saliency implies high consistency: if summaries before and after perturbation are similar to the reference summary, they will be similar to each other, leading to low Δz_s and thus low Δz_c. Similarly, high saliency implies high faithfulness: if the model-generated summary is similar to the reference summary, it will also be factually consistent with the input dialogue, leading to low Δz_s and thus low Δz_f. However, if Δz_s is large, the model could remain faithful under perturbation (small Δz_f): summaries can be different from the reference summary yet consistent with the input dialogue. Thus, conceptually,

Model	Utterance Perturbations			Dialogue Perturbations		
	$\Delta z_c\%$	$\Delta z_s\%$	$\Delta z_f\%$	$\Delta z_c\%$	$\Delta z_s\%$	$\Delta z_f\%$
DIAL-BART0	30.37±0.39	21.80±3.54	37.09±2.57	34.30±0.44	26.44±8.31	47.13±7.51
FLAN-T5	38.23±0.57	41.36±9.10	46.80±14.53	44.12±0.71	39.89±9.09	48.23±11.44
LLAMA-2-70B	47.10±0.17	35.16±0.01	33.19±0.09	54.53±0.48	33.59±0.03	31.69±0.02

Table 5: Robustness of zero-shot summarizers on the TweetSum dataset.

the relation can be explained in only one direction, but empirically the dimensions are highly correlated. Nevertheless, our findings are insightful in their own right, suggesting that the high correlation among all dimensions could be valuable for future robustness studies. For instance, the consistency or faithfulness dimension can serve as reference-free measures of robustness. Consistency is also the easiest to use for human evaluation, as it only requires comparing two summaries.

6 Improving Robustness

One solution to address robustness issues could be to employ reverse heuristics to remove perturbations from dialogues. However, not all perturbations can be easily discovered and removed. For example, in repetition or time delay perturbations, the repeated utterance may include less information or be paraphrased compared to the original. While greetings and closing remarks might be simpler to remove, we include these perturbations as they offer a systematic approach to investigating model behavior, such as potential lead and recency biases.

Another potential solution to address robustness issues can be to use recent large language models to pre-process dialogues by removing errors and repetitions. However, this approach suffers from two challenges: (1) During deployment, additional pre-processing could increase latency, and (2) language models may hallucinate content, posing the risk of introducing factual errors in the input dialogue.

Finally, we examine if training with perturbations can help to mitigate robustness issues. We fine-tune BART on the training data augmented with perturbations and re-evaluate its performance. We create multiple training datasets, each modified by a specific kind of perturbation (typographical errors and language use variations for utterance level; repetitions, split utterances, and greetings for dialogue level), using TweetSum's training split. These modified datasets, with 5-50% of dialogues perturbed, are used to fine-tune BART, which we then test on a similarly altered TweetSum's test split.[7] We hypothesize that training with more perturbed dialogues will initially improve performance until a threshold, after which overfitting may reduce effectiveness.

[7]We experimented with training and evaluating a single model on data with all perturbations. However, since different perturbations can have varied impacts on model performance, we found perturbation-wise analysis more interpretable.

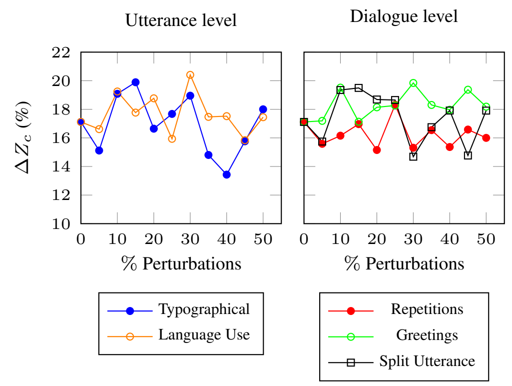


Figure 3: Impact of fine-tuning with perturbations.

Figure 3 shows the change in model consistency when fine-tuned with perturbations. The lower the change in consistency, the higher the model robustness to the perturbations. One takeaway is that different perturbations necessitate varying amounts of perturbed examples in the training set to achieve maximum performance gain. For example, typographical errors and language use variations yield the largest drop in Δz_c when approximately 40% and 45% of the dialogues are perturbed during training. In contrast, dialogue-level perturbations require significantly less perturbed data during training, with approximately 30% split-utterances, 15% greetings, and only 5% repetitions being sufficient. Overall, the results demonstrate that fine-tuning with perturbed data does not yield consistent performance improvements, warranting more detailed exploration as part of future work.

7 Conclusion

We investigate the impact of naturally occurring variations on state-of-the-art dialogue summarization models using three publicly available datasets. To simulate variations, we introduce utterance-level and dialogue-level perturbations. We conduct our analysis using three dimensions of robustness: consistency, saliency, and faithfulness, which capture different aspects of the summarization model's performance. Our results show that both fine-tuned and instruction-tuned models are affected by perturbations, with instruction-tuned models being more susceptible, particularly to dialogue-level perturbations, spurring the need for future research.

