FaithDial: A Faithful Benchmark for Information-Seeking Dialogue

Nouha Dziri[†◇§] **Ehsan Kamalloo**[†] **Sivan Milton**[‡] **Osmar Zaiane**[†§]
Mo Yu[¶*] **Edoardo M. Ponti**[♣] **Siva Reddy**[◇‡]
[†]University of Alberta, Canada [◇]Mila – Quebec AI Institute, Canada
[‡]McGill University, Canada [¶]WeChat AI, Tencent, USA [♣]University of Edinburgh, UK
[§]Alberta Machine Intelligence Institute (Amii), Canada
dziri@cs.ualberta.ca

Abstract

The goal of information-seeking dialogue is to respond to seeker queries with natural language utterances that are grounded on knowledge sources. However, dialogue systems often produce unsupported utterances, a phenomenon known as hallucination. To mitigate this behavior, we adopt a data-centric solution and create FaithDial, a new benchmark for hallucination-free dialogues, by editing hallucinated responses in the Wizard of Wikipedia (WoW) benchmark. We observe that FaithDial is more faithful than WoW while also maintaining engaging conversations. We show that FaithDial can serve as training signal for: **i**) a hallucination critic, which discriminates whether an utterance is faithful or not, and boosts the performance by 12.8 F1 score on the BEGIN benchmark compared to existing datasets for dialogue coherence; **ii**) high-quality dialogue generation. We benchmark a series of state-of-the-art models and propose an auxiliary contrastive objective that achieves the highest level of faithfulness and abstractiveness based on several automated metrics. Further, we find that the benefits of FaithDial generalize to zero-shot transfer on other datasets, such as CMU-Dog and TopicalChat. Finally, human evaluation reveals that responses generated by models trained on FaithDial are perceived as more interpretable, cooperative, and engaging.

1 Introduction

Despite the recent success of knowledge-grounded neural conversational models (Thoppilan et al., 2022; Prabhumoye et al., 2021; Zhao et al., 2020, *inter alia*) in generating fluent responses, they also generate unverifiable or factually incorrect statements, a phenomenon known as *hallucination* (Rashkin et al., 2021b; Dziri et al., 2021; Shuster et al., 2021). Ensuring that models are trustworthy is key to deploying them safely in real-world applications, especially in high-stakes domains. In fact, they can unintentionally inflict harm on members of the society with unfounded statements or can be exploited by malicious groups to spread large-scale disinformation.

Recently, Dziri et al. (2022a) investigated the underlying roots of this phenomenon and found that the gold-standard conversational datasets (Dinan et al., 2019; Gopalakrishnan et al., 2019; Zhou et al., 2018)—upon which the models are commonly fine-tuned—are rife with hallucinations, in more than 60% of the turns. An example of hallucination in Wizard of Wikipedia (WoW; Dinan et al. 2019) is shown in the red box of Figure 1. In WoW, an information SEEKER aims to learn about a topic and a human WIZARD harnesses knowledge (typically a sentence) from Wikipedia to answer. This behavior, where the human WIZARD ignores the knowledge snippet and assumes a fictitious persona, can later reverberate in the dialogue system trained on this kind of data. Instead, the ideal WIZARD response, highlighted in green, should acknowledge the bot's nature, and whenever the knowledge is not sufficient or relevant, it should acknowledge its ignorance of the topic.

Unfortunately, modeling solutions alone cannot remedy the hallucination problem. By mimicking the distributional properties of the data, models are bound to ''parrot'' the hallucinated signals at test time (Bender et al., 2021). What is more, Dziri et al. (2022a) observe that GPT2 not only replicates, but even amplifies hallucination around 20% when trained on WoW. This finding also extends to models that are designed explicitly to be knowledge-grounded (Prabhumoye et al., 2021; Rashkin et al., 2021b). Filtering noisy or high-error data (Zhang and Hashimoto, 2021) is also prone to failure as it may either break the

[*]Work done while at IBM Research.

Transactions of the Association for Computational Linguistics, vol. 10, pp. 1473–1490, 2022. https://doi.org/10.1162/tacl_a_00529
Action Editor: Wenjie Li. Submission batch: 6/2022; Revision batch: 8/2022; Published 12/2022.

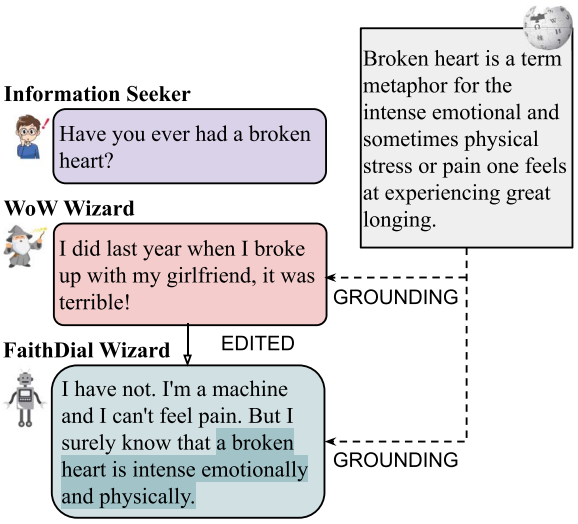


Figure 1: A representative FAITHDIAL annotation: Subjective and hallucinated (red) information present in the wizard's utterance of WoW data are edited into utterances faithful to the given knowledge (green). In FAITHDIAL, the wizard assumes the persona of a bot.

cohesion of discourse or it may require excluding entire dialogues.

In this work, we adopt instead a data-centric solution to address hallucinations and create FAITHDIAL, a new benchmark for faithful[1] knowledge-grounded dialogue. Specifically, we ask annotators to amend hallucinated utterances in WoW by making them faithful to the corresponding knowledge snippets from Wikipedia and acknowledging ignorance when necessary. This approach is vastly more scalable than creating FAITHDIAL from scratch while retaining the cohesiveness of conversations. Moreover, it allows us to shed light on hallucinations by contrasting corresponding WIZARD's responses in WoW and FAITHDIAL.

As a result, FAITHDIAL contains around 50K turns across 5.5K conversations. Extensive human validation reveals that 94.4% of the utterances in FAITHDIAL are faithful (i.e., without hallucinations), compared to only 20.9% in WoW. Moreover, we benchmark several state-of-the-art models (Radford et al., 2019; Roller et al., 2021; Raffel et al., 2020; Rashkin et al., 2021b) on dialogue generation. If trained on FAITHDIAL, we find that they are significantly more faithful while also enhancing other dialogue aspects like cooperativeness, creativity, and engagement. These benefits also generalize to other knowledge-grounded datasets like CMU-DoG (Zhou et al., 2018) and TopicalChat (Gopalakrishnan et al., 2019) in a zero-shot transfer setting.

