Pointwise Mutual Information Based Metric and Decoding Strategy for Faithful Generation in Document Grounded Dialogs

Yatin Nandwani and **Vineet Kumar** and **Dinesh Raghu** and **Sachindra Joshi** and **Luis A. Lastras**
IBM Research, AI
{yatin.nandwani@, vineeku6@in, diraghu1@in, jsachind@in, lastrasl@us}.ibm.com

Abstract

A major concern in using deep learning based generative models for document-grounded dialogs is the potential generation of responses that are not *faithful* to the underlying document. Existing automated metrics used for evaluating the faithfulness of response with respect to the grounding document measure the degree of similarity between the generated response and the document's content. However, these automated metrics are far from being well aligned with human judgments. Therefore, to improve the measurement of faithfulness, we propose a new metric that utilizes (Conditional) Point-wise Mutual Information (PMI) between the generated response and the source document, conditioned on the dialogue. PMI quantifies the extent to which the document influences the generated response – with a higher PMI indicating a more faithful response. We build upon this idea to create a new decoding technique that incorporates PMI into the response generation process to predict more faithful responses. Our experiments on the BEGIN benchmark demonstrate an improved correlation of our metric with human evaluation. We also show that our decoding technique is effective in generating more faithful responses when compared to standard decoding techniques on a set of publicly available document-grounded dialog datasets.

1 Introduction

Document–grounded dialog agents converse with users by using a specific set of documents provided to them. These agents are designed to be faithful to the information present in the grounding document and refrain from offering any information that cannot be verified through it. As most existing document–grounded dialog agents (Prabhumoye et al., 2021; Wu et al., 2021) are built by fine-tuning large language models, ensuring faithful response generation is a major challenge.

To measure the ability of dialog agents to generate faithful responses, several automatic metrics have been proposed. These metrics take as input the response generated by the agent and the grounding document to quantify faithfulness. These metrics are based on lexical overlap (e.g., BLEU, unigram-F1), semantic overlap (BERTScore) or a trained classifier (Dziri et al., 2022a). Recently, Honovich et al. (2021) proposed Q^2, a metric that classifies a response as faithful if questions generated from the response predict the same answers when grounded on both the response and the document.

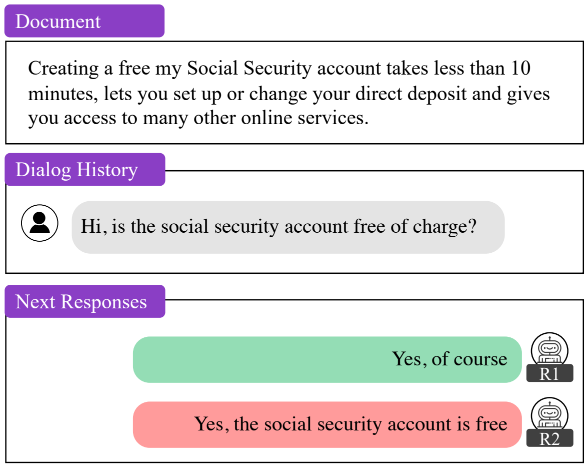

Figure 1: An example document grounded dialog with two types of responses: sentential response (R2) and non-sentential response (R1).

A major concern with the existing faithfulness metrics is their inability to measure the faithfulness of non-sentential responses. These are incomplete responses whose meaning can be understood only with the help of the dialog-so-far. For example, the response *R1* in Figure 1 is non-sentential, whereas the response *R2* is its corresponding sentential form. Existing automated metrics often correctly classify faithful sentential responses as faithful, but fail to classify non-sentential faithful responses as faithful. The main difference is that non-sentential responses cannot be understood in the absence of the dialog context. As existing metrics do not use the dialog history while predicting faithfulness, they fail to correctly classify non-sentential responses. For example, as non-sentential responses do not have domain words in them, similarity–based metrics such as unigram-F1 and BERTScore fail. Also, as generating QA pairs using these responses typically results in question/answer capturing incomplete information, Q^2 would also fail. Non-sentential utterances are quite frequent in dialog data. Schlangen (2003) estimates the frequency of non-sentential utterances to be 20% of all the utterances.

To overcome this problem, we propose a new metric that quantifies the faithfulness of a generated response with respect to both the document and the dialog history. Our metric is grounded in information theoretic concepts and captures the association of the response with the given document using Conditional Pointwise Mutual Information (CPMI). We call our metric PMI–FAITH, which uses CPMI between the generated response and the document, conditioned on the dialogue history, for quantifying faithfulness. PMI–FAITH captures the intuition that for a response to be grounded in the document, the probability of its generation given the document should be higher than the probability of its generation without the document.

A significant advantage of our metric PMI–FAITH is that it can be factorized the same way as the likelihood of a response can be factorized in auto regressive models. We take advantage of this property to propose a novel decoding strategy, PMI–DECODE. The goal of PMI–DECODE is to maximize not just the response's likelihood but a score that combines its likelihood and faithfulness.

To summarize, our contributions are threefold:

1. We propose PMI–FAITH, a novel metric which quantifies faithfulness as a conditional PMI between the response and the document.
2. We propose a novel decoding strategy, PMI–DECODE, which can aid in generating faithful responses.
3. Our experiments show that PMI–FAITH correlates with human judgments better than any existing metrics on BEGIN (Dziri et al., 2022b). We also show that PMI–DECODE generates more faithful responses than greedy decoding on three standard document–grounded dialog datasets.

2 Related Works

In this section, we discuss related works in the areas of quantifying faithfulness and mitigating hallucination in document–grounded dialogue systems.

