Evaluating Groundedness in Dialogue Systems: The BEGIN Benchmark

Nouha Dziri[*†♢] **Hannah Rashkin**[*§] **Tal Linzen**[†♣] **David Reitter**[§]
[♢]University of Alberta [§]Google Research [♣]New York University
{hrashkin,reitter}@google.com
dziri@cs.ualberta.ca
linzen@nyu.edu

Abstract

Knowledge-grounded dialogue agents are systems designed to conduct a conversation based on externally provided background information, such as a Wikipedia page. Such dialogue agents, especially those based on neural network language models, often produce responses that sound fluent but are not justified by the background information. Progress towards addressing this problem requires developing automatic evaluation metrics that can quantify the extent to which responses are grounded in background information. To facilitate evaluation of such metrics, we introduce the Benchmark for Evaluation of Grounded INteraction (BEGIN). BEGIN consists of 8113 dialogue turns generated by language-model-based dialogue systems, accompanied by humans annotations specifying the relationship between the system's response and the background information. These annotations are based on an extension of the natural language inference paradigm. We use the benchmark to demonstrate the effectiveness of adversarially generated data for improving an evaluation metric based on existing natural language inference datasets.

1 Introduction

Neural network language models (Vaswani et al., 2017; Radford et al., 2019) have been increasingly adopted as a central part of open-domain dialogue systems (Wolf et al., 2019; Zhang et al., 2020; Roller et al., 2020; Adiwardana et al., 2020). Utterances sampled from such language models sound natural, as reflected in these systems' high scores in human evaluations focused on measures such as "engagingness" or "human-likeness". While fluent, however, the responses generated by these systems are often only locally coherent or contain confabulated statements (see the red portions of the response in Figure 1 for illustration).

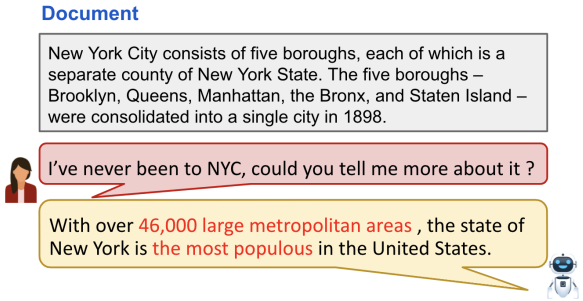

Figure 1: An example of a hallucinated response generated by the GPT2 language model fine-tuned on the Wizard of Wikipedia dataset (Dinan et al., 2019). The response was conditioned on a passage of text from Wikipedia and the previous utterance. The response (in yellow) is on topic and appears to be plausible, but the information it contains is not supported by the document.

In this work, we introduce a new classification task and benchmark for evaluating *knowledge-grounded dialogue systems*, systems that are expected to conduct a conversation based on a particular source of information, with the goal of making unstructured information more accessible to a user. It is critical for such a system to avoid producing utterances that appear to convey information but in reality are not supported by the document or even contradict it. The system should also avoid responses that fail to respond to the user's question, because they are accurate but off-topic (e.g., *"The University of Michigan is located in Ann Arbor"*), or because they are excessively general (e.g., *"I don't know much about NYC"* in Figure 1).

Existing automatic evaluation metrics for dialog, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and MAUDE (Sinha et al., 2020), are ill-suited to detecting these issues, and correlate poorly with human judgments (Liu et al., 2016; Dziri et al., 2019b; Sai et al., 2019). To facilitate progress towards reliable evaluation metrics for grounded dialog, we propose a new classifi-

[*]Equal Contribution.
[†]Work done while at Google Research.

cation task extending the Natural Language Inference (NLI) paradigm (Dagan et al., 2005; Bowman et al., 2015; Williams et al., 2018). NLI seeks to determine, given a premise p and a hypothesis h, whether p entails h, contradicts it, or is neutral with respect to it. For our taxonomy, we adopt the *entailment* and *contradiction* labels from the NLI paradigm, but split the *neutral* label into three sub-categories: *hallucination* responses, which are topical but include unverifiable information; *off-topic* responses; and *generic* responses, which are too vague to be verified. As a testbed for evaluation metrics based on this taxonomy, we create the Benchmark for Evaluation of Grounded INteraction (BEGIN), a dataset of 8113 dialogue responses generated by language models fine-tuned on the Wizard of Wikipedia dataset (Dinan et al., 2019), and ask annotators to categorize these responses using the proposed taxonomy.