8 Limitations

We list some of the limitations of our study which researchers and practitioners would hopefully benefit from when interpreting our analysis. 1) Our analysis uses automatic metrics to measure semantic similarity. Established metrics such BERTScore are imperfect (Deutsch et al., 2022). However, they are widely used in the summarization literature, and also correlate with human judgements of summary quality, and thus are useful for comparing system-level performance. To validate our findings, we also conduct a human evaluation to better understand trends observed due to various perturbations. The investigation of better-automated metrics for natural language generation is an active field of research, and we hope to integrate novel performance metrics in future work. (2) While our perturbations are motivated by real-life scenarios, they are still synthetic in nature. However, we take care wherever possible to avoid unrealistic changes to the dialogues. (3) Our study limits to only open-sourced models and does not investigate the robustness of proprietary LLMs (e.g., ChatGPT), which may be more robust. We decided to limit our study to open-sourced models as it allows us to carefully control what is in the training data, which is not possible with proprietary LLMs and the possibility of data contamination also makes it hard to draw conclusions. (4) Our study mainly focuses on text-based dialogue summarization datasets and does not include spoken conversations, which would bring in very different and diverse nuances of spoken conversations compared to text-based conversations, and is currently out of the scope of this paper. (5) Our study proposes one possible method to measure robustness, and we acknowledge that there can be many other viable ways to quantify robustness. However, quantifying the robustness of tasks involving text generation (e.g., summarization) is an active area of research (Wang et al., 2022) and we hope our work will spur further investigation as part of future work. (6) We did not investigate the robustness of models under both utterance and dialogue level perturbations occurring together in a single dialogue, as that would result in a large number of possible combinations to consider. We leave this for future work.

9 Ethics Statement

All annotators in our human evaluation were recruited via Appen platform and were presented with a consent form prior to the annotation. They were also informed that only satisfactory performance on the screening example will allow them to take part in the annotation task. None of the material/examples they looked at had any hateful or abusive content. We also ensured that the annotators were paid fair amount of wages using Appen's Fair Pay Price Per Judgment which equates to an hourly rate matching a little over the minimum wage of annotators in their respective countries. All the datasets used in this work are publicly available under the CDLA-Sharing license and do not contain any private information.

References

Jiaao Chen and Diyi Yang. 2021. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hyundong Cho, Chinnadhurai Sankar, Christopher Lin, Kaushik Sadagopan, Shahin Shayandeh, Asli Celikyilmaz, Jonathan May, and Ahmad Beirami. 2022. Know thy strengths: Comprehensive dialogue state tracking diagnostics. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5345–5359, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, Haoyue Shi, Yiwen Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Gloria Wang, Zijie J. Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, M. Yee, Jing Zhang, and Yue Zhang. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation.

Arash Einolghozati, Sonal Gupta, Mrinal Mohit, and Rushin Shah. 2019. Improving robustness of task oriented dialog systems. *3rd Conversational AI Workshop at 33rd Conference on Neural Information Processing Systems*.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. TWEETSUMM - a dialog summarization dataset for customer service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *ArXiv*, abs/2107.03175.

Tim Fischer, Steffen Remus, and Chris Biemann. 2022. Measuring faithfulness of abstractive summaries. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 63–73, Potsdam, Germany. KONVENS 2022 Organizers.

Jatin Ganhotra, Robert Moore, Sachindra Joshi, and Kahini Wadhawan. 2020. Effects of naturalistic variation in goal-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4013–4020, Online. Association for Computational Linguistics.

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Lu Wang, and Arman Cohan. 2022. Exploring the challenges of open domain multi-document summarization. *arXiv preprint arXiv:2212.10526*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongyan Jing, Daniel Lopresti, and Chilin Shih. 2003. Summarization of noisy documents: A pilot study. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 25–32.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Z. Hakkani-Tür. 2021. "how robust r u?": Evaluating task-oriented dialogue systems on spoken conversations. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1147–1154.

Kundan Krishna, Yao Zhao, Jie Ren, Balaji Lakshminarayanan, Jiaming Luo, Mohammad Saleh, and Peter J Liu. 2022. Improving the robustness of summarization models by detecting and removing input noise. *arXiv preprint arXiv:2212.09928*.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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

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

Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. 2021. Robustness testing of language understanding in task-oriented dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2467–2480, Online. Association for Computational Linguistics.

Ailsa Meechan-Maddon. 2019. The effect of noise in the training of convolutional neural networks for text summarisation.