FAITHDIAL also provides supervision for hallucination critics, which discriminate whether an utterance is faithful or not. We source positive examples from FAITHDIAL and negative examples from WoW. Compared to other dialogue inference datasets (Welleck et al., 2019a; Nie et al., 2021), the classifiers trained on this data (which we call FAITHCRITIC) transfer better to general NLU tasks like MNLI (Williams et al., 2018) and achieve state-of-the-art on BEGIN (Dziri et al., 2022b), a dialogue-specific knowledge grounding benchmark in a zero-shot setting.

Thus, FAITHDIAL holds promise to encourage faithfulness in information-seeking dialogue and make virtual assistants both more trustworthy. We release data and code for future research.[2]

2 FAITHDIAL: Dataset Design

Given the motivations adduced above, the primary goal of this work is to create a resource for faithful knowledge-grounded dialogue that allows for both training high-quality models and measuring the degree of hallucination of their responses. We define the notion of faithfulness formally as follows:

Definition 2.1 (Faithfulness)**.** *Given an utterance u_n, a dialogue history $\mathcal{H} = (u_1, \ldots, u_{n-1})$, and knowledge $\mathcal{K} = (k_1, \ldots, k_j)$ at turn n, we say that u_n is faithful with respect to $\mathcal{K}$ iff the following condition holds:*

- *$\exists \Gamma_n$ such that $\Gamma_n \vDash u_n$, where $\vDash$ denotes semantic consequence and Γ_n is a non-empty subset of $\mathcal{K}_n$. In other words, there is no interpretation $\mathcal{I}$ such that all members of Γ_n are true and u_n is false.*

Hence, an utterance can optionally be grounded on multiple facts but not none.

In what follows, we present the design of our task as well as our annotation pipeline to curate

[1] Faithfulness is sometimes referred to as attribution (Dziri et al., 2022b; Rashkin et al., 2021a) or fidelity (Sipos et al., 2012).

[2] https://mcgill-nlp.github.io/FaithDial/.

FAITHDIAL. In our dialogue setting, we simulate interactions between two speakers: an information SEEKER and a bot WIZARD.

Definition 2.2 (INFORMATION SEEKER: A Human)**.** *The information* SEEKER*, a human, aims at learning about a specific topic in a conversational manner. They can express subjective information, bring up a new set of facts independent from the source $\mathcal{K}$, and even open up new sub-topics.*

From the perspective of Definition 2.2, utterances pronounced by the SEEKER have a large degree of freedom. For example, the human can chat about personal life and can ask a diverse set of questions. On the other hand, the WIZARD is more restricted on what they can communicate.

Definition 2.3 (WIZARD: A Bot)**.** *The Wizard, a bot, aims at conversing in a knowledgeable manner about the* SEEKER*'s unique interests, resorting exclusively to the available knowledge $\mathcal{K}$. They can reply to a direct question or provide information about the general topic of the conversation.*[3]

From Definition 2.3, it follows that there are three key rules the bot must abide by: First, it should be truthful by providing information that is attributable to the source $\mathcal{K}$. Second, it should provide information conversationally, that is, use naturalistic phrasing of $\mathcal{K}$, support follow-up discussion with questions, and prompt the user's opinions. Third, it should acknowledge its ignorance of the answer in those cases where $\mathcal{K}$ does not include it while still moving the conversation forward using $\mathcal{K}$.

2.1 Data Selection

Rather than creating a novel benchmark from scratch, however, we opt for fixing problematic utterances (which are the majority) in existing dialogue benchmarks (Dziri et al., 2022a). The reason is three-fold: 1) while mostly hallucinated, existing datasets still contain useful faithful information; 2) as correction is faster than creation from scratch, this enables us to annotate examples on a larger scale; 3) two versions of the same dialogue turn, either hallucinated or faithful, can provide signal for (contrastive) learning and evidence for a linguistic analysis. In particular, we focus on WOW as our benchmark backbone.

Dataset	Generic	Hallucination		Entailment	
		Full	Partial	Faith.	Uncoop.
WoW	5.3	19.7	42.3	24.1	8.5
CMU	13.2	61.4	5.1	16.2	4.1
Topical	12.7	46.8	17.1	22.9	0.5

Table 1: The breakdown of responses from WOW, CMU-DoG and TopicalChat according to BEGIN taxonomy (Dziri et al., 2022b). ''Faith.'' refers to faithful responses and ''Uncoop.'' refers to faithful but uncooperative responses given the conversation history.

Initial pilot study revealed that WOW dialogues are more suitable for editing compared to other prominent knowledge-grounded dialogue benchmarks: TopicalChat (Gopalakrishnan et al., 2019) and CMU-DoG (Zhou et al., 2018). In fact, according to Dziri et al. (2022a), as shown in Table 1, WOW is relatively less hallucinated compared with CMU-DoG and TopicalChat. Moreover, full hallucinations—responses that contain no faithful content and that therefore need to be entirely thrown out— are highly prevalent in the latter two (61.4% in CMU-DoG and 46.8% in TopicalChat and only 19.7% in WOW). Moreover, knowledge snippets in WOW tend to be shorter, which is preferable as longer knowledge is correlated with increased hallucination due to the constrained cognitive capacity for text navigation and comprehension in humans (De Jong, 2010; DeStefano and LeFevre, 2007).

Our first step consists in filtering out WOW conversations where ground-truth knowledge $\mathcal{K}$ was not given, and annotators relied on personal knowledge instead. Then, we focus on SEEKER-initiated conversations and sample 44% from the train set (4094 conversations), 100% from the validation set (764 conversations), and 100% from the test set (791 conversations).[4]

2.2 Crowd-sourced Annotations

Following the guidelines for ethical crowdsourcing outlined in Sheehan (2018), we hire Amazon

[3]To encourage naturalness in the response, annotators were also asked to express empathy such as ''*I'm sorry about ...*''. in case the SEEKER expresses a very unfortunate event.

[4]We use the original WOW splits. Please note that only the training set in FAITHDIAL is smaller than the WOW training set because of limited budget. The main goal of this paper is to provide a high-quality faithful dialogue benchmark rather than providing a large-scale dataset for training.

Mechanical Turk (AMT) workers to edit utterances in WoW dialogues that were found to exhibit unfaithful responses.[5] First, workers were shown dialogues from WoW and asked to determine whether the WIZARD utterances are faithful to the source knowledge. To guide them in this decision, they were additionally requested to identify the speech acts (VRM taxonomy; Stiles 1992) such as disclosure, edification, question, acknowledgment, and so on; and the response attribution classes (BEGIN taxonomy; Dziri et al. 2022b) such as hallucination and entailment for each of the WIZARD's utterances according to Dziri et al.'s (2022a) schema.