We establish baseline performance on this benchmark using two pretrained transformer models, BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020). We fine-tune them on two existing NLI datasets: MNLI (Williams et al., 2018) and DNLI (Welleck et al., 2019). As these NLI datasets only support the coarse-grained distinction between *entailment*, *contradiction* and *neutral*, we additionally use perturbation techniques to generate examples for each of the five categories of our taxonomy, and fine-tune the pretrained models on this extended dataset. We find that there is considerable room for improvement between the best model we train (slightly above 70%) and human performance estimated from inter-annotator agreement on the benchmark (around 90% agreement with majority), suggesting that there are opportunities for developing stronger metrics for grounded dialog evaluation.

The main contributions of this work are:

(a) We propose a taxonomy of responses generated by a knowledge-grounded conversation system; this taxonomy extends the Natural Language Inference framework.

(b) We present a new benchmark, BEGIN, consisting of knowledge-grounded dialogue system responses annotated according to this taxonomy.

(c) We establish baseline performance on BEGIN using BERT and T5 fine-tuned on standard NLI datasets, and improve upon it using our own adversarially-created dataset.

2 Constructing the BEGIN Dataset

To generate dialogue responses for the BEGIN dataset, we fine-tuned two dialogue agents on an existing knowledge-grounded dialogue dataset. Based on error patterns commonly produced by these dialogue agents, we created a taxonomy of grounded dialog responses. We then constructed BEGIN by asking human annotators to categorize a sample of dialogue system responses according to this taxonomy.

2.1 Training Knowledge-Grounded Agents

Dataset We fine-tune our knowledge-grounded dialogue models on the Wizard of Wikipedia dataset (WoW; Dinan et al. 2019). WoW consists of crowdsourced English dialogues between a "Wizard" and an "Apprentice", where the goal of the Wizard is to convey to the apprentice information about a particular topic. The Apprentice, in turn, is expected to seek information about the topic. At each turn of the conversation, the Wizard is presented with passages from Wikipedia and chooses an evidence span (typically a sentence) to use as supporting evidence in their response. Not all utterances in WoW are grounded in external evidence: unlike the Wizard, the Apprentice is not presented with Wikipedia sentences when produce an utterance; and the Wizard is allowed to produce an utterance that does not use the evidence.

Task Formulation We trained our models to generate the Wizard's response based on a concatenation of two inputs: an evidence span (the Wikipedia sentence presented to the Wizard) and the previous dialogue turn, produced by the Apprentice. We omitted (evidence span, previous turn, response) triples in which the Wizard did not explicitly select a passage as evidence for the response. We used 82722 triples for training, 8800 triples for development, and 8690 triples for test.

Models We fine-tuned the base GPT-2 model (Radford et al., 2019) and the base version of T5 (Raffel et al., 2020) on the WoW dataset. We filtered out responses that the Google Perspective API deemed to have a greater than 50% likelihood of containing toxic language.

For GPT-2, we framed response generation as language modeling (cf. DialogGPT, Zhang et al. 2020): we concatenated the evidence, the previous turn and the response, and continued training GPT-2 using cross-entropy over the response tokens. We

used the same hyper-parameters as in Wolf et al. (2019). We refer to this model as GPT2-WOW.

For T5, which is an encoder-decoder model, the encoder was provided with the concatenation of the evidence span and the previous turn, and the decoder was trained to predict each token in the response given the previous tokens. We used the hyperparameters found using grid search on the dev set. We refer to this model as T5-WOW. Neither of these fine-tuned models is intended to advance the state of the art in grounded dialogue generation; rather, we used them to investigate the errors made by typical neural dialogue systems.

2.2 Response Taxonomy for Grounded dialogue Systems

A manual inspection of 200 of the responses generated by GPT2-WOW revealed that, in addition to responses that were entailed by the document as desired, there were five common error types. This section defines and illustrates each of the six response types using examples from the manually inspected sample of responses.

Entailment Entailed responses are those that are faithfully supported by the evidence span:

> **Evidence**: *Robert Norman Ross* (October 29, 1942 – July 4, 1995) was an *American painter*, art instructor, and *television host*.
>
> **Previous turn**: I would like to know more about *Bob Ross*.
>
> **Response**: Yes, absolutely! He was an *American painter* and *television host*.