Simon Mille, Kaustubh Dhole, Saad Mahamood, Laura Perez-Beltrachini, Varun Gangal, Mihir Kale, Emiel van Miltenburg, and Sebastian Gehrmann. 2021. Automatic construction of evaluation suites for natural language generation datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Robert J. Moore and Raphael Arar. 2019. *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework*. Association for Computing Machinery, New York, NY, USA.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Conference on Empirical Methods in Natural Language Processing*.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544, Online. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2020. Raddle: An evaluation benchmark and analysis platform for robust task-oriented dialog systems. *ArXiv*, abs/2012.14666.

Denis Peskov, Joe Barrow, Pedro Rodriguez, Graham Neubig, and Jordan Boyd-Graber. 2019. Mitigating noisy inputs for question answering. *arXiv preprint arXiv:1908.02914*.

Lihua Qian, Lin Qiu, Weinan Zhang, Xin Jiang, and Yong Yu. 2019. Exploring diverse expressions for paraphrase generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3173–3182, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Harvey Sacks, Emanuel A. Schegloff, and Gail D. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696 – 735.

Sailik Sengupta, Jason Krone, and Saab Mansour. 2021. On the robustness of intent classification and slot labeling in goal-oriented dialog systems to real-world noise. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 68–79, Online. Association for Computational Linguistics.

Xin Tian, Xinxian Huang, Dongfeng He, Yingzhan Lin, Siqi Bao, H. He, Liankai Huang, Qiang Ju, Xiyuan Zhang, Jianyue Xie, Shuqi Sun, Fan Wang, Hua Wu, and Haifeng Wang. 2021. Tod-da: Towards boosting the robustness of task-oriented dialogue modeling on spoken conversations. *ArXiv*, abs/2112.12441.

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

Larry Wasserman. 2004. *All of statistics: a concise course in statistical inference*, volume 26. Springer.

Linzi Xing, Wen Xiao, and Giuseppe Carenini. 2021. Demoting the lead bias in news summarization via alternating adversarial learning. In *Annual Meeting of the Association for Computational Linguistics*.

Alex Yorke. butter-fingers.

Lin Yuan and Zhou Yu. 2019. Abstractive dialog summarization with semantic scaffolds. *arXiv preprint arXiv:1910.00825*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Lulu Zhao, Fujia Zheng, Keqing He, Weihao Zeng, Yuejie Lei, Huixing Jiang, Wei Wu, Weiran Xu, Jun Guo, and Fanyu Meng. 2021. Todsum: Task-oriented dialogue summarization with state tracking. *ArXiv*, abs/2110.12680.

Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. 2020. Convlab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In *Annual Meeting of the Association for Computational Linguistics*.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Brooke Anderson, and Diyi Yang. 2022. Value: Understanding dialect disparity in nlu. *ArXiv*, abs/2204.03031.

A Appendix

A.1 Details/Examples of Perturbations

See Table 6.

A.2 Details of annotation guidelines of quality validation in §5.2

For annotation collection, we only allowed annotators proficient in English from a small group of the most experienced annotators adjudicated by the Appen platform; from any country. We also used hidden test questions for quality control and required annotators to maintain at least 80% accuracy throughout the job on these hidden test questions. These test questions are pre-labeled and are used before and during the task to quiz the annotator. We selected 15 test questions from the validation split of each dataset ensuring that these questions do not overlap with questions seen by the annotators for the actual annotation task. Figure 4 shows the annotation guidelines and Figure 5 shows examples provided for this task.

A.3 Details of annotation guidelines for the validity of trends in §5.6

Quality Control: For this task, as well we only allowed annotators proficient in English from a small group of the most experienced annotators adjudicated by the Appen platform; from any country. We also used hidden test questions for quality control and required annotators to maintain at least 80% accuracy throughout the job on these hidden test questions. Figure 6 shows the annotation guidelines, and Figure 7 shows examples provided for this task.

Number of annotations: In the main task, each annotator was shown 5 examples per page with one hidden test example. For each example, we collected three annotations. In cases where there was no agreement among the initial three annotations, we obtained additional annotations. A maximum of five annotations was considered.

Noise Filtering: Before computing consistency scores, we took several steps to filter out noisy annotations. The Appen platform estimates the trust score for each worker (by calculating accuracy on hidden test examples) and also marks examples as tainted if it is annotated by an annotator whose accuracy score has fallen below the minimum accuracy threshold. To retain only the highest quality annotations, we remove annotations that were marked as tainted and only keep annotations from workers