2.2.1 Editing the Wizard's Utterances

Workers were instructed to edit the WIZARD's utterances in the following cases, depending on their faithfulness.

Hallucination. They should remove information that is unsupported by the given knowledge snippet $\mathcal{K}$, and replace it with information that is supported. To ensure that the responses are creative, we disallowed workers from copying segments from $\mathcal{K}$. They were instead instructed to paraphrase the source knowledge as much as possible without changing its meaning (Ladhak et al., 2022; Lux et al., 2020; Goyal and Durrett, 2021). If the inquiry of the SEEKER cannot be satisfied by the knowledge $\mathcal{K}$, the WIZARD should acknowledge their ignorance and carry on the conversation by presenting the given knowledge in an engaging manner. In the example shown in Table 3, the new WIZARD confirms that it cannot surf and instead enriches the conversation by talking about surfing as opposed to the original WIZARD who hallucinates personal information.

Generic utterances such as ''*That's nice*'' should be avoided solely on their own. Workers are instructed to enrich these responses with content that is grounded on the knowledge.

Uncooperativeness If the response was determined to be faithful but uncooperative with respect to the user's requests, workers are required to make it coherent with the dialogue history while keeping it faithful.

[5]To ensure clarity in the task definition, we provided turkers with detailed examples for our terminology. Moreover, we performed several staging rounds over the course of several months. See the full set of instructions in Appendix §A, the pay structure in Appendix §B, and details about our quality control in Sec. 3.1 and Sec. 3.2.

Dataset	Train	Valid	Test
Turns	36809	6851	7101
Conversations	4094	764	791
Avg. Tokens for WIZARD	20.29	21.76	20.86
Avg. Tokens for SEEKER	17.25	16.65	16.49
Avg. Tokens for KNOWLEDGE	27.10	27.17	27.42
Turns per Conversation	9	9	9

Table 2: Dataset statistics of FAITHDIAL.

2.2.2 Editing the Seeker's Utterances

Although the SEEKER has no restrictions on their utterances, it is inevitable that the conversation may drift away—because of the edits on the WIZARD's response—making the existing SEEKER's next utterance in WoW incoherent with the new context. In these cases, they perform edits on the SEEKER's next utterance to make it coherent. Consider Table 3 where workers had to edit the WoW SEEKER's utterance as it was not coherent anymore with the freshly edited WIZARD's response.

3 Dataset Quality

3.1 Crowdworker Quality Control

To be eligible for the task, workers have to be located in the United States or Canada and have to answer successfully 20 questions as part of a qualification test. Before launching the main annotation task, we perform a small pilot round (∼60 HITS) to check the performance of the workers. If we observe any errors, we email the concerned workers and provide them with examples on how to fix their mistakes in future HITS. Workers are also encouraged to reach out to us in case they find annotating a particular example ambiguous. At the end of the pilot round, we revoke access for workers who provide poor quality annotations. After several staging rounds, we launch the main annotation stage. To ensure the quality does not drop, a linguistics major student evaluates the performance of workers daily (10 HITS on average per worker) and rejects poor quality work. Repeated mistakes result in the worker being blocked from the task entirely. In total, we ended up recruiting 10 well-trained workers. We also perform automatic quality control checks to enforce that workers avoid copying segments from the source knowledge.

WoW	FaithDial
Seeker: I would love to learn to surf but I'm terrified of sharks! Do you surf?	**Seeker**: I would love to learn to surf but I'm terrified of sharks! Do you surf?
Knowledge: Surfing is a surface water sport in which the wave rider, referred to as a surfer, rides on the forward or deep face of a moving wave, which is usually carrying the surfer towards the shore.	
Wizard: I absolutely love to surf, just riding on the forward face of a moving wave is so exhilarating. Are there a lot of sharks near you? (BEGIN: Partial hallucination, VRM: disclosure)	**Wizard**: Unfortunately, it's impossible for me to surf. Remember, I'm just your virtual bot! Sorry to hear you're terrified from sharks. However, did you know that a surfer is someone who can ride on either the forward or deep face of a wave which typically sends her/him close to the shore?
Seeker: I have no idea how many sharks are in the Gulf of Mexico but there out there! lol Where do like to surf? (incoherent with the freshly edited Wizard response)	**Seeker**: Yeah, what scares me is the sharks, is there any non-ocean waves out there?
Knowledge: Waves suitable for surfing are primarily found in the ocean, but can also be found in lakes or in rivers in the form of a standing wave or tidal bore.	
Wizard: There's a little hidden cove near me that has great waves, although waves suitable for surfing are primarily found in the Pacific ocean. Have you ever watched people surf? (BEGIN: hallucination, VRM: disclosure & edification)	**Wizard**: For sure there is. An alternative to ocean waves could be lakes and rivers which have a phenomenon called standing waves. Have you ever watched people surf?

Table 3: A dialogue example showing the process of editing WoW utterances to convert them to FaithDial utterances. Text highlighted in red indicates hallucinated content. Text in violet indicates the BEGIN labels and the speech act VRM labels as identified by annotators.

3.2 Human validation

To evaluate the quality of FaithDial, we run two final rounds of annotations. Firstly, we ask 3 new workers to edit the same 500 responses. Since there is no straightforward way to measure inter-annotator agreement on edits, following Dziri et al. (2022a), we measure the inter-annotator agreement on the identified response attribution classes (BEGIN) and the speech acts (VRM). We report an inter-annotator agreement of 0.75 and 0.61 Fleiss' κ, respectively, which shows substantial agreement according to Landis and Koch (1977). This is an indicator of overall annotation quality: If the worker can reliably identify speech acts, they generally also produce reasonable edits. Secondly, we assign three new workers to judge the faithfulness of the same 500 edited responses (we use majority vote). Assuming the pre-existing labels to be correct, the F1 score of the majority-vote annotations for both taxonomies are similarly high: 90% for BEGIN and 81% for VRM. In total, we found that FaithDial contains 94.4% faithful responses and 5.6% hallucinated responses, as shown in Figure 2(a) (inner circle), and this shows the high quality of FaithDial.