Faithfulness: Researchers have used various terms such as faithfulness (Cao et al., 2018), factual consistency (Cao et al., 2020; Santhanam et al., 2021), factual accuracy (Goodrich et al., 2019), fidelity (Chen et al., 2020), attribution (Rashkin et al., 2021a) and hallucination (i.e., the lack of faithfulness) (Xiao and Wang, 2021) to define and quantify faithfulness of a model's generated text to a given knowledge.

Most of the works focusing on evaluating faithfulness propose to train a classifier for the task (Goodrich et al., 2019; Kryscinski et al., 2020; Dziri et al., 2022a). Whereas our proposed metric doesn't require any training and is agnostic to the underlying data. Recently, Honovich et al. (2021) proposed Q^2 for quantifying faithfulness. It uses a question generator to first generate question–answer (QA) pairs from the generated response. Then a QA system is used to find an answer, to the generated question, from the document. Finally, an NLI system is used to compare the two answers. Though Q^2 uses the given document to check the faithfulness of a response, it ignores the dialog history. Thus, it may fail at handling non–sentential responses as depicted in fig. 1. Our metric PMI–FAITH addresses this issue.

Mitigating Hallucination: Hallucination can be reduced by pre–processing data and removing instances from training data that are factually incorrect (Shen et al., 2021). It can also be reduced by training control tokens and using them during response generation (Filippova, 2020; Rashkin et al., 2021b). We focus on decoding and do not require explicit changes to the training procedure or data.

In addition to the above two areas, many recent works (Dziri et al., 2022b; Honovich et al., 2022) have released different benchmarks that can be used to evaluate the performance of faithfulness metrics. While Honovich et al. aim to standardize benchmark datasets across different generation tasks, Dziri et al. focus on document–grounded dialogues, and thus we use it to compare our metric with various baselines. Finally, Ji et al. (2022) presents a survey that discusses hallucination in various natural language generation tasks, including document–grounded dialogue response generation.

3 Background

In this section, we first review the task of document-grounded dialog response generation, followed by the definition of the faithfulness metric.

Document Grounded Response Generation: Let dialog history $\mathbf{h} = [u_1, \cdots u_m]$ be a sequence of m utterances in the dialog so far and $\mathbf{d}$ be the document on which the next response is grounded. The task of document-grounded dialog response generation is to predict the next response, $\mathbf{r} = \langle r_1 r_2 \ldots r_T \rangle$, one token at a time, given the dialog history $\mathbf{h}$ and the document $\mathbf{d}$. Here, $\forall i, r_i \in \mathcal{V}$, where $\mathcal{V}$ is the vocabulary of all possible tokens. The underlying model learns a probability distribution $P(\mathbf{r}|\mathbf{d}, \mathbf{h})$ over all possible responses $\mathbf{r} \in \mathcal{V}^+$. Typically, this distribution is factorized over the tokens of $\mathbf{r}$ as:

$$P(\mathbf{r}|\mathbf{d}, \mathbf{h}) = \prod_{t=1}^{T} P(r_t|\mathbf{d}, \mathbf{h}, \mathbf{r}_{1:t-1}) \tag{1}$$

Faithfulness Metric: Most of the existing definitions (and metrics) for faithfulness focus mainly on document $\mathbf{d}$ and response $\mathbf{r}$ but ignore the history $\mathbf{h}$ (Dziri et al., 2022a,b). This may be for the sake of uniformity across different tasks such as summarization, grounded dialogue generation, and paraphrase generation. We qualify the definition of faithfulness specifically for the task of document–grounded dialogue generation. Formally, a response $\mathbf{r}$ is considered 'faithful' to a given document $\mathbf{d}$ and the dialogue history $\mathbf{h}$ iff $\mathbf{d}, \mathbf{h} \vDash \mathbf{r}$, where $\vDash$ represents logical entailment.

A faithfulness metric should quantify the faithfulness of the response $\mathbf{r}$ to the document $\mathbf{d}$ and dialogue history $\mathbf{h}$. In general, such a metric should take $\mathbf{r}, \mathbf{d}$ and $\mathbf{h}$ as its input and compute a score, $F(\mathbf{r}, \mathbf{d}, \mathbf{h}) \in \mathbb{R}$, such that a higher value of $F(\mathbf{r}, \mathbf{d}, \mathbf{h})$ indicates a more faithful response.

4 Approach

In this section, we describe our proposed metric for faithfulness – PMI–FAITH. We then propose a novel decoding strategy PMI–DECODE based on our metric, with the objective of generating relevant and faithful responses.

4.1 PMI–FAITH

PMI–FAITH is based on the information–theoretic concept of Pointwise Mutual Information. We use the notion of CPMI between generated response $\mathbf{r}$ and the document $\mathbf{d}$ given the context $\mathbf{h}$ to capture the influence of the document in generating the response. We define our metric, PMI–FAITH, for faithfulness of the response $\mathbf{r}$ to the document $\mathbf{d}$ as:

$$\begin{aligned} \text{PMI–FAITH}(\mathbf{r}, \mathbf{d}, \mathbf{h}) &= \text{CPMI}(\mathbf{r}; \mathbf{d}|\mathbf{h}) \\ = \log \frac{P(\mathbf{r}, \mathbf{d}|\mathbf{h})}{P(\mathbf{r}|\mathbf{h})P(\mathbf{d}|\mathbf{h})} &= \log \frac{P(\mathbf{r}|\mathbf{d}, \mathbf{h})}{P(\mathbf{r}|\mathbf{h})} \end{aligned} \tag{2}$$

Mathematically, PMI is a measure of the strength of the association between two random events. A positive value of CPMI in eq. (2) implies that the probability of generating the response given the document and the dialogue history is higher than the probability of generating the response given only the dialogue history. Hence, the response is likely to be grounded in the document. On the other hand, if the response $\mathbf{r}$ is not faithful to the document $\mathbf{d}$, the probability of its generation given the document and the dialogue history is likely to be similar to the probability of its generation without the document, resulting in a lower value of PMI–FAITH. We use pre–trained language models such as BLOOM (Scao et al., 2022) or GPT2 (Radford et al., 2019), to compute these conditional probabilities $P(\mathbf{r}|\mathbf{d}, \mathbf{h})$ and $P(\mathbf{r}|\mathbf{h})$.