Hallucination Hallucinated responses are topical, but state information that cannot be verified using the source document, e.g.:

> **Evidence**: *The Tucson metro* area is served by many local television stations and is the 68^{th} largest designated market area (DMA) in the U.S. with 433,310 homes (0.39% of the total U.S).
>
> **Previous turn**: Have you ever heard about the *Tuscon metro* area?
>
> **Response**: Yes, one of the most visited *US* cities is *the Tucson metro* which was largely influenced by the *television* culture.

We also include in the hallucination category responses that convey personal or subjective opinions, experiences, feelings, and so on, as such statements cannot be verified using the given evidence; some examples are *I love to take walks in the park*, *I've been to the hospital before*, or *I think Justin Trudeau won the elections in 2015*. While such responses may be acceptable in human dialogues, in the context of a knowledge-grounded dialogue system we consider them to be hallucinations.

Generic These are neutral sentences that are general enough to fit into a large number of possible contexts. Even if they are ostensibly about the same topic as the document, they are vague and do not provide information (true or not), e.g.:

> **Evidence**: A *ski* is a narrow strip of semi-rigid material worn underfoot to glide over snow.
>
> **Previous turn**: *Skiing* is really a lot of fun. Nothing like finding some fresh powder.
>
> **Response**: It seems you might like *skiing* in the winter.

Off-topic In contrast to hallucinations, which are semantically close to the evidence but add small amounts of unverifiable information, off-topic responses are unrelated to the evidence, e.g.:

> **Evidence**: *Toyota* is the world's market leader in sales of hybrid electric vehicles, and one of the largest companies to encourage the mass - market adoption of hybrid vehicles across the globe.
>
> **Previous turn**: Hi. I have never had a *Toyota* but I know a lot of people like them.
>
> **Response**: The era of *manufactured fibers* began with the development of rayon in France in the 1890s..

Contradiction These are informative sentences that directly contradict the evidence, e.g.:

> **Evidence**: Although most *law schools* only offer the traditional *three - year program*, several U.S. law schools offer an Accelerated JD program.
>
> **Previous turn**: Because I did not want to go through all the schooling? Do you know how long *law school* is?
>
> **Response**: Most *law schools* only offer a *10 year program* .

Uncooperative responses Human dialogues are characterized by cooperative responses that follow the conversational expectations set out by Grice's maxims (Grice, 1989): typical speakers do not intentionally mislead the listener, change the topic, or act unhelpfully in any other way. Dialogue system responses may be inappropriate if they are not cooperative in this sense, even when they are entailed by the evidence document; for example:

Evidence: Starbucks was founded in Seattle, Washington in 1971.

Previous turn: I love eating at Starbucks, have you been there?

Response: The original location of Starbucks is in Seattle, which was founded in 1971.

Ultimately, dialogue evaluation metrics should take cooperativeness into account, and as such we asked our raters to score the cooperativeness of dialogue responses. For the time being, however, we limit the scope of our experiments below to evaluating faithfulness to the document, and leave modeling cooperativeness to future work.

2.3 Annotation Procedure

Rather than asking raters to explicitly classify responses based on the proposed taxonomy, we broke down the task into hierarchical questions with Likert scales (from 1 to 5). We summarize this procedure below, and provide the exact questions in Appendix A. Responses often consist of multiple sentences; because each sentence may display different degrees of faithfulness, we asked annotators to rate each sentence in the response separately.

First, we asked annotators to judge whether the response was about the same topic as the evidence, and if not, whether it was best described as generic or off-topic. We additionally asked them if the response was cooperative. In the next stage, we asked the raters if the response was objective or contained personal and subjective opinions. If the rater judged that the response was objective, we asked them if in their judgment the response was intended to provide information, about the evidence in the document or anything else. If the answer to the last question was affirmative, we presented them with two follow-up questions: first, we asked them if the response was fully supported by the evidence; and second, we asked if any part of the response contradicted the evidence.