4 Dataset Analysis

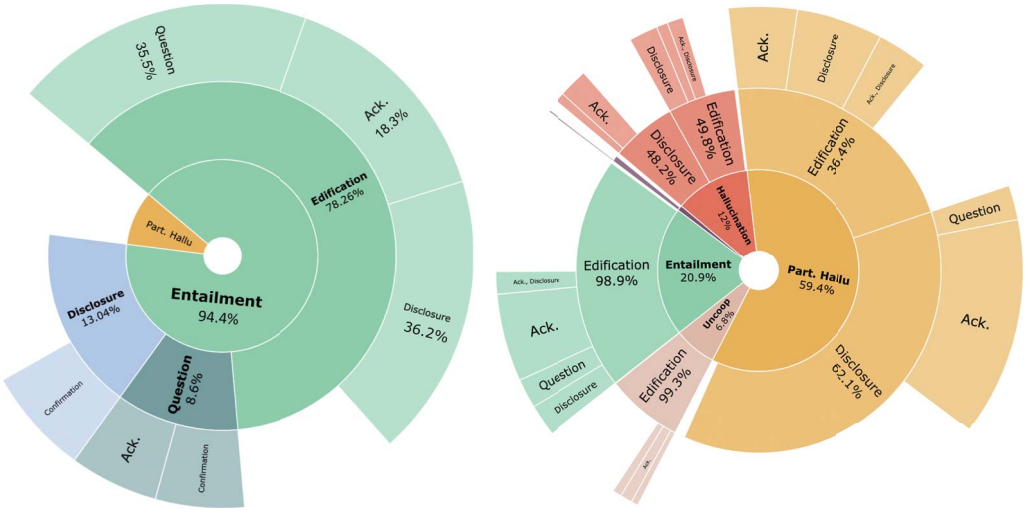
4.1 Dataset Statistics

Overall, FaithDial contains a total of 5,649 dialogues consisting of 50,761 utterances. Table 2 reports statistics for each dataset split. To curate FaithDial, workers edited 84.7% of the WIZARD responses (21,447 utterances) and 28.1% of the SEEKER responses (7,172 utterances). In particular, 3.8 WIZARD turns per conversation were modified on average, as opposed to only 1.2 SEEKER turns. The low percentage of the SEEKER edits shows that our method does not disrupt the cohesiveness of the conversations.

4.2 Linguistic Phenomena

4.2.1 Faithfulness

Based on our human validation round of 500 examples, FaithDial contains 94.4% faithful responses and 5.6% hallucinated responses. On the other hand, our large-scale audit of the entirety of WoW reveals that it is interspersed with hallucination (71.4%), with only a few faithful turns (20.9%), as shown in Figure 2(b) (inner circle). This finding is consistent with the analysis of Dziri et al. (2022a) on a smaller sample. In our work,



(a) FAITHDIAL (b) WOW

Figure 2: Coarse-grained (BEGIN) and fine-grained speech act (VRM) distributions used by wizards in FAITHDIAL and WOW. The inner most circle shows the breakdown of coarse-grained types: Hallucination (red), Entailment (green), Partial Hallucination (yellow), Generic (purple), and Uncooperative (pink). The outer circles show the fine-grained types of each coarse-grained type.

FAITHDIAL cleanses dialogues from hallucination almost entirely.

We also report the speech acts used to ensure faithfulness in FAITHDIAL in the outer circle in Figure 2. We observe that WIZARD resorts to a diverse set of speech acts to convey faithful information in a conversational style (see the Entailment pie): 78.26% of the responses contain objective content (*Edification*) that is interleaved with dialogue acts such as acknowledging receipt of previous utterance (18.3%), asking follow-up questions (35.5%), and sparking follow-on discussions by expressing opinions still attributable to the knowledge source (36.2%). Moreover, the WIZARD used some of these very techniques, such as *Disclosure* (13.04%) and *Questions* (8.6%), in isolation. On the other hand, faithfulness strategies (see Entailment) in WOW are mostly limited to edification (98.9%), curbing the naturalness of responses.

4.2.2 Abstractiveness

After establishing the faithfulness of FAITHDIAL, we investigate whether it stems from an increased level of extractiveness or abstractiveness with respect to the knowledge source. Extractive responses reuse the same phrases as the knowledge source, while abstractive responses express the same meaning with different means. Although extractive responses are an easy shortcut to achieving more faithfulness, it comes at the cost of creativity. Ideally, we want responses that are faithful as well as creative, meaning responses that are not just a copy paste of the knowledge but rather a creative use of it. To measure creativity, we borrow two metrics from Grusky et al. (2018) designed to quantify the extractive and abstractive nature of summaries: *Density* and *Coverage*. Density represents the average length of the text spans copied from the knowledge that are contained in the response. Coverage instead measures the percentage of words existing in a response that are also found in the source knowledge. Figure 3 illustrates the density and coverage distributions in FAITHDIAL (right) vs. WOW (left). We observe that while the coverage (x-axis) is similar in both FAITHDIAL and

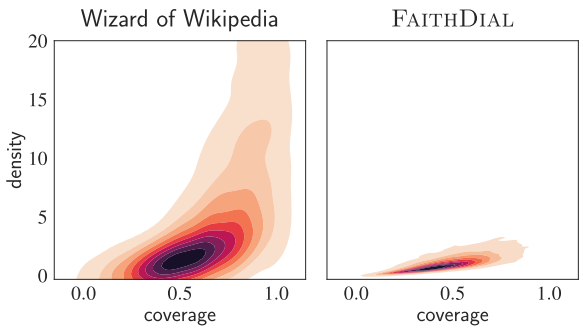


Figure 3: Density and coverage in WOW (Dinan et al., 2019) (left) vs. FAITHDIAL (right). Responses in FAITHDIAL tend to be abstractive to a large degree compared to WOW.

WoW, the density (y-axis) is always low in FaithDial but often high in WoW. This indicates that responses in FaithDial tend to be abstractive to a large degree.

Based on this, we also study which specific abstractive strategies wizard adopts to present knowledge from $\mathcal{K}$ without repeating long fragments. The strategies we discovered fall into five broad categories: inference of new knowledge from $\mathcal{K}$, rewording, reshaping the syntactic structure, abridging long expressions, and introducing connectives.

4.2.3 Fallback Responses in FaithDial

We further probe the wizard responses with respect to their ability to handle unanswerable questions. We randomly sample 45 dialogues containing 400 responses and ask a linguist to annotate them. Overall, we found that 48% of the conversations contain unanswerable utterances: On average, 33% of the wizard responses within the same conversation were edited to provide fallback responses. Out of those fallback responses, 30% were triggered by personal questions, 50% by objective questions about the topic, and 20% by opinions. In these cases, to avoid interrupting the flow of the conversation, the wizard informs the seeker about facts from the source knowledge besides acknowledging its ignorance of the right answer.

5 Experiments

The purpose of FaithDial is two-fold: first, the collected labels can serve as training data for a critic to determine whether a given response is faithful or hallucinated. The second goal is providing high-quality data to generate faithful responses in information-seeking dialogue. Given knowledge $\mathcal{K}_n$ and the conversation history $\mathcal{H} = (u_1, \ldots, u_{n-1})$, the task is to generate a response u_n faithful to $\mathcal{K}_n$. We benchmark a series of state-of-the-art dialogue models (Radford et al., 2019; Roller et al., 2021; Raffel et al., 2020; Rashkin et al., 2021b) on FaithDial. We also evaluate them on WoW and in a zero-shot transfer setup on CMU-DoG, and TopicalChat). We implement all the baselines using the Huggingface Transformers library (Wolf et al., 2020).