4.2 PMI–DECODE

PMI–DECODE is a decoding strategy whose objective is to generate responses that are both relevant and faithful. Typically, the goal of any decoding strategy is to select a response that has the maximum (log) likelihood:

$$\mathbf{r} = \arg\max_{\tilde{\mathbf{r}} \in \mathcal{V}^+} \log P(\tilde{\mathbf{r}}|\mathbf{d}, \mathbf{h})$$

The objective of PMI–DECODE is to select a response that is highly likely and faithful. This is achieved by maximizing a combination of likelihood and faithfulness quantified using an appropriate metric F. With $\alpha \in [0, 1]$, and a linear scoring function, we get:

$$\mathbf{r} = \arg\max_{\tilde{\mathbf{r}} \in \mathcal{V}^+} (1 - \alpha) \log P(\tilde{\mathbf{r}}|\mathbf{d}, \mathbf{h}) + \alpha F(\tilde{\mathbf{r}}, \mathbf{d}, \mathbf{h}) \tag{3}$$

With an auto–regressive model that generates the response one token at a time, we use decoding strategies, such as greedy decoding, beam search, nucleus sampling (Holtzman et al., 2020), or beam sampling as a heuristic to find the maxima. For ease of description, we use the greedy decoding below, though our approach is agnostic to the choice of heuristic for maximising the objective function. It just modifies the standard log–likelihood objective with an additional term corresponding to faithfulness. Our choice of PMI–FAITH as function F for quantification of faithfulness keeps the decoding heuristic tractable as shown below.

With eq. (3) as our modified objective, greedy decoding would sample the next token r_t such that it maximizes the overall score of the partial response $\langle r_1 r_2 \dots r_t \rangle$:

$$\begin{aligned} r_t &= \arg\max_{v \in \mathcal{V}} (1-\alpha) \log P(\langle r_1 \dots r_{t-1} v \rangle | \mathbf{d}, \mathbf{h}) \\ &\qquad + \alpha F(\langle r_1 \dots r_{t-1} v \rangle, \mathbf{d}, \mathbf{h}) \\ &= \arg\max_{v \in \mathcal{V}} (1-\alpha) [\log P(\langle r_1 r_2 \dots r_{t-1} \rangle | \mathbf{d}, \mathbf{h}) \\ &\qquad + \log P(v | \mathbf{d}, \mathbf{h}, \mathbf{r}_{1:t-1})] \\ &\qquad + \alpha F(\langle r_1 \dots r_{t-1} v \rangle, \mathbf{d}, \mathbf{h}) \end{aligned} \tag{4}$$

In eq. (4), the likelihood term has been factorized and notice that its first term is independent of the next token candidate v and thus can be dropped while taking $\arg\max$. Not all faithfulness metrics can be decomposed the same way as the likelihood term. One advantage of PMI–FAITH is that it can be decomposed the same way as likelihood as follows:

$$\text{PMI–FAITH}(\langle r_1 \dots r_{t-1} v \rangle, \mathbf{d}, \mathbf{h})$$

$$= \log \frac{P(\langle r_1 \dots r_{t-1} v \rangle | \mathbf{d}, \mathbf{h})}{P(\langle r_1 \dots r_{t-1} v \rangle | \mathbf{h})} \tag{5}$$

$$\begin{aligned} &= \log \frac{P(\langle r_1 r_2 \dots r_{t-1} \rangle | \mathbf{d}, \mathbf{h})}{P(\langle r_1 r_2 \dots r_{t-1} \rangle | \mathbf{h})} \\ &\quad + \log \frac{P(v | \mathbf{d}, \mathbf{h}, \mathbf{r}_{1:t-1})}{P(v | \mathbf{h}, \mathbf{r}_{1:t-1})} \end{aligned} \tag{6}$$

$$\begin{aligned} &= \text{PMI–FAITH}(\mathbf{r}_{1:t-1}, \mathbf{d}, \mathbf{h}) \\ &\quad + \text{CPMI}(v; \mathbf{d} | \mathbf{h}, \mathbf{r}_{1:t-1}) \end{aligned} \tag{7}$$

Now replacing $F(\langle r_1 \dots r_{t-1} v \rangle$ in eq. (4) by PMI–FAITH from eq. (7), and dropping the two terms which are independent of v, we get:

$$\begin{aligned} r_t = \arg\max_{v \in \mathcal{V}} & (1-\alpha) \log P(v | \mathbf{d}, \mathbf{h}, \mathbf{r}_{1:t-1}) \\ & + \alpha \text{CPMI}(v; \mathbf{d} | \mathbf{h}, \mathbf{r}_{1:t-1}) \end{aligned} \tag{8}$$

To compute CPMI in eq. (8), the same language model can be used to get the conditional probabilities $P(v|\mathbf{d}, \mathbf{h}, \mathbf{r}_{1:t-1})$ and $P(v|\mathbf{h}, \mathbf{r}_{1:t-1})$ by separately passing **d**,**h**, $\mathbf{r}_{1:t-1}$ and **h**, $\mathbf{r}_{1:t-1}$, respectively, through the model.