We collected annotations for 8113 dialogue responses, which we split into a development (10% of the examples) and test (90% of examples) set; we release the data at https://github.com/google/BEGIN-dataset. Examples were randomly divided into dev. and test set partitions in such a way that examples using the same input context would only appear in the same partition. We did not create a training set to discourage the development of evaluation metrics that overfit to the specific features of BEGIN.

In post-processing, we converted the numerical ratings assigned by the annotators to one of our category labels using the procedure described in Appendix B. We note that the categories are not always mutually exclusive. For example, in a conversation about bees, the response *They have pretty big personalities* would be both off-topic and generic. The appropriate label may also depend on linguistic ambiguity that cannot be resolved from the context given to the annotators. In one example, the response *Oppenheimer as he is known as I think in neonatal med / ophthalmology* was generated about a document that says *He was the Director of Pediatric Neurosurgery at Johns Hopkins Hospital in Maryland from 1984 until his retirement in 2013*. Because the pronoun *he* in the document doesn't resolve to an antecedent, it is hard to determine whether this utterance is better described as a hallucination (attributing a medical specialty not mentioned in the document) or off-topic (this document was probably not about Oppenheimer at all), which is what the rater ultimately selected. Based on the annotations, 78% of the generic responses and 71% of the off-topic responses in our development set may also contain hallucinated information, but we label these overlapping cases as generic or off-topic respectively since these broader issues often subsume the hallucination problems.

We include examples from the development set in Table 1 along with the label breakdown. We note that the labels are unevenly distributed. Hallucinations and generic comments make up two of the biggest categories of responses. By contrast, contradictions make up a small fraction of the distribution. This suggests that LM-based dialogue agents like GPT2-WoW and T5-WoW are more likely to add extra confabulated information rather than directly contradict the evidence.

To evaluate inter-annotator agreement, we obtained two additional annotations (from the same pool of raters) for approximately 15% of the responses. The average Krippendorf's alpha on the responses to the different questions was around 0.41. This denotes relatively low-to-moderate agreement. One factor that may impact the scoring is slight disagreements between Likert scores (such as the difference between a 4 and a 5) which were counted as "partial" agreements scaled by the distance between the responses. Further reducing agreement are cases where a disagreement in an earlier question propagated to the follow-up ques-

Category	Frequency	Example	
Contradiction	0.6%	Evidence:	Broccoli is often boiled or steamed but may be eaten raw.
		Previous turn:	It actually does. vegetable of course, like broccoli and spinach.
		Response:	Broccoli can be eaten raw but not boiled
Off-topic	6.1%	Evidence:	A programming language is a formal language that specifies a set of instructions that can be used to produce various kinds of output.
		Previous turn:	Is it programming?
		Response:	Yes and I think that is why the United States still has the highest population of any country
Generic	27.5%	Evidence:	Almost as soon as automobiles had been invented, races of various sorts were organised, with the first recorded as early as 1867.
		Previous turn:	That's so cool. I guess that's just our competitive spirit.
		Response:	That's what makes it such a interesting history to behold
Hallucination	32.2%	Evidence:	Generally, the heavier the ball, the more pins that will topple on two equivalent shots.
		Previous Turn:	Yea but sadly bowling just isn't as demanding and difficult a sport as football so its fallen off while football is as prominent as ever.
		Response:	When they first came out in the 70s, they used to have at least ten percent of the ball on two equivalent shots
Entailment	33.6%	Evidence:	Elvis Aaron Presley (January 8, 1935 – August 16, 1977) was an American singer, musician, and actor.
		Previous turn:	I like to listen to Elvis, the king of rock and roll.
		Response:	Elvis Presley was born in 1935, and his death was August 16, 1977

Table 1: Examples of each of the categories from the development set of BEGIN, and their frequency in the development set. All responses were generated from GPT-2 fine-tuned on Wizard of Wikipedia based on the evidence and the previous turn shown in the table. Labels were annotated by human raters.

tions (e.g. they answered differently to whether the response was intended to be informative and so one of them left blank the follow-up question on how supported the information is) but we counted agreement on each of these questions as if they were independent of each other. Another factor that may affect human performance is ambiguous cases or instances where response categories are not always mutually exclusive, as mentioned above.