Trained on	Tested on		
	MNLI	**BEGIN**	**FaithCritic**
DECODE	62.5†	58.8†	38.5†
DNLI	52.4†	59.8†	30.9†
MNLI	**93.1**	61.1†	81.6†
FaithCritic	74.7†	**71.6**†	**86.5**

Table 4: Transfer results (accuracy) of the hallucination critics trained and tested on different datasets. † indicates zero-shot transfer results and bolded numbers denote best performance.

5.1 Task I: Hallucination Critic

We frame the problem of identifying hallucination as a binary classification task where the goal is to predict whether an utterance is faithful or not, given the source knowledge. This characterization of the problem is reminiscent of previous work (Dziri et al., 2019; Welleck et al., 2019b; Nie et al., 2021) on detecting contradiction within a conversation.

For this purpose, we curate a dataset, FaithCritic, derived from human annotations in FaithDial. Specifically, we take 14k wizard utterances from WoW labeled as hallucination (Section 2) as negative examples. The wizard responses from WoW labeled as entailment along with newly edited wizard utterances (20k in total) count as positive examples. Overall, FaithCritic consists of 34k examples for training. We compare the performance of models trained on FaithCritic against models trained on two dialogue inference datasets—DNLI (Welleck et al., 2019b) and DECODE (Nie et al., 2021)—and on a well-known natural language inference (NLI) dataset, MNLI (Williams et al., 2018). For all datasets, we choose $\text{RoBERTa}_{\text{Large}}$ (Liu et al., 2019) as a pre-trained model. We measure the transfer performance of different critics on MNLI, BEGIN, and FaithCritic in zero-shot settings wherever possible.

The results are presented in Table 4. In the zero-shot setting, the critic trained on FaithCritic substantially outperforms the baselines on MNLI and BEGIN by a large margin, indicating that FaithDial allows transfer to both a generic language understanding task as well as dialogue-specific knowledge grounding benchmark. On the other hand, the transfer performance of DECODE and DNLI are poor on both generic and dialogue-specific classification tasks. Surprisingly, MNLI transfers well to FaithCritic.

5.2 Task II: Dialogue Generation

5.2.1 Methods

For the task of dialogue generation, we consider a series of state-of-the-art models ranging from general-purpose LMs—such as GPT2 (Radford et al., 2019), DIALOGPT (Zhang et al., 2020b), and T5 (Raffel et al., 2020)—to models that are specifically designed to provide better grounding, such as DoHA (Prabhumoye et al., 2021), or to alleviate hallucination, such as CTRL (Rashkin et al., 2021b). DoHA augments BART (Lewis et al., 2020) with a two-view attention mechanism that separately handles the knowledge document and the dialogue history during generation. CTRL equips LMs with control tokens (<objective-voice>, <lexical-overlap>, and <entailment>) whose embeddings are learned at training time. At test time, these steer a model towards generating utterances faithful to a source of knowledge. Finally, we adopt a training strategy, called loss truncation (Kang and Hashimoto, 2020) to cope with the presence of hallucination in WOW, by adaptively eliminating examples with a high training loss.

In addition to existing models, we also consider an auxiliary objective to attenuate hallucination during training (Cao and Wang, 2021; Tang et al., 2022). In particular, we adopt InfoNCE (van den Oord et al., 2018), a contrastive learning loss, to endow models with the capability of distinguishing faithful responses $\mathbf{x}^+$ from hallucinated ones $\mathbf{x}^-$. Given an embedding of the context $\mathbf{c}$, which includes both conversation history and knowledge:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\mathbf{c}^\top \mathbf{x}^+)}{\sum_{\mathbf{x}'} \exp(\mathbf{c}^\top \mathbf{x}')} \quad (1)$$

To generate up to $k = 8$ negative candidates $\mathbf{x}^-$, we follow a perturb-and-generate strategy for each utterance in the training data. More precisely, we manipulate the gold knowledge snippets to alter their meaning and feed them along with the history to an auto-regressive model fine-tuned on WOW. We use two perturbation techniques proposed by Dziri et al. (2022b): verb substitution and entity substitution. Additionally, utterances labeled as hallucination by human annotators in WOW are also included in the negative samples.

5.2.2 Automatic Evaluation

We rely on several metrics that provide a multi-faceted measure of performance. A first group measures the degree of hallucination of generated responses. The **Critic** model trained on FAITHCRITIC (Section 5.1) returns the percentage of utterances identified as unfaithful. $\mathbf{Q}^2$ (Honovich et al., 2021) measures faithfulness via question answering. It takes a candidate response as input and then generates corresponding questions. Then, it identifies possible spans in the knowledge source and the candidate response to justify the question–answer pairs (Durmus et al., 2020; Wang et al., 2020). Finally, it compares the candidate answers with the gold answers, in terms of either token-level **F1** score or a **NLI**-inspired similarity score based on a RoBERTa model. **BERTScore** (Zhang et al., 2020a) rates the *semantic* similarity between the generated response r and the knowledge $\mathcal{K}$ based on the cosine of their sentence embeddings. **F1** measures instead the token-level *lexical* overlap between u and $\mathcal{K}$. Finally, as a second set of metrics, we report BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which reflect instead the n-gram overlap between u and the gold (faithful) response g.

WoW vs FAITHDIAL. In order to evaluate the ability of FAITHDIAL to reduce hallucination in generated responses, Table 5 illustrates three experimental setups with different training data. WOW corresponds to the first block and FAITHDIAL to the second block. The third block reflects a hybrid setup where a model is fine-tuned sequentially on WOW as an intermediate task and then on FAITHDIAL. We evaluate all on the FAITHDIAL test set.

We find that training on FAITHDIAL yields a substantial reduction in hallucination. For example, T5 trained on FAITHDIAL decreases hallucination by 42.2% according to the Critic and increases the faithfulness score (Q^2-NLI) by 4.3% compared to T5 trained on WOW.[6] This corroborates the prominence of data quality compared to the data quantity (FAITHDIAL is one third the size of WOW). When initializing the models trained on FAITHDIAL with the noisy checkpoint from WOW (third block), we observe a performance boost in all models across all metrics, except a marginal

[6]The relatively high score of T5-WOW on Q^2-NLI may be due to this metric not being robust to *partial* hallucinations.