Perturbation Type	Perturbation Category	Perturbation Name	Examples
Utterance Level	Typographical Errors	remove punctuation	`great!` → `great`
		remove/add whitespace	`Customer` → `Custo mer`
		change letter casing	`action` → `actIon`
		common substitutions expansions	`n't` → `not`
		common substitutions contractions	`I am` → `I'm`
	Grammatical Errors	dropping determiners	`a, the, an`
		subject-verb disagreements	`She likes apples.` → `She like apples.`
		homophone swaps	`their` → `there`
	Spoken Language Errors	filler words and disfluencies	`uhm, uh, erm, ah, er, err,` `actually, like, you know` `I think/believe/mean, I would say` `maybe, perhaps, probably, possibly,` `most likely`
Dialogue Level	Repetitions	N/A	`'Sorry, I couldn't hear you, can you repeat?'`
	Time Delays	N/A	`'Just give me a few minutes..'` `'sure', 'yup!'` `'Thanks for waiting.'`
	Greeting and closing remarks	greeting (Customer Support)	`'Hi! I am your customer support assistant. How may I help you today?'`
		greeting (friends)	`'Hi!' or 'Hey there!'`
		closing (Customer Support)	`'Thank you for contacting us. Have a nice day!'`
		closing (friends)	`'Cool, talk to you later!', 'Bye.'`

Table 6: Examples of each perturbation

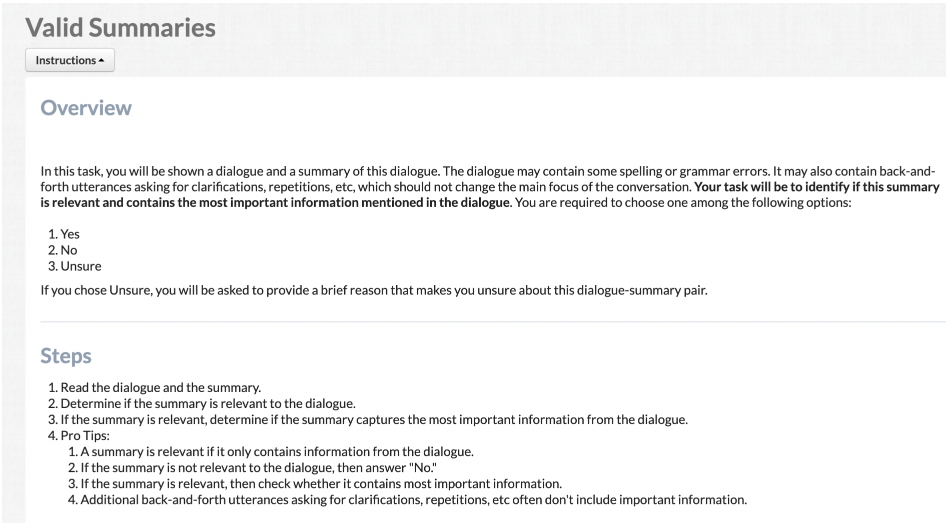


Figure 4: Annotation guidelines for quality validation of perturbed dialogue-summary pairs.

whose trust score is 100%. On qualitatively examining the annotations we also found cases where the two summaries were word-by-word the same, yet the annotator did not give a rating of 4 (highly similar or exact match). Since this is a case of obvious noise, we remove such cases. If an example has less than 3 annotations left after the filtering step, we drop the example. After this filtering, we finally use 314 annotations to conduct our analysis.

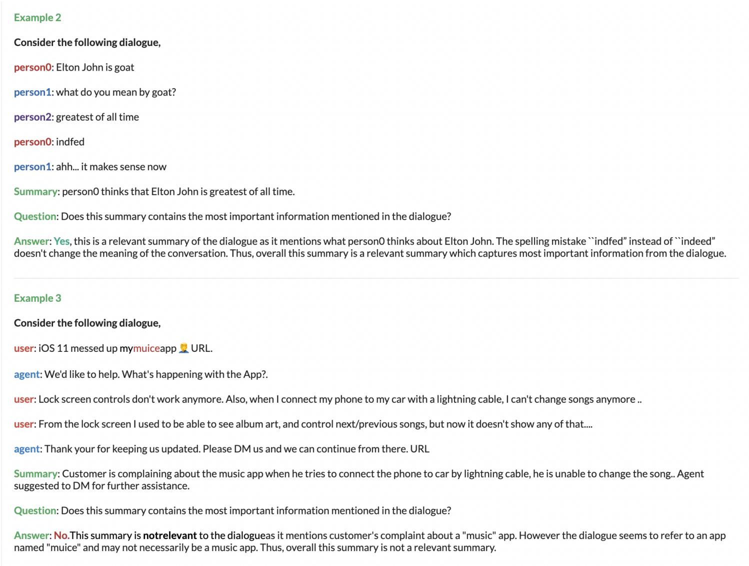
A.4 Targeted dialogue perturbations to investigate the repetition bias

To delve deeper into the repetition bias observed in the models, we conducted targeted perturbations, where we repeat utterances based on whether the information conveyed in those utterances was considered important by the reference summary. Specifically, we identify utterances that are highly relevant and least relevant to the reference summary. To measure relevance, we compute semantic similarity[8] between each utterance and each sentence in the reference summary. For each summary sentence, we then determine the most (least) relevant utterance by selecting the one with the highest (lowest) similarity with the summary sentence. When perturbing the most relevant utterance, we perturb the utterances that were identified as relevant to at least one summary sentence. When perturbing

Dataset	Model	Repeated Utterance Most Relevant	Repeated Utterance Least Relevant	Random
TweetSum	BART	12.40	14.53	14.46
	Pegasus	13.49	16.68	14.22
	T5	9.26	11.46	10.84
TODSum	BART	1.94	4.32	3.52
	Pegasus	2.05	2.05	2.92
	T5	1.85	3.66	3.50

Table 7: Saliency scores of fine-tuned models with targeted perturbations. Perturbing the least relevant utterance results in the highest change in saliency, suggesting that the model exhibits repetition bias.