3 Classifying Responses Using Existing NLI Datasets

We establish baseline performance on BEGIN using models based on BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020). We first fine-tune the models on existing NLI datasets. Since these datasets are only labelled for the traditional three-way NLI classification (entailment, contradiction and neutral), in these experiments we collapse the three sub-labels that correspond to *neutral* in our fine-grained classification. In Section 4, we train a classifier for the full 5-way categorization scheme by creating adversarial data.

The first dataset we use, MNLI (Williams et al., 2018), is a collection of 433k premise-hypothesis pairs, where three hypotheses, one for each label, were generated by crowdworkers based on a premise drawn from a corpus. The second is DNLI (Welleck et al., 2019), which consists of 343k pairs of dialogue utterances and "persona attributes" curated from the PersonaChat dataset (Zhang et al., 2018) (e.g., a persona-describing attribute like *I have two cats* may be contradicted by a conversation utterance saying *I don't have any pets*). Each persona sentence and each utterance were associated with human-labeled triples (subject, predicate, object). A number of approaches were used to form NLI examples by linking dialogue utterances and persona sentences. For example, each unique pair of sentences that shared the same triple were labeled as entailment and each pair of sentences that were from contradictory triples were labeled as contradiction.

4 Adversarially Augmented Training Set

The experiments described in Section 3 were based on existing NLI datasets, which support a coarse-grained three-way classification, but not the full five-way taxonomy we introduced in Section 2.2. In this section we introduce a strategy to automat-

Evidence	Previous Turn	Response	Method	Label
The island's geography comprises relatively low-lying mountains surrounding a central plain, with several navigable rivers extending inland.	yes! is there many mountains there?	The geography is mostly low lying mountains which surround a central plain and many rivers extending inland.	None	Entailment
A 2009 report found that the skateboarding market **is worth an estimated 4.8 billion** in annual revenue with 11.08 million active skateboarders in the world.	I have heard that skateboarding is fun	The skateboarding market is not estimated to be 4.8 billion dollars.	Response negation	Contradiction
It is located on the Cumberland River in the **north central** part of the state.	I would really like to visit that river someday.	It is also in the south peripheral part of the state	Antonym	Contradiction
Elvis Aaron Presley (January 8, 1935 – August 16, 1977) was an American singer, musician, and actor.	Elvis Presley was ahead of his time and really brought in a new era of music.	I don't know much about music.	GPT2-WoW (with low temperature)	Generic
The fia formula one world championship has been one of the premier forms of racing around the world since its inaugural season in 1950.	What is the formula 1 championship called?	He is right. i've been to one, but they have tvs set up. the cars go fast though because there is minimum weight that includes driver.	Random utterance	Off-topic

(a) Examples generated without perturbing the evidence (by perturbing the response, by generating a new response using GPT2-WoW, or by selecting existing responses from the dataset).

Original evidence	Perturbed evidence	GPT2-WoW response	Method	Label
The **Chihuahua** is the smallest breed of dog and is named after the state of **Chihuahua** in **Mexico**.	The Chihuahua is the smallest breed of dog and is named after the state of **Maine** in **Rochester**.	The Chihuahua is known to be very friendly and Maine's mayor love them.	Entity swapping	Hallucination
Baking chocolate, also called bitter chocolate, contains **cocoa solids** and cocoa butter in varying proportions, without any added sugars.	**Cocoa solids**, also called bitter chocolate, contains **baking chocolate** and cocoa butter in varying proportions, without any added sugars.	The consumption of coco solids has many benefits.	Subject-object inversion	Hallucination

(b) Examples generated by perturbing the evidence: the response is a hallucination with respect to the original evidence.

Table 2: Examples from the automatically-generated adversarial data. Bold text highlights segments in the source documents that are either supported, contradicted or neutral by the response. Red text highlights the justification for why the responses are annotated with a specific label.

ically create "silver" training data for a classifier that produces the full taxonomy. We avoid training on BEGIN, because, as we mentioned before, we see it as a test-only dataset; recent work has shown that neural networks can overfit to irrelevant features of the dataset when trained on one part of it and tested on another part. We generated a balanced dataset of 7900 (evidence, dialogue history, response) triples (i.e. each label constituted 20% of the data), using the procedures described for each target label in the remainder of this section. See Table 2 for examples of our silver data.

Entailment We use the original human generated responses, but to avoid opinions or subjective experiences, we subsample from the portion of examples where the response doesn't use first person pronouns (selected from a word list) and at least 25% of the words in the response are in the evidence (to avoid responses that are only tangentially related to the evidence).