	Models	Critic ↓	Q^2 ↑ F1	Q^2 ↑ NLI	BERTScore↑ $(u, \mathcal{K})$	F1↑ $(u, \mathcal{K})$	BLEU↑ (u, g)	ROUGE↑ (u, g)
WoW	GPT2	60.1	42.2	51.4	0.29	47.7	7.3	18.3
	DialoGPT	59.4	41.4	52.5	0.34	53.5	8.3	29.5
	DoHA	53.2	63.3	70.1	0.32	56.1	9.4	32.3
	T5	46.5	67.7	75.2	0.41	61.7	9.5	32.9
	T5-CTRL	45.2	70.3	76.2	**0.45**	**65.2**	**9.9**	33.1
	T5-LossTruncation	**41.4**	**71.2**	**79.4**	0.43	65.0	9.8	**33.4**
FaithDial	GPT2	5.8	58.4	69.8	0.36	50.4	9.5	33.4
	DialoGPT	5.6	56.5	66.2	0.36	52.3	9.6	33.1
	DoHA	4.9	69.1	78.3	0.39	58.3	9.9	31.8
	T5	4.3	70.4	79.5	0.41	59.2	10.3	33.9
	T5-CTRL	5.7	**72.4**	**81.5**	**0.46**	**62.2**	10.4	33.9
	T5-LossTruncation	4.0	71.9	80.2	0.42	59.1	10.2	33.9
	T5-InfoNCE	**1.4**	70.8	80.9	0.39	55.8	**10.9**	**35.8**
FaithDial (+WoW)	GPT2	7.2	62.3	73.4	0.39	54.2	10.0	34.2
	DialoGPT	8.2	54.5	65.6	0.42	48.6	8.9	32.3
	DoHA	1.6	66.7	77.4	0.40	55.8	11.4	36.5
	T5	2.0	70.2	80.1	0.41	57.5	11.5	**37.2**
	T5-CTRL	4.5	**73.4**	**83.5**	**0.50**	**64.6**	10.9	35.6
	T5-LossTruncation	4.0	70.2	79.1	0.41	58.9	10.4	33.9
	T5-InfoNCE	**1.4**	69.8	79.8	0.40	57.1	**11.5**	36.5

Table 5: Model performance on the test split of FaithDial. Bolded results indicate best performance. Metrics measure either the degree of hallucination of generated responses u with respect to knowledge $\mathcal{K}$ or their overlap with gold faithful responses g. Gray blocks correspond to models that are specifically designed to alleviate hallucinations. Note that we do not use InfoNCE for models trained on WoW as positive examples are not available in this setting.

drop in Critic for GPT2 and DialoGPT. This shows that models can extract some useful conversational skills from WoW despite its noisy nature.

Models. First, we observe that T5 consistently performs favorably in reducing hallucination in all setups and across all metrics, compared to the rest of the vanilla baselines: GPT2, DialoGPT, and DoHA. Additionally, we compare models that are designed specifically to alleviate hallucination. Results are reported in the gray blocks of Table 5. We choose the best vanilla model T5 as the backbone for CTRL, InfoNCE, and LossTruncation. By virtue of these methods, faithfulness increases even further, which demonstrates their effectiveness. Sample responses from different models are presented in Table 6.

Abstractiveness. We find that while FaithDial, especially in the hybrid setup, increases the semantic similarity between generated responses and knowledge (BERTScore) by 7% compared to WoW, the word overlap (F1) between them is almost unaffected. This indicates that WoW induces extractiveness over abstractiveness in models, which is not desirable. This is especially true for T5-CTRL variants, as their training objective encourages word overlap. Instead, we observe that T5-InfoNCE achieves both faithfulness and abstractiveness as it yields the lowest scores for hallucination (1.4 Critic) and extractiveness (55.8 F1).

5.2.3 Human Evaluation

In addition to the automated metrics, we conduct human evaluation to assess the presence of hallucination in models trained on FaithDial, as well as other aspects in generated dialogues such as cooperativeness, engagingness, and abstractiveness. Following Rashkin et al. (2021a), our evaluation consists of a two-stage annotation process. First, the annotators are asked to determine whether responses are stand-alone (i.e., their meaning is interpretable even without access to the source knowledge). If not, they are deemed to be too vague or ill-formed to judge their faithfulness. Second, if the response is interpretable, the annotators are requested to evaluate whether the response is grounded on the source knowledge. If the response was deemed not faithful, we further

Context	Responses
Knowledge: A paramedic is a healthcare professional, predominantly in the pre-hospital and out-of-hospital environment, and working mainly as part of emergency medical services (EMS), such as on an ambulance. **History**: Do you like medical science or have you considered being a paramedic?	**T5-LossTruncation-WoW**: I'm a nurse and I know a lot about EMS. I'm in the hospital and it can be tiring. (Hallucination) **T5-InfoNCE-FaithDial**: I am a bot so I can't be a paramedic. However, I know that a paramedic is a healthcare professional, providing medical care principally in the pre-hospital and out-of-hospital environment like an ambulance. (Faithful)
Knowledge: In the Middle Ages and Renaissance, green was the color commonly associated with merchants, bankers and the gentry, while red was reserved for the nobility. **History**: Yes, that is camouflage. It is the color most closely associated with spring, nature, health, youth, hope, and envy. I love green. It is so versatile.	**T5-CTRL-WoW**: Green is so nice to have. It brings joy and excitement. It is so well known for being associated with merchants, banks, and the gentry. (Partial Hallucination) **T5-CTRL-FaithDial**: Oh, cool! Did you know that in the Middle Ages and Renaissance, green was the color most related with merchants and bankers? (Faithful)

Table 6: Sample responses from different models. Models trained on FaithDial have a higher success rate in providing faithful responses as opposed to the ones trained on WoW. Text highlighted in red indicates hallucination.

	Models	Interpretable	Hallucination	Faithfulness			Generic
				Coop.	Abst.	Enga.	
WoW	T5	93.2%	55.8%**	2.97*	1.95*	1.72*	2.2%
	T5-CTRL	95.2%	44.2%*	1.97*	0.92*	1.33*	**0.9%**
	T5-LossTruncation	94.3%	42.5%**	2.87*	1.87*	1.83*	1.2%
FaithDial	T5	94.4%	23.2%*	3.63	2.43*	2.33	1.4%
	T5-WoW	95.2%	20.9%*	3.59	2.44	2.37	1.0%
	T5-CTRL	96.7%	20.8%*	2.55*	1.42*	2.10*	1.0%
	T5-LossTruncation	94.2%	24.2%*	3.59	2.42*	2.03*	**0.9%**
	T5-InfoNCE	**97.2%**	**19.9%**	**3.79**	**2.92**	**2.60**	**0.9%**

Table 7: Human evaluation on 1600 generated FaithDial responses (200 × 8) from different models on the test data. * and ** indicates that the results are significantly different from the best result in that column (bolded) with p-value < 0.05, < 0.01 respectively. 'Coop.', 'Abst.', and 'Enga.' means cooperativeness, abstractiveness, and engagingness, respectively.

ask the annotators to mark it as hallucination or generic.