We observed that using CPMI in the scoring function sometimes results in selecting tokens from the document which may interfere with the grammar. To mitigate this, instead of maximizing over the entire vocabulary $\mathcal{V}$ at each step t, we propose to maximize only over the '*top p*' subset from the likelihood distribution, $\mathcal{V}_{p,t}$, defined as the minimum cardinality subset of tokens with sum of their probabilities as p. We call this as top p masking:

$$\begin{aligned} r_t = \arg\max_{v \in \mathcal{V}_{p,t}} & (1-\alpha) \log P(v | \mathbf{d}, \mathbf{h}, \mathbf{r}_{1:t-1}) \\ & + \alpha \text{CPMI}(v; \mathbf{d} | \mathbf{h}, \mathbf{r}_{1:t-1}) \end{aligned} \tag{9}$$

The intuition here is that while CPMI has a positive influence on generating a more faithful response, it may negatively impact the grammatical structure. Therefore by restricting the vocabulary to $\mathcal{V}_{p,t}$, we use only highly probable tokens to form a response and thus are likely to generate responses that are faithful as well as grammatically correct.

5 Experiments

The goals of our experiments are: (1) to validate the ability of our metric, PMI–FAITH, to identify faithful responses (section 5.1), and (2) to demonstrate that our novel decoding strategy, PMI–DECODE, that optimizes a combination of likelihood and faithfulness indeed generates more faithful responses than greedy decoding that maximizes only the likelihood (section 5.2). For all our experiments, we use datasets, code and libraries that are publicly available under appropriate licenses for research use. We cite them wherever appropriate.

5.1 Evaluation of Faithfulness Metrics

We exploit an existing dataset of human annotations of faithfulness proposed by Dziri et al. (2022b) and show that PMI–FAITH performs better than existing metrics in identifying faithful responses.

Dataset: We experiment using recently proposed BEGIN benchmark (Dziri et al., 2022b) for evaluating the ability of PMI–FAITH to identify faithful responses. The benchmark uses three document grounded datasets, *viz*, CMU–DoG (Zhou et al., 2018), TopicalChat (Gopalakrishnan et al., 2019), and WoW (Dinan et al., 2019). It contains over $11,059$ responses generated by three different models, GPT2 (Radford et al., 2019), DoHA (Prabhumoye et al., 2021) and T5 (Raffel et al., 2020),

		Lexical			Semantic	Trained Model	QA-QG	Our
		F1-U	BLEU	RougeL	BERT	faithc.	Q^2	PMI–FAITH
All	P	0.401	0.478	0.487	0.459	**0.684**	0.517	0.607
	R	0.785	0.479	0.552	0.673	0.492	0.744	**0.818**
	F1	0.531	0.479	0.518	0.546	0.573	0.610	**0.697**
	Acc.	0.677	0.757	0.760	0.739	0.829	0.779	**0.834**
CMU	F1	0.462	0.027	0.140	0.294	0.156	0.543	**0.663**
TC	F1	0.429	-	0.050	0.432	0.039	0.487	**0.584**
WoW	F1	0.587	0.620	0.647	0.612	**0.794**	0.705	0.771

Table 1: Precision (P), Recall (R), F1 score (F1), and Accuracy (Acc.) of different faithfulness metrics on the BEGIN Benchmark. Bottom three rows report F1 score (F1) over each of the three contributing datasets in the Benchmark, *viz.*, CMU–DoG (CMU), TopicalChat (TC), and Wizards of Wikipedia (WoW). F1-U: Unigram F1, faithc.: faithcritic.

on randomly selected samples from test splits of the 3 datasets. Each generated response is annotated by humans and classified into either 'Fully–attributable', 'Generic', or 'Not fully–attributable'. Overall, 23.3% of the total response have been classified as 'Fully–attributable' by human annotators.

Baselines: In addition to Q^2, we compare against various lexical and semantic similarity-based, and trained classifiers as faithfulness metrics. Specifically, we use Unigram-F1, SacreBLEU(Post, 2018; Papineni et al., 2002), and RougeL (Lin, 2004) to capture lexical overlap between **d** and generated response **r**; BERTScore(Zhang et al., 2020) to capture **r**'s semantic similarity with **d**. We use the code[1] provided by Honovich et al. (2021) for all the above baselines.[2] We also compare against *faithcritic*[3](Dziri et al., 2022a), which is a pre–trained classifier to predict faithfulness of a response.

Evaluation protocol: Our objective is to identify faithful responses. Accordingly, we consider 'Fully–attributable' as the positive class and both 'Generic' and 'Not fully–attributable' as the negative class. For each metric, we identify a threshold and classify a response as faithful iff its value is greater than the threshold. We use F1 score over the devset to identify the optimal thresholds.

Computation of PMI–FAITH: To measure PMI–FAITH$(\mathbf{r}, \mathbf{d}, \mathbf{h})$, we need to compute two conditional probabilities: $P(\mathbf{r}|\mathbf{d}, \mathbf{h})$, and $P(\mathbf{r}|\mathbf{h})$. To do so, we use pretrained LLMs available off the shelf from huggingface library (Wolf et al., 2019). To quantify the impact of using one language model over the other, we compute the performance of PMI–FAITH using eight LLMs of varying sizes: five BLOOM (Scao et al., 2022) models with up to 7 billion parameters, and three GPT2 (Radford et al., 2019) models up to GPT2-large (774 million). We observe a robust and consistent performance with a variability of only 0.02 points in the F1 score. Hence, for all further experiments, we use BLOOM-560m.

Results: Table 1 reports the precision, recall, F1 score, and accuracy achieved by different metrics on the test split of BEGIN benchmark. We first observe that PMI–FAITH performs better than all other faithfulness metrics by a huge margin across all reported performance measures, with the absolute gains ranging from 21.8% to 8.7% in F1 score. Even against the strong baseline of Q^2, it achieves an absolute gain of 5.6% and 8.7% in accuracy and F1 score, respectively. As expected, all the lexical overlap and semantic similarity based metrics achieve poor performance, with accuracy worse than even the majority–class classifier's accuracy of 76.7%. Next, we notice that all metrics, except faithcritic, have higher recall than precision, indicating that they tend to be lenient while classifying a response as faithful, whereas faithcritic tends to be conservative and classifies most of the responses as not faithful. Comparing the next two best metrics, we observe that Q^2 has better F1 and recall, but worse accuracy and precision than faithcritic.