Off-topic Off-topic responses are sampled from WoW responses that are based on other pieces of evidence. To avoid having off-topic responses that would be trivial to spot based on lexical cues, we sample from conversations that were about the same topic as the target conversation.

Generic Generic sentences are generated from GPT2-WoW with a low softmax temperature (0.4).

Hallucination We perturb evidence spans from the WoW test set and then feed them to GPT2-WoW; in general, this results in responses that could be considered hallucinations with respect to the original evidence (see Table 2b). We use three perturbation methods, each applied to a different evidence document. All of these perturbations substantially alter the truth of the sentence while keeping it on topic. First, we swap the subject and the object of the original evidence. Second, we replace up to two verbs in the sentences by verbs of the same tense. Finally, based on an error analysis that showed that most hallucination errors made by our dialogue system involved incorrect entities, we extract all mentioned entities from different dialogue examples using the SpaCy NER tagger (Honnibal and Montani, 2017), and replace up to two randomly chosen entities in the original evidence document with entities of the same type (e.g., Person, Location or Organization).

Contradiction Adversarially-generated contradiction examples include two types of cases. The first is negated sentences, created based on an English Resource Grammar (ERG) parse (Flickinger et al., 2014); for example, *The skateboarding market is estimated to be around eight billion dollars* was replaced with *The skateboarding market is **not** estimated to be around eight billion dollars*. In the second type of case, adjectives are replaced with their WordNet antonyms (Miller, 1998): *Ancient Greece was home to the **first** pentathlon* is replaced with *Ancient Greece was home to the **last** pentathlon that was documented*.

Unlike hallucination examples, where we perturb the document and then feed it to GPT2-WoW, the contradiction examples are generated by directly perturbing the human response from the WoW dataset: initial experiments indicated that GPT2-WoW is not sensitive to these perturbations when applied to the evidence, in contrast with its sensitivity to the more substantial perturbation we used to generate hallucination examples. This indicates that this dialogue systems responses are only grounded in the document to a fairly limited extent.

5 Experimental Set-up

Each classifier takes as input the context—the evidence and the previous conversation turn—concatenated with a separating delimiter and the response. For BERT, we train a three-way or five-way classifier over the output `[CLS]` token. For T5, we follow the MNLI set-up used in the paper that introduced T5 (Raffel et al., 2020): the string "premise: " is concatenated with the context, the string "hypothesis: " is concatenated with the response, and the concatenation of the two strings is then passed as input to T5.

In addition to separate experiments evaluating models fine-tuned on MNLI and models fine-tuned on our adversarially augmented training data, we also investigated the performance of models trained first on MNLI and then on our adversarial data.

All models were trained with a batch size of 32 over 3 epochs, using the Adam optimizer and a learning rate of 2×10^{-5}. We evaluated the classifiers' performance via accuracy and macro-averaged F1 (i.e. computing F1 on each category before averaging) on BEGIN.

		Development set				*Test set*			
		3-way		5-way		3-way		5-way	
Model	Training Data	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BERT	DNLI	50.7	33.9	-	-	51.1	34.6	-	-
	MNLI	67.2	48.3	-	-	66.8	48.2	-	-
	Adversarial	70.9	48.5	43.7	30.8	**69.7**	48.1	40.9	29.5
	MNLI+Advers	69.6	50.4	**46.4**	**33.5**	68.5	47.1	**43.8**	31.0
T5	DNLI	51.2	33.8	-	-	51.9	36.0	-	-
	MNLI	69.1	50.3	-	-	68.5	49.1	-	-
	Adversarial	69.3	50.5	45.2	32.4	68.3	49.1	42.2	30.6
	MNLI+Advers	**71.0**	**52.6**	45.5	33.5	69.3	**49.5**	43.2	**31.4**

Table 3: Accuracy and Macro-F1 scores on the development and test sets for the 3-way and 5-way classification tasks.

6 Results

Table 3 summarizes the results of our experiments. MNLI is clearly a better fit to this task than DNLI. Fine-tuning on the adversarial data on its own is fairly effective even though it is a significantly smaller resource than DNLI or MNLI. Finally, fine-tuning first on MNLI and then on the adversarial data produces higher accuracy than training on the adversarial data alone.