[8]using sentence transformers [CITE]



Example 2

Consider the following dialogue,

person0: Elton John is goat

person1: what do you mean by goat?

person2: greatest of all time

person0: indfed

person1: ahh... it makes sense now

Summary: person0 thinks that Elton John is greatest of all time.

Question: Does this summary contains the most important information mentioned in the dialogue?

Answer: Yes, this is a relevant summary of the dialogue as it mentions what person0 thinks about Elton John. The spelling mistake ``indfed" instead of ``indeed" doesn't change the meaning of the conversation. Thus, overall this summary is a relevant summary which captures most important information from the dialogue.

Example 3

Consider the following dialogue,

user: iOS 11 messed up mymuiceapp 🤦 URL.

agent: We'd like to help. What's happening with the App?.

user: Lock screen controls don't work anymore. Also, when I connect my phone to my car with a lightning cable, I can't change songs anymore ..

user: From the lock screen I used to be able to see album art, and control next/previous songs, but now it doesn't show any of that....

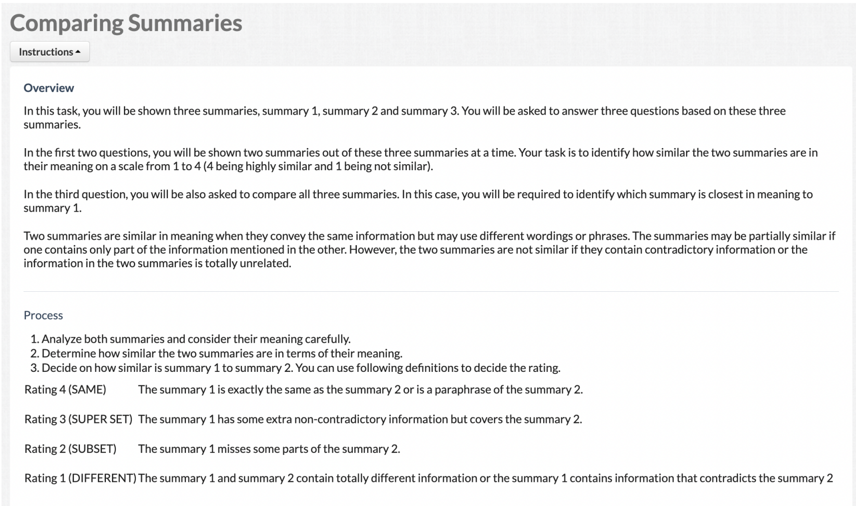
agent: Thank your for keeping us updated. Please DM us and we can continue from there. URL

Summary: Customer is complaining about the music app when he tries to connect the phone to car by lightning cable, he is unable to change the song.. Agent suggested to DM for further assistance.

Question: Does this summary contains the most important information mentioned in the dialogue?

Answer: No.This summary is **notrelevant** to the dialogueas it mentions customer's complaint about a "music" app. However the dialogue seems to refer to an app named "muice" and may not necessarily be a music app. Thus, overall this summary is not a relevant summary.

Figure 5: Examples provided as part of annotation guidelines for quality validation of perturbed dialogue-summary pairs



Comparing Summaries

Instructions

Overview

In this task, you will be shown three summaries, summary 1, summary 2 and summary 3. You will be asked to answer three questions based on these three summaries.

In the first two questions, you will be shown two summaries out of these three summaries at a time. Your task is to identify how similar the two summaries are in their meaning on a scale from 1 to 4 (4 being highly similar and 1 being not similar).

In the third question, you will be also asked to compare all three summaries. In this case, you will be required to identify which summary is closest in meaning to summary 1.

Two summaries are similar in meaning when they convey the same information but may use different wordings or phrases. The summaries may be partially similar if one contains only part of the information mentioned in the other. However, the two summaries are not similar if they contain contradictory information or the information in the two summaries is totally unrelated.

Process

1. Analyze both summaries and consider their meaning carefully.
2. Determine how similar the two summaries are in terms of their meaning.
3. Decide on how similar is summary 1 to summary 2. You can use following definitions to decide the rating.