To identify dataset specific biases, Table 1 also reports F1 score separately for each of the three contributing datasets. We observe that PMI–FAITH achieves the highest F1 on CMU–DoG and Top-

[1] *https://github.com/orhonovich/q-squared*

[2] For BERTScore, we change the underlying model to the current best *microsoft/deberta-xlarge-mnli*.

[3] *https://huggingface.co/McGill-NLP/roberta-large-faithcritic*

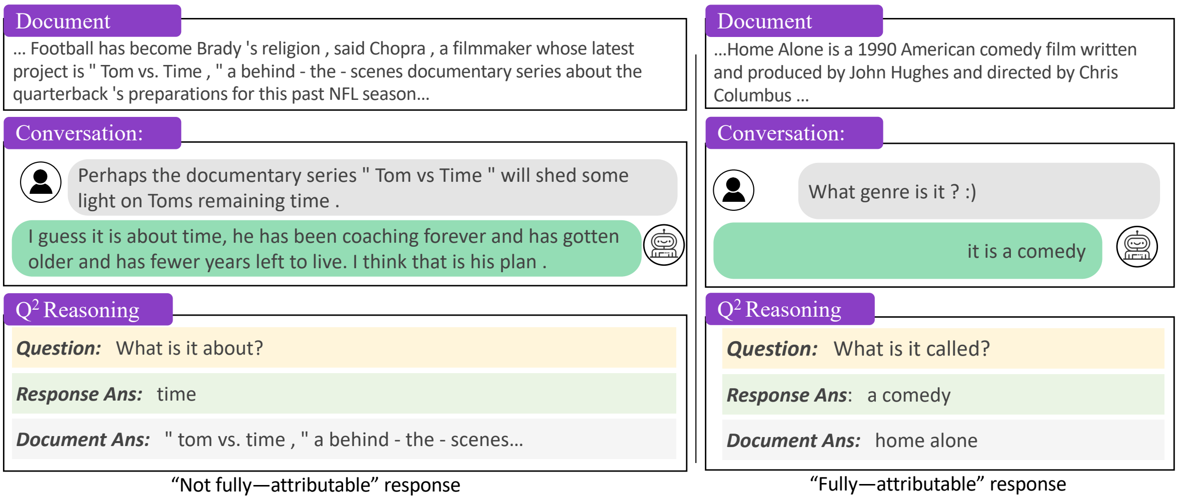

Figure 2: A 'Fully attributable' (right) and a 'Not fully attributable' (left) example, incorrectly classified by Q^2 and correctly classified by PMI–FAITH.

icalChat with more than 12% and 9.7% absolute gain, respectively, over the other metrics. Surprisingly, faithcritic achieves the best F1 score of 0.794 points on WoW whereas its F1 on TopicalChat and CMU–DoG is only 0.039 and 0.156 points, respectively. This over-fitting on WoW could be due to the fact that faithcritic is a learned metric and the training data for it has been adapted from WoW, and its low performance on the other two datasets demonstrates its lack of generalization.

Subjective Analysis: Figure 2 shows two examples where PMI–FAITH correctly identifies the faithfulness (or lack of it) whereas the strongest baseline Q^2 fails to do so. In the case of 'Fully–attributable' response (right), the pronoun '*it*' in the response is an anaphora, referring back to the antecedent 'home alone' (movie name), which is difficult to infer without the dialogue context. However, Q^2 doesn't take the dialogue history into account, and thus it considers the pronoun *it* in the response as a cataphor, referring to its postcedent 'comedy'. As a result, the QA system correctly answers the generated question 'What is *it* called?' with the postcedent '*comedy*', when presented with the response, and correctly outputs its antecedent (*home alone*) when presented with the document. But the overall Q^2 system fails, as the two answers do not match. On the other hand, by virtue of considering dialogue history during computation, PMI–FAITH has information that the question is about the genre and not the movie name, and hence it can correctly classify the response as 'Fully–attributable'.

The other example highlights two issues: (1) when the response is partially hallucinated, QG system may generate question from the faithful part of the response and may incorrectly declare the whole response as faithful. In this example, most of the response contains an opinion, which is not faithful to the document, but the QG system focused on '*I guess it is about time*'. (2): the other issue is that the NLI system fails to capture that the single word answer '*time*' from the response is not entailed by the long answer from the document, resulting in the incorrect prediction by the overall system. On the other hand, PMI–FAITH considers the response as a whole, instead of separately focusing on parts of it. As a result, it is correctly able to identify the given response as not faithful.

5.2 Evaluation of Decoding Strategies

We fine-tune a BART-Large (Lewis et al., 2019) LM separately for three document-grounded dialog datasets and compare the responses generated by PMI–DECODE with responses from standard likelihood-based decoding. In the experiments below, we use beam search with beam size 4 to optimize both the scoring functions. We compare the decoding strategies using various automated metrics for faithfulness and relevance, including PMI–FAITH and Q^2, followed by a human study on randomly selected samples from two of the three datasets.