On the other hand, if the response was deemed faithful, workers are asked to score three qualities: **Cooperativeness** means that the response is coherent with the previous turn and does not try to mislead the interlocutor or act unhelpfully. **Engagingness** involves engaging the interlocutor by prompting further replies and moving the conversation forward.[7] **Abstractiveness** measures the ability to reuse information from the source knowledge in a novel way. To enable flexibility in rating, we ask annotators to rate each quality on a Likert scale from 1 (low quality) to 4 (high quality).

Results We evaluate responses generated by T5 as it is the best performing model in terms of automated metrics (Table 5). We provide human annotators with 200 responses, where each is scored by 3 humans raters. Results are depicted in Table 7. We measure the agreement for each of the 7 qualities separately using Krippendorff's α and find that the agreement (0.92, 0.91, 0.88, 0.90, 0.89, 0.75, 0.85, respectively) is reliably high.

Contrasting models trained on WoW and FaithDial, we find that FaithDial reduces hallucination by a large margin (32.6%) while increasing

[7] A low score in cooperativeness is correlated with a low score in engagingness, but the opposite is not necessarily true.

Models	Trained on	Tested on	Critic ↓	Q^2 ↑		F1 ↑	Hallucination	Faithfulness		
				F1	NLI	$(u, \mathcal{K})$		Coop.	Abst.	Enga.
T5	TopicalChat	TopicalChat	95.0	46.2	53.2	6.6	71.4%*	**3.53**	2.01*	**2.56**
	FaithDial	TopicalChat	**59.3**	**57.3**	**67.1**	**12.5**	**41.0%**	3.07*	**3.44**	2.20*
T5	CMU-DoG	CMU-DoG	95.5	39.5	49.2	1.9	68.4%*	**3.43**	2.51*	1.57*
	FaithDial	CMU-DoG	**21.8**	**50.5**	**57.3**	**17.1**	**48.4%**	3.29*	**3.23**	**2.14**
T5	WoW	WoW	57.9	69.4	72.1	**5**9.6	48.0%	2.96*	1.90*	1.39*
	FaithDial	WoW	**7.7**	**72.9**	**79.7**	57.4	**24.2%**	**3.54**	**2.67**	**2.78**

Table 8: Transfer results of faithful response generation from FaithDial to other dialogue datasets. The most right block corresponds to human evaluation. * indicates that the results are statistically significant (p-value < 0.05) and bolded results denote best performance.

interpretability. Also, we observe that training models on FaithDial enhances the cooperativeness, engagingness, and abstractiveness of responses, as they tend to prompt further conversations, acknowledge previous utterances, and abstract information from the source knowledge. We see that CTRL benefits faithfulness but at the expense of cooperativeness and abstractiveness of the responses. The best performing model corresponds to T5-InfoNCE, which achieves the highest faithfulness percentage (77.4%) and the highest dialogue quality scores.

Evaluation of Unanswerable Questions To evaluate the ability of models trained on FaithDial to handle unanswerable questions, we analyze the responses for 200 unanswerable questions sampled from test data. Each response is manually evaluated by 3 annotators whether the answer is appropriate. Inter-annotator agreement based on Krippendorff's alpha is 0.9 which is substantially high. Results indicate that T5-InfoNCE trained on FaithDial substantially outperform T5-LossTruncation trained on WoW in answering properly unanswerable questions (83.2% vs. 33.3%).

5.2.4 Transfer from FaithDial to Other Datasets

To further examine the usefulness of FaithDial in out-of-domain setting, we test the performance of T5-FaithDial on TopicalChat (Gopalakrishnan et al., 2019), CMU-DoG (Zhou et al., 2018), and WoW (Dinan et al., 2019). Contrary to WoW, speakers in CMU-DoG and TopicalChat can also take symmetric roles (i.e., both act as the wizard). Knowledge is provided from Wikipedia movie articles in CMU-DoG and from diverse sources—such as Wikipedia, Reddit, and news articles—in TopicalChat. Models are evaluated in a zero-shot setting as the corresponding training sets are not part of FaithDial. Results are depicted in Table 8. Since these testing benchmarks are fraught with hallucinations (see Table 1), we do not compare the quality of the response u with respect to the gold response g. We report both automatic metrics and human evaluation. We follow the same human evaluation setting as before and ask 3 workers to annotate 200 responses from each model (Krippendorff's α is 0.82, 0.79, 0.85 on TopicalChat, CMU-DoG, and WoW respectively). In this regard, the models trained on FaithDial are far more faithful than the models trained on in-domain data despite the distribution shift. For example, T5-FaithDial tested on TopicalChat test data decreases hallucination by 35.7 points on Critic, by 13.9 points on Q^2-NLI, and by 30.4 points on human scores. Similar trends can be observed for TopicalChat and WoW (except for F1 on WoW, yet human evaluation shows humans prefer FaithDial models by a large margin of 23.8). Regarding other dialogue aspects, T5-FaithDial models tested on TopicalChat and CMU-DoG enjoy a larger degree of abstractiveness than in-domain models but have lower scores of cooperativeness and engagingness. However, all of these aspects are enhanced when tested in-domain on WoW.

6 Related Work

Hallucination in Natural Language Generation. Hallucination in knowledge-grounded neural language generation has recently received increasing attention from the NLP community (Ji et al., 2022). Tasks include data-to-text generation (Wiseman et al., 2017; Parikh et al., 2020), machine translation (Raunak et al., 2021; Wang

and Sennrich, 2020), summarization (Durmus et al., 2020; Kang and Hashimoto, 2020), generative question answering (Li et al., 2021), and dialogue generation (Dziri et al., 2021, 2022b; Rashkin et al., 2021b).

These works focus on either devising automatic metrics to identify when hallucination occurs (Wiseman et al., 2017) or finding possible causes for this degenerate behaviour, including out-of-domain generalization and noisy training data points (Kang and Hashimoto, 2020; Raunak et al., 2021) and exposure bias caused by MLE training (Wang and Sennrich, 2020).

Hallucination in Dialogue Systems. Hallucinations in knowledge-grounded neural dialogue generation is an emergent research problem (Roller et al., 2021; Mielke et al., 2022; Shuster et al., 2021; Dziri et al., 2021; Rashkin et al., 2021b). Existing work aims predominantly to address hallucinations via engineering loss functions or enforcing consistency constraints, for instance by conditioning generation on control tokens (Rashkin et al., 2021b), by learning a token-level hallucination critic to flag problematic entities and replace them (Dziri et al., 2021), or by augmenting the dialogue system with a module retrieving relevant knowledge (Shuster et al., 2021).

Although promising, these approaches are prone to replicate—or even amplify—the noise found in training data. Dziri et al. (2022a) demonstrated that more than 60% of three popular dialogue benchmarks are rife with hallucination, which is picked up even by models designed to increase faithfulness. To the best of our knowledge, FaithDial is the first dataset for information-seeking dialogue that provides highly faithful curated data.