Rating 4 (SAME)	The summary 1 is exactly the same as the summary 2 or is a paraphrase of the summary 2.
Rating 3 (SUPER SET)	The summary 1 has some extra non-contradictory information but covers the summary 2.
Rating 2 (SUBSET)	The summary 1 misses some parts of the summary 2.
Rating 1 (DIFFERENT)	The summary 1 and summary 2 contain totally different information or the summary 1 contains information that contradicts the summary 2

Figure 6: Annotation guidelines for the validity of trends; to collect similarity annotations for pair of summaries.

the least relevant utterance, we perturb the utterances that were identified as least relevant to all the summary sentences.

As shown in Table 7, we observe that the model exhibits the highest change in saliency scores when we perturb the least relevant utterance, which further demonstrates the model's tendency to consider repeated information as important, even though it was not considered important as per the reference summary. In contrast, repetition of the most relevant utterance shows the least change in the scores, since the model already focuses on the most relevant information before perturbation and after repeating that utterance, it still remains important to be included in the summary.

A.5 Sensitivity to perturbation rate

A.6 Perturbation-wise impact on zero-shot models

See Table 8 and Table 9

A.7 Correlation analysis

Table 10 shows the Pearson correlations between pairs of dimensions on the `TweetSum` dataset. Correlations scores are also visualized in Figures 10, 11, 12. Similar correlation are also observed on `SAMSum` (Figures 14, 15, 13) and `TODSum` datasets (Figures 17, 18, 16).

A.8 Analysis using ROUGE-L and SummaC scores

Example 1 (SAME):

Summary 1: The customer is upset about the train delay. According to the agent, delay repay compensation is available to those affected by disruption who have been delayed for more than 30 minutes.

Summary: 2 Customer is complaining about the delay in the train. Agent states that the delay repay compensation is available to those who have been caught up in disruption and delayed by over 30 minutes.

Answer: 4 (SAME). Both summaries convey the exact same information and have the same meaning. Thus, these summaries are very similar. On a scale of 1-4, these two summaries score 4.

Example 2 (SUPER SET):

Summary 1: Customer is complaining about the fluctuation in the internet which goes in and out and fluctuates speed is also extremely variable. Agent requests to speak privately via direct message so that they can look further into this issue.

Summary 2: The customer is complaining that he was disappointed with the internet services. The agent asked the customer to directly message them for further assistance.

Answer: 3 (SUPER SET) In this example, the summary 1 is a super set of the summary 2. The summary 2 mentions that the customer is complaining about internet services. This information is also mentioned in the summary 1 which provides more details about the complaint (e.g., fluctuating speed). In other words, the summary 1 contains some additional information not present in the summary 2 and thus summary 1 is a superset of summary 2. Both summary 1 and summary 2 mention that the agent requested the user to directly message, thus this portion of the summary 2 exactly matches with the summary 1. Overall, the summary 1 is a super set of the summary 2. On a scale of 1-4, these two summaries score 3.

Example 3 (SUB SET):

Summary 1: Customer is complaining that he is unable to change the reservation. Agent requests to provide confirmation number via DM so that they can assist further.

Summary 2: The customer is complaining that he was trying to contact Delta about changing the reservation. the agent asked to provide their confirmation number via DM so that they can assist

Answer: 2 (SUB SET) In this example, the summary 2 slightly differs from the summary 1. The summary 2 mentions that the customer was trying to contact Delta about changing the reservation. However, the summary 1 mentions that the customer is unable to change the reservation and doesn't mention anything about Delta. At the same time, both summaries are similar when mentioning the agent's response. Thus, overall, the two summaries are only partially similar, with summary 1 missing some information compared to the summary 2. On a scale of 1-4, these two summaries score 2.

Figure 7: Examples provided as part of annotation guidelines to collect similarity annotations for pair of summaries.

Model	Perturbation					
	repetitions	time_delays	greetings	Closing remarks	split_utterances	combined_utterances
DIAL-BART0	35.30	31.15	35.02	23.07	35.10	18.31
FLAN-T5	45.65	32.88	60.10	48.11	41.45	20.34

Table 8: Change in consistency scores due to dialouge-level perturbations on instruction-tuned models when used as zero-shot summarizers. Models are more affected due to repetitions, time-delays, greetings, and split utterances compared to closing remarks and combined utterances.

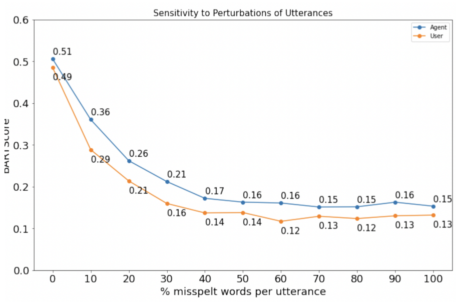


Figure 8: Consistency scores for spelling error perturbation, when varying the percentage of words perturbed per utterance. We perturb all utterances in a dialogue. A perturbation rate of 20% also causes a considerable drop in model performance.