None of our models exceeds 71% accuracy in the three-way classification setting, or 46.5% accuracy in five-way classification. By comparing individual annotator ratings and the majority-voted label in the triple-annotated subset of BEGIN, we estimate that human accuracy is about 90% for 3-way classification and 75% for 5-way classification. While humans do not agree perfectly, there is still a lot of room for improvement between these models and human performance. Finally, we note that accuracy is similar across BERT and T5, despite the fact that T5 contains orders of magnitude more parameters and was pretrained on a much larger corpus. This suggests that scaling up the pretrained model may not be sufficient to make progress on this task using the training datasets we have explored.

6.1 Error Analysis

In Table 4, we show some examples of errors made by different BERT-based models. One possible source of mistakes is misleading lexical cues. In the first example, all three models label the response as entailed, which may be due to the high degree of lexical overlap with the evidence. However, the semantic meaning of the information in the response is a bit different from the evidence, which is why humans annotated it as a hallucination. In the second example, two of the BERT models misidentify the response as a contradiction, which may be due to the models relying too much on the negative word *didn't* in the response.

6.2 Label Confusion

Figure 2 compares the predictions of three fine-tuned versions of BERT to the gold labels. All three variants are poor at predicting contradiction. The model fine-tuned on MNLI over-predicted contradictions, while the model fine-tuned on the adversarial data has low recall for contradiction prediction. The model trained on both resources performed better at predicting contradictions, but still misclassified most of them as hallucinations. The off-topic examples are also difficult for the models to classify; while they are only rarely confused with contradiction or entailment, they were often confused with the other neutral subcategories (*hallucination* and *generic*).

All three models displayed a substantial degree of confusion between the *hallucination* and *entailment* labels, but not as much between *entailment* and *generic* or *off-topic* examples. This difficulty of distinguishing hallucinated and faithful information would be obscured in the 3-way NLI categorization scheme where *hallucination* is aggregated with *off-topic* and *generic*. In other words, our 5-way taxonomy makes it possible to identify this as a particular set of *neutral* examples that is more likely to be confused with entailment.

7 Related Work

NLI for dialogue system evaluation Welleck et al. (2019) propose using NLI to improve dialogue consistency. They create the dialogue NLI

Evidence:	Early skiers used one long pole or spear.	G: Hallucination
Previous turn:	That's amazing. i sometimes wonder why it's not as popular as football, basketball and the rest though. it's an amazing sport	M: Entailment A: Entailment
Response:	they were the first to use a long pole or spear	MA: Entailment
Evidence:	Historically, archery has been used for hunting and combat.	G: Hallucination
Previous turn:	Yes it does take skill. I did not know what I was doing and ended up burning or scratching the inside of my arm with the tail of the arrow! Im not sure if its something you can develop with practice or not	M: Contradiction A: Contradiction MA: Hallucination
Response:	You did'nt do archery while hunting or doing combat !	
Evidence:	Skiing can be a means of transport, a recreational activity or a competitive winter sport in which the participant uses skis to glide on snow.	G: Off-topic M: Neutral
Previous turn:	yea seems easy but it isnt	A: Hallucination
Response:	The easiest way to get in shape is to go down to the slopes of a volcano and practice snowboarding	MA: Hallucination
Evidence:	Cheerleading is an activity wherein the participants (referred to as "cheerleaders") cheer for their team as a form of encouragement.	G: Hallucination M: Neutral
Previous turn:	I did a little bit of cheerleading when i was younger. how about you?	A: Off-topic
Response:	No but i do have a lot of experience cheering for my team in the olympics.	MA: Off-topic
Evidence:	Skiing can be a means of transport, a recreational activity or a competitive winter sport in which the participant uses skis to glide on snow.	G: Generic M: Contradiction
Previous turn:	My brother is a professional skier.	A: Off-topic
Response	I have never tried skiing.	MA: Off-topic

Table 4: Example outputs of BERT-based models on the development set (G: Gold label; M: BERT fine-tuned on MNLI; A: BERT fine-tuned on the adversarial data; MA: BERT fine-tuned on MNLI, and then on the adversarial data). M was trained on the three-way classification task (*entailment*, *neutral*, *contradiction*) while A and MA were trained on the full five-way classification.