Datasets: We perform our experiments on three document-grounded dialog datasets: Multi-Doc2Dial (Feng et al., 2021), TopicalChat (Gopalakrishnan et al., 2019) and FaithDial (Dziri

	Num. samples			Avg. words		
	Train	**Dev.**	**Test**	**Doc.**	**Hist.**	**Resp.**
MD2D	24,603	4,699	4,567	166	93	18
TC	131,555	8,183	8,301	241	199	20
FD	18,357	3,417	3,539	23	69	18

Table 2: Number of train, dev and test samples for the three datasets, and average number of words in the document (doc.), dialogue history (Hist.), and gold response (Resp.) in the test split.

et al., 2022b). The dialogs in MultiDoc2Dial (MD2D) are grounded on multiple domain-specific documents. Each dialog is between a user and an agent. Only the agent has access to the documents. So, we only model the agent responses for this dataset. TopicalChat (TC) consists of dialogs between two parties, where each party may have a different set of documents on the same topics. These documents can be news articles, Reddit snippets, or Wikipedia snippets. We exclude the utterances which used 'personal knowledge' as one of the document sources. We use the 'rare' version of TC for our experiments. FaithDial (FD), a faithful adaptation of WoW(Dinan et al., 2019), in which one participant can ask a wide range of questions and the other participant can only provide information from Wikipedia. See table 2.

Decode Method	Faithfulness (Knowledge)				Relevance (Gold)	
	PMIF	**F1-U**	**BERT**	**Q2**	**BLEU**	**RougeL**
	MultiDoc2Dial					
Greedy	0.59	0.25	0.15	0.63	**30.56**	**0.49**
PMI–D	0.64	0.29	0.19	0.65	28.95	0.47
PMI–D_{NM}	**0.69**	**0.33**	**0.24**	**0.69**	24.59	0.43
PMI–D_{EQ}	0.65	0.29	0.19	0.65	27.72	0.46
	TopicalChat					
Greedy	0.49	0.10	0.02	0.68	**6.63**	**0.22**
PMI–D	0.55	0.12	0.05	0.72	5.92	0.20
PMI–D_{NM}	**0.57**	**0.12**	**0.06**	**0.73**	5.60	0.20
PMI–D_{EQ}	0.56	0.12	0.06	0.72	5.40	0.19
	FaithDial					
Greedy	0.54	0.64	0.55	0.83	**13.53**	**0.40**
PMI–D	0.59	0.75	0.66	0.85	12.95	0.40
PMI–D_{NM}	**0.63**	**0.79**	**0.71**	**0.87**	12.17	0.39
PMI–D_{EQ}	0.60	0.74	0.66	0.84	12.39	0.39

Table 3: Faithfulness and relevance metrics computed on the test set for various datasets after finetuning bart-large. BLEU is computed on a scale of $[0, 100]$. All other metrics lie in the range $[0, 1]$. PMI–D uses $\alpha = 0.25$ and *top* $p = 0.6$; PMI–D_{NM} uses $\alpha = 0.25$ without masking; PMI–D_{EQ} uses $\alpha = 0.5$ and *top* $p = 0.6$.

Training Details: For each of the three datasets,

	Multi-Doc2Dial			Topical Chat		
	Fai	Rel	Gra	Fai	Rel	Gra
Greedy	0.52	0.72	0.96	0.69	0.70	0.96
PMI-D	0.60	0.75	0.92	0.80	0.67	0.93

Table 4: Human evaluation on different decoding strategies. We evaluate faithfulness (Fai), relevance (Rel) and grammar (Gra).

we separately finetune a bart-large[4] model using the code[5] made available by Dziri et al. (2022a). Unless otherwise specified, we use the default parameter values. For both MD2D and TC, we use a batch size of 16 and train for 50 and 10 epochs, with patience of 20 and 5 respectively. For all three datasets, we use the model checkpoint which returns the best perplexity score on the dev set. Each model takes less than 14 hours of fine–tuning on an Nvidia A-100 GPU with 80GB memory.

Results: In table 3, we report unigram F1(lexical), BERTScore(semantic), Q^2, and PMI–FAITH, averaged over the entire test set, as the faithfulness metrics. For relevance, we report BLEU-4 and RougeL of the generated response computed *w.r.t.* the gold response. Note that PMI–FAITH is the difference of two log probabilities, and to normalize it between 0 and 1, we use the min.(-2.1) and max.(6.4) value of PMI–FAITH computed on the devset of BEGIN benchmark. For PMI–DECODE (PMI–D), we use $\alpha = 0.25$ as the weight of CPMI (eq. (9)), and mask with top $p = 0.6$. We report ablations on both the hyper–parameters: PMI–D_{NM} doesn't use *top p* masking and sets top $p = 1.0$, and PMI–D_{EQ} ablates on α with $\alpha = 0.5$.

We first observe that all versions of PMI–DECODE improve faithfulness compared to standard likelihood-based greedy decoding. We find that while PMI–D_{NM} has the highest faithfulness metrics, its relevance scores are the lowest. Looking at some examples, we observed that a few of them were not even grammatically correct. This is because CPMI adversely penalizes the tokens that have high probability with or without the document, *e.g.*, helping verbs after the word 'this'. Masking helps in focusing only on those tokens which are originally ranked high by the model, thereby circumventing this issue. PMI–D_{EQ} generates better relevance scores for two out of three datasets while taking a hit on the

[4] https://huggingface.co/facebook/bart-large
[5] https://github.com/McGill-NLP/FaithDial

Document	**What happens if I had a lapse of insurance coverage during active duty?** You are required to maintain continuous coverage or surrender your plates before deployed. If you were unable to do so , DMV has a procedure in place to exempt you from administrative actions upon your return. You must provide a notarized statement that the vehicle was not used during the time in question, a copy of your military ID, proof of current insurance or surrender of plates, and signed and dated copies of deployment and return papers, or DD-214 if separated from service...
Context	**User**: lost my insurance while on active duty and have some questions **Agent**: You are required to maintain continuous coverage or surrender your plates before being deployed. Were you unable to keep your insurance while on active duty? **User**: yes, it just wasn't on my mind and I don't get notices and anything like that while deployed
Greedy	**Agent**: Unfortunately, no relevant information is found.
PMI:	**Agent**: DMV has a procedure in place to exempt you from administrative actions upon your return. You must provide a notarized statement that the vehicle was not used during the time in question, a copy of your military ID , proof of current insurance or surrender of plates, and signed and dated copies of deployment and return papers , or DD-214 if separated from service.