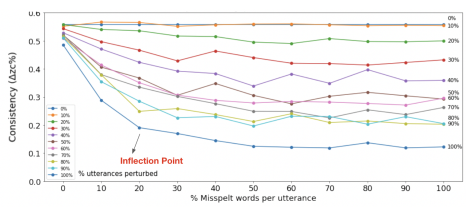


Figure 9: Consistency scores for spelling error perturbation, when varying the percentage of words perturbed per utterance. We also vary the number of utterances being perturbed. Perturbing more than 30% utterances also causes a considerable drop in model performance.

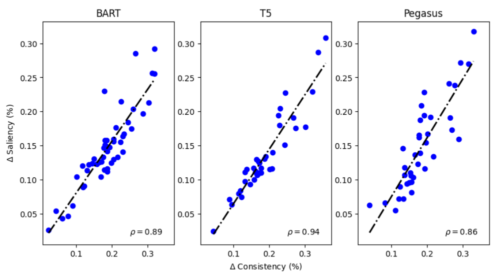


Figure 10: Correlation between consistency and saliency dimensions on TweetSum dataset.

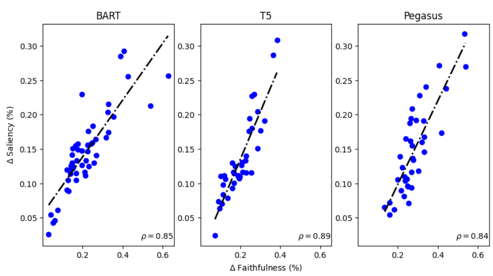


Figure 11: Correlation between faithfulness and saliency dimensions on TweetSum dataset (Outliers excluded for the purpose of visualization).

Model	Perturbation			
	typographical	**grammar**	**language_use**	**speech_recognition**
DIAL-BART0	33.74	32.26	27.53	30.33
FLAN-T5	42.60	48.03	39.75	33.86

Table 9: Change in consistency scores due to utterance-level perturbations on instruction-tuned models when used as zero-shot summarizers. Models are equally affected due to all perturbations.

Model	Pair of dimensions		
	$(\Delta z_c, \Delta z_s)$	$(\Delta z_c, \Delta z_f)$	$(\Delta z_f, \Delta z_s)$
BART	0.89	0.91	0.85
T5	0.94	0.93	0.89
Pegasus	0.86	0.85	0.84

Table 10: Pearson correlations between pairs of dimensions on the `TweetSum` dataset. Similar correlation observed on `SAMSum` and `TODSum` (Appendix A.7).

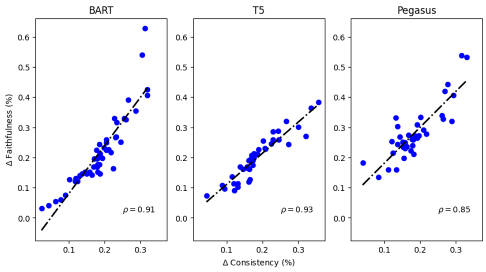


Figure 12: Correlation between faithfulness and consistency dimensions on TweetSum dataset.

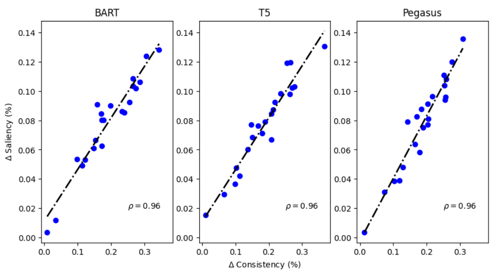


Figure 13: Correlation between consistency and saliency dimensions on SAMSum dataset.

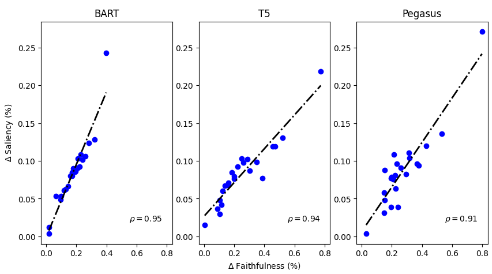


Figure 14: Correlation between faithfulness and saliency dimensions on SAMSum dataset (Outliers excluded for the purpose of visualization).

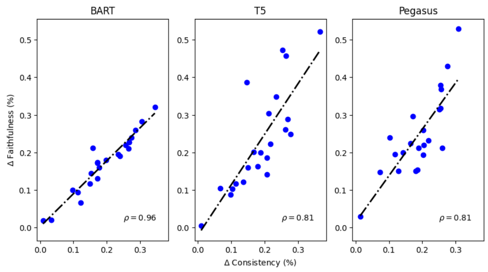


Figure 15: Correlation between faithfulness and consistency dimensions on SAMSum dataset.