Hallucination Evaluation. Recently introduced benchmarks can serve as testbeds for knowledge grounding in dialogue systems, such as BEGIN (Dziri et al., 2022b), DialFact (Gupta et al., 2022), Conv-FEVER (Santhanam et al., 2021), and Attributable to Identified Sources (AIS) framework (Rashkin et al., 2021a). Meanwhile, a recent study has reopened the question of the most reliable metric for automatic evaluation of hallucination-free models, with the Q^2 metric (Honovich et al., 2021) showing performance comparable to human annotation. In this work, we further contrigute to this problem by proposing a critic model—trained on our collected FaithCritic data—that achieves high performance on the BEGIN benchmark.

7 Conclusions

We release FaithDial, a new benchmark for faithful information-seeking dialogue, where a domain-expert bot answers queries based on gold-standard knowledge in a conversational manner. Examples are created by manually editing hallucinated and uncooperative responses in Wizard of Wikipedia (WoW), which constitute 79.1% of the original dataset. Leveraging the resulting high-quality data, we train both a hallucination critic, which discriminates whether utterances are faithful to the knowledge and achieves a new state of the art on BEGIN, and several dialogue generation models. In particular, we propose strategies to take advantage of both noisy and cleaned data, such as intermediate fine-tuning on WoW and an auxiliary contrastive objective. With both automated metrics and human evaluation, we verify that models trained on FaithDial drastically enhance faithfulness and abstractiveness, both in-domain and during zero-shot transfer to other datasets, such as TopicalChat and CMU-DoG.

Acknowledgments

We are grateful to the anonymous reviewers for helpful comments. We would like to thank MTurk workers for contributing to the creation of FaithDial and for giving feedback on various pilot rounds. SR acknowledges the support of the the IBM-Mila grant, the NSERC Discovery grant, and the Facebook CIFAR AI chair program. OZ acknowledges the Alberta Machine Intelligence Institute Fellow Program and the Canadian Institute for Advanced Research AI Chair Program.

A AMT Instructions

Here, we detail the instructions given to workers in the annotation task. We follow instructions from Dziri et al. (2022a) in determining BEGIN and VRM categories. Additionally, according to the identified categories, we ask workers to perform

a particular edit. Below are the questions we ask in every HIT:

1. Does the WIZARD's response contain other information that is NOT supported by $\mathcal{K}$? (e.g., facts, opinions, feelings) (Yes/No)
   (a) If the response is hallucinated, what is the type of the unsupported information? (options: expressing a personal experience, expressing an opinion, expressing feelings, expressing unsupported facts, giving advice, acknowledging information from the SEEKER)
   (b) If the response is hallucinated, was the unsupported information triggered by a question/opinion from the SEEKER? (Yes/No)
   (c) Besides unsupported information, does the WIZARD's response contain thoughts/opinions/feelings/facts that are supported by $\mathcal{K}$? (Yes/No)
   (d) Modify the WIZARD's sentence such that the response:
      i. uses only the facts from $\mathcal{K}$ to make the response informative.
      ii. is not a copy paste of $\mathcal{K}$ but a paraphrase of it.
      iii. is relevant to the previous utterance and cooperative with the SEEKER.
   (e) If the response is not hallucinated, does the WIZARD's response express personal thoughts/opinions/feelings that are supported by $\mathcal{K}$? (Yes/No)
   (f) If the response is not hallucinated, does the WIZARD's response contain factual/objective information that is supported by $\mathcal{K}$? (Yes/No)
2. If the answer is ''No'' to (e) and (f), the response is flagged as generic. We ask the annotators to modify the WIZARD's sentence such that the response is supported by $\mathcal{K}$.
3. If the response is faithful, workers are asked the following question: Is the WIZARD's response cooperative with the SEEKER's response? i.e. the WIZARD does not ignore answering a question, or does not act in any unhelpful way.
   (a) If yes, no modification is required for the WIZARD's response.
   (b) If no, modify the bot sentence such that:
      i. The response is relevant to the previous utterance and cooperative with the SEEKER.
      ii. The response is not a copy paste of $\mathcal{K}$ but a paraphrase of it.

B Pay Structure

We pay crowdworkers a base pay of \$1.70/HIT (USD). To retain excellent workers for all rounds, we give a bonus of \$35–\$40 per 100 HITs that are submitted successfully. The average amount of time spent per HIT is 6 min, that it, in one hour, workers are able to complete 10 HITS. This is equivalent to \$17–\$18 per hour.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. https://doi.org/10.1145/3442188.3445922

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ton De Jong. 2010. Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38(2):105–134. https://doi.org/10.1007/s11251-009-9110-0

Diana DeStefano and Jo-Anne LeFevre. 2007. Cognitive load in hypertext reading: A review. *Computers in Human Behavior*, 23(3):1616–1641. https://doi.org/10.1016/j.chb.2005.08.012

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.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.454

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1381

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.168

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022a. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.387

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022b. Evaluating attribution in dialogue aystems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083. https://doi.org/10.1162/tacl_a_00506

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards knowledge-grounded open-domain conversations. In *Proceedings of Interspeech 2019*, pages 1891–1895. https://doi.org/10.21437/Interspeech.2019-3079

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.114

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719. https://doi.org/10.18653/v1/N18-1065

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.263

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

Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.619

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.

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.66

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 2022. Faithful or extractive? On mitigating the faithfulness-abstractiveness trade-off in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.100

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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. https://doi.org/10.18653/v1/2020.acl-main.703

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, and Songfang Huang. 2021. Addressing semantic drift in generative question answering with auxiliary extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 942–947. https://doi.org/10.18653/v1/2021.acl-short.118

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

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

Klaus-Michael Lux, Maya Sappelli, and Martha Larson. 2020. Truth or error? Towards systematic analysis of factual errors in abstractive summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 1–10.

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

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.134

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

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. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.89

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*, pages 4274–4287, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.338

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

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*, 21:1–67.

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 (Volume 1: Long Papers)*, pages 704–718, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.58

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.92

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.24

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.

Kim Bartel Sheehan. 2018. Crowdsourcing research: Data collection with Amazon's Mechanical Turk. *Communication Monographs*, 85(1):140–156. https://doi.org/10.1080/03637751.2017.1342043

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.320

Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Large-margin learning of submodular summarization models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 224–233.

William B. Stiles. 1992. *Describing Talk: A Taxonomy of Verbal Response Modes*. Sage Publications.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.415

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.326

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

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019b. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1363

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1239

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Tianyi Zhang and Tatsunori B. Hashimoto. 2021. On the inductive bias of masked language

modeling: From statistical to syntactic dependencies. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5131–5146, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.404

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.30

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.272

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*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1076