Table 5: An example from the test set of Multi-Doc2Dial dataset where Greedy generates a '*I don't know*' response and PMI–DECODE generates a relevant and faithful response.

faithfulness metrics. PMI–D results in the best tradeoff between faithfulness and relevance, thus we use it for human evaluation. Though PMI–D generates more faithful responses, its relevance scores seem to be lower than the greedy approach. Data analysis reveals that this is mostly due to inherent solution multiplicity (Nandwani et al., 2021) in most of the dialogues, where more than one correct responses are possible, but we report relevance *w.r.t.* only the provided gold response. This is further confirmed by human evaluation. Finally, table 5 presents an example from MD2D where likelihood based decoding returns a '*no info. found*', a response which is present for around 1800 training samples. The same model returns the correct response when PMI–DECODE is used instead of only likelihood for decoding, demonstrating the capability of PMI–DECODE to shift the score (or probability distribution) in favour of the words present in the document.

Human Evaluation: We perform human evaluation experiments to compare (1) *relevance* - the ability to generate responses that are contextually correct given the dialog history, (2) *faithfulness* - the ability to generate responses that are faithful to the given document, and (3) *grammar* – ability to generate grammatically correct and fluent responses. All three dimensions were categorically labeled as agree, neutral, or disagree. We sampled 100 random (document, dialog history, response) tuples, 50 each from MD2D and TC. We evaluate the responses generated by two decoding techniques: greedy and PMI–DECODE. Out of six in–house annotators used (3 per dataset), four were experts in dialog research and two were beginners.

The results are summarized in Table 4. For each dimension, we report the percentage of responses that were rated *agree*. As expected, PMI–DECODE generates more faithful responses compared to greedy. We observe a 15% improvement in faithfulness compared to greedy decode on both datasets. Further, PMI–DECODE improves relevance on MD2D but slightly deteriorates on TC.

As PMI–DECODE maximises not just the likelihood of responses, but a combination of likelihood and faithfulness, we expected the responses to contain grammatical errors compared to greedy decode. To counter this issue, we proposed to use a weighted combination of likelihood and faithfulness during decode, with a higher weight on likelihood. We also restricted the vocabulary during each decode step to just the *top-p* subset. The human study shows that these mitigation techniques helped in reducing the grammatical mistakes made by PMI–DECODE. We see that the grammar is only slightly inferior to greedy on both the datasets.

We use Fleiss Kappa (Fleiss and Cohen, 1973) to measure the inter-annotator agreement, which is substantial for relevance (0.63) and faithfulness (0.63), and almost perfect (0.88) for grammar.

6 Conclusion

In this paper, we present a novel metric, PMI–FAITH, to measure quantify faithfulness of responses generated by document grounded dialog systems. It uses conditional PMI between the response and the document given the dialog history to quantify faithfulness. We extend the idea of PMI–FAITHto propose a novel decoding strategy, PMI–DECODE which encourages responses to be faithful to the given document by maximizing both the likelihood and faithfulness of the decoded response. Our experiments on the BEGIN benchmark prove that our proposed metric better correlates with human judgments compared to existing metrics. On three document-grounded dialog datasets, our novel decoding strategy generates more faithful responses than greedy decoding, as measured using automated metrics and a human study.

Limitations

Though our decoding technique generates more faithful responses, we observed its inability to respond to generic chit–chat or pleasantries, like 'Hello!' or 'Good–bye'. It is possible to combine it with other techniques, like training with CTRL tokens (Rashkin et al., 2021b), which can enable it to generate both generic as well as faithful responses depending upon the dialogue context. But identifying when to generate a particular kind of response may require more insights and we leave this overall thread for future work. Next, to compute CPMI, we need to pass **d**, **h**, and **h** separately to the decoder. Though it can be done in parallel, but it may still reduce the throughput of the overall system by half. Finally, as demonstrated by the human evaluation, PMI–DECODE at times generates grammatically incorrect responses, even though the pre–trained language models are very good at generating fluent and coherent English. While we presented two knobs: α and *top* p masking to overcome this, we believe there could be other ways of handling this.

Ethics Statement

Our work does not introduce any new ethical concerns per se, other than the ones already faced by large language models. Our decoding strategy works on top of any trained language model and generates the text which is more faithful to a given input document. This can act as a double–edged sword: on one hand, if the document itself contains profanity, it may enhance the model's likelihood of generating similar content. But on the other hand, providing a valid document may also reduce the inherent likelihood of the model to generate profane content. Therefore, we recommend using it with responsibility and caution.

References

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6251–6258. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.

Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. 2020. Logic2text: High-fidelity natural language generation from logical forms. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2096–2111. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022a. FaithDial: A Faithful Benchmark for Information-Seeking Dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022b. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Trans. Assoc. Comput. Linguistics*, 10:1066–1083.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. Multidoc2dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6162–6176. Association for Computational Linguistics.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 864–870. Association for Computational Linguistics.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613 – 619.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 166–175. ACM.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, DialDoc@ACL 2022, Dublin, Ireland, May 26, 2022*, pages 161–175. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. \$qˆ2\$: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7856–7870. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *CoRR*, abs/2202.03629.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

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

Yatin Nandwani, Deepanshu Jindal, Mausam, and Parag Singla. 2021. Neural learning of one-of-many solutions for combinatorial problems in structured output spaces. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W. Black, and Ruslan Salakhutdinov. 2021. Focused attention improves document-grounded generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4274–4287. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021a. Measuring attribution in natural language generation models. *CoRR*, abs/2112.12870.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021b. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 704–718. Association for Computational Linguistics.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *CoRR*, abs/2110.05456.

Teven Le Scao, Angela Fan, Christopher Akiki, and et al. 2022. Bloom: A 176b-parameter open-access multilingual language model.

David Schlangen. 2003. A coherence-based approach to the interpretation of non-sentential utterances in dialogue.

Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. 2021. Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 1598–1608. ACM.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, and Bill Dolan. 2021. A controllable model of grounded response generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14085–14093. AAAI Press.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2734–2744. Association for Computational Linguistics.

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

Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 708–713. Association for Computational Linguistics.

A Human Evaluation

The screenshot of a sample task is shown in Figure 3. Out of six in–house annotators used, four were experts in dialog research and two were beginners. All the annotators provided their consent over an appropriate official communication channel, *e.g.*, official email or slack channels. The following were the instructions provided to the human evaluators.

What is the task? There are 50 incomplete dialogs along with a document over which the dialog is grounded on. For each (document, incomplete dialog) pair we provide the next response predicted by 2 different dialog systems (shuffled in random order). You are requested to judge the response generated by these 2 systems along three dimensions: faithfulness, relevance and grammar. Each dimension has to be evaluated using the following scale: Agree (A), Neutral (N), and Disagree (D).

How to judge relevance? Relevance measures how apt is the response given the dialog context and the knowledge. Please select *agree* when the response is apt and does not convey any incorrect information. Select *neutral* when it is hard to decide whether it is right or wrong and *disagree* otherwise.

How to judge faithfulness? The faithfulness of a response is only dependent on the grounding document and it is independent of the dialog. A system response can be marked disagree for relevance and still be marked agree for faithfulness. Please select *agree* when the complete response can be inferred from the document. Select *neutral* when it is hard to decide whether it can be inferred from the document or not and *disagree* when a major portion of the response cannot be inferred from the document. For the case where the response is something like "No information is present". The judgement should be *agree* if there is no information about that in the document provided and "disagree" if there is information available in the document, but the system didn't pick it up. For cases where the user initiates a chit-chat (say the user says "hi, how are you"), the agent responds with chit-chat ("I am doing good"), please can mark faithfulness as *neutral*.

How to judge grammar? The grammar score for a response is independent of the dialog or the document. A system response can be marked as disagree for relevance and still be marked agree for grammar. Please select *agree* when the response looks like how an expert human writes. Select *neutral* when there is a major issue with how the response

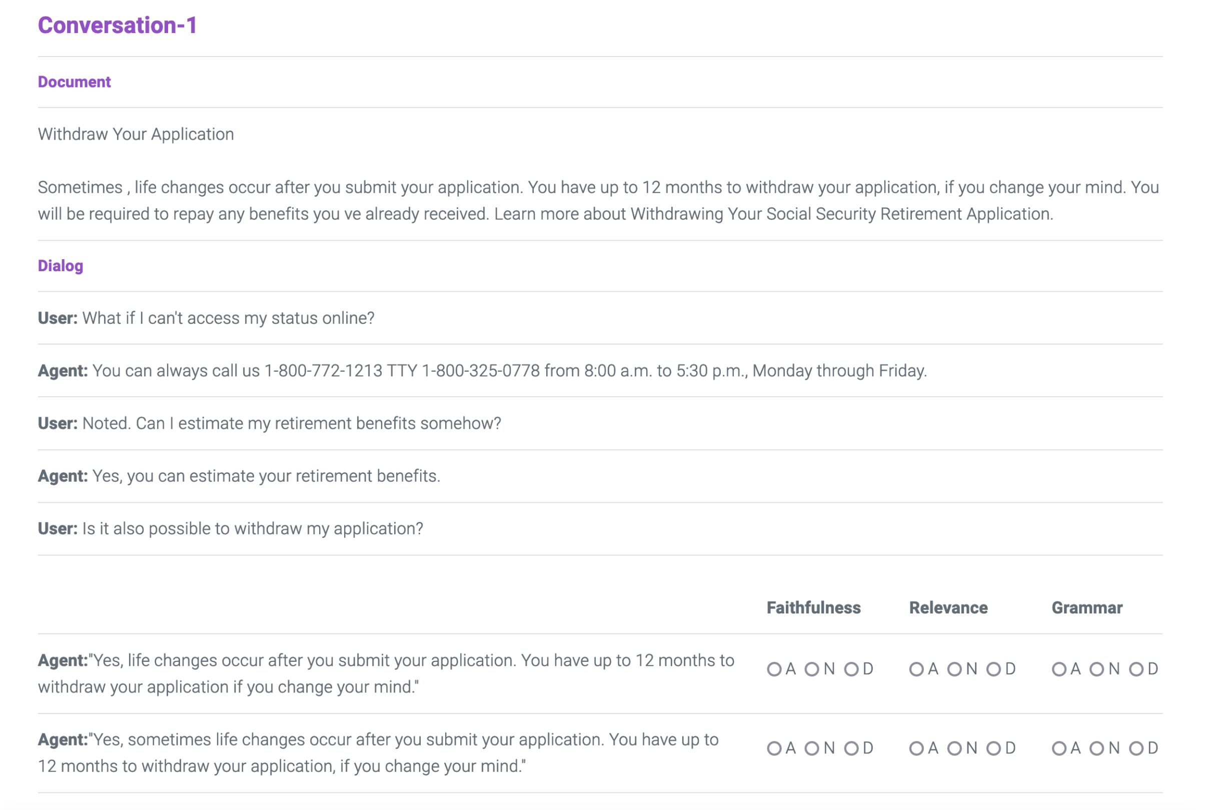

Figure 3: An example of the task provided to the human judges.

reads but it still understandable and *disagree* when the response makes no sense.