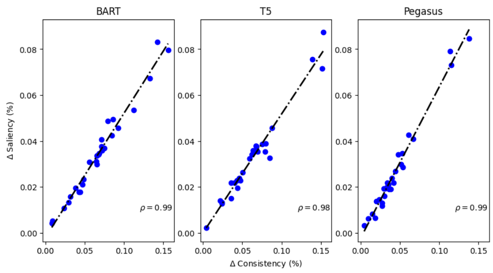


Figure 16: Correlation between consistency and saliency dimensions on TODSum dataset.

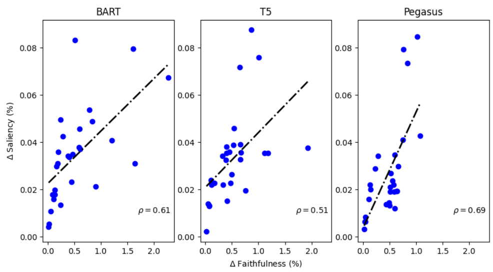


Figure 17: Correlation between faithfulness and saliency dimensions on TODSum dataset (Outliers excluded for the purpose of visualization).

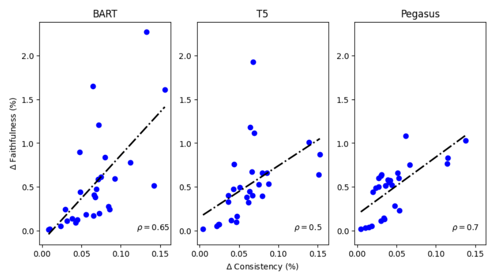


Figure 18: Correlation between faithfulness and consistency dimensions on TODSum dataset.

Model	Utterance Perturbations			Dialogue Perturbations		
	Consistency	Saliency	Faithfulness	Consistency	Saliency	Faithfulness
BART Large	14.00±0.22	10.91±0.01	9.18±0.01	14.37±0.37	10.37±0.01	8.97±0.01
BART Base	14.18±0.29	10.65±0.01	9.60±0.01	15.40±0.31	9.74±0.01	9.04±0.09
Pegasus	13.50±0.46	13.24±0.01	11.29±0.02	14.78±0.39	12.14±0.02	9.80±0.01
T5 Base	14.72±0.36	13.43±0.01	11.01±0.01	13.88±0.42	12.27±0.02	9.79±0.01
T5 Small	14.66±0.33	14.40±0.01	10.11±0.01	15.75±0.31	10.99±0.01	8.72±0.08
DIAL-BART0	29.72±0.36	22.70±0.01	20.53±0.01	34.09±0.30	26.3±0.02	23.29±0.01
FLAN-T5	34.06±0.55	34.63±0.01	36.67±0.02	39.84±0.53	36.98±0.03	40.82±0.06
LLAMA-2	47.1±0.17	35.16±0.01	33.19±0.09	54.53±0.48	33.59±0.03	31.69±0.02

Table 11: Results on TweetSum using ROUGE-L

Model	Utterance Perturbations			Dialogue Perturbations		
	Consistency	Saliency	Faithfulness	Consistency	Saliency	Faithfulness
BART Large	19.18±0.35	6.66±0.01	3.37±0.01	20.85±0.60	7.70±0.02	2.11±0.01
BART Base	19.35±0.41	6.67±0.01	4.23±0.02	21.08±0.47	5.34±0.02	3.07±0.01
Pegasus	19.67±0.50	8.33±0.02	3.75±0.01	21.70±0.53	7.43±0.03	3.67±0.03
T5 Base	19.20±0.50	7.81±0.03	3.87±0.03	21.40±0.58	7.76±0.04	3.44±0.01
T5 Small	20.77±0.55	8.44±0.06	3.69±0.01	21.17±0.63	5.93±0.01	2.38±0.04
DIAL-BART0	43.05±0.52	12.8±0.03	4.55±0.01	51.75±0.47	16.05±0.02	6.32±0.03
FLAN-T5	39.54±0.64	14.96±0.00	5.95±0.01	45.93±0.65	15.35±0.04	7.72±0.02
LLAMA-2	45.05±0.44	20.51±0.04	18.06±0.02	56.32±0.43	20.58±0.11	12.79±0.06

Table 12: Results on TweetSum using SummaC

Dimension	Repetitions	Time Delays	Greetings	Conclusion	Split Utterances	Combine Utterances
Consistency	31.03±0.52	25.73 ±0.77	36.89±1.07	18.17±0.95	13.34±0.75	8.7±0.62
Saliency	12.16±0.66	9.64±0.97	16.72±2.36	5.62±0.73	11.63±1.05	6.62±0.77
Faithfulness	10.17±0.45	7.54±0.58	10.84±0.93	5.3±0.69	8.96±0.6	5.33±0.49

Table 13: Impact of Dialouge Perturbations on TweetSum using ROUGE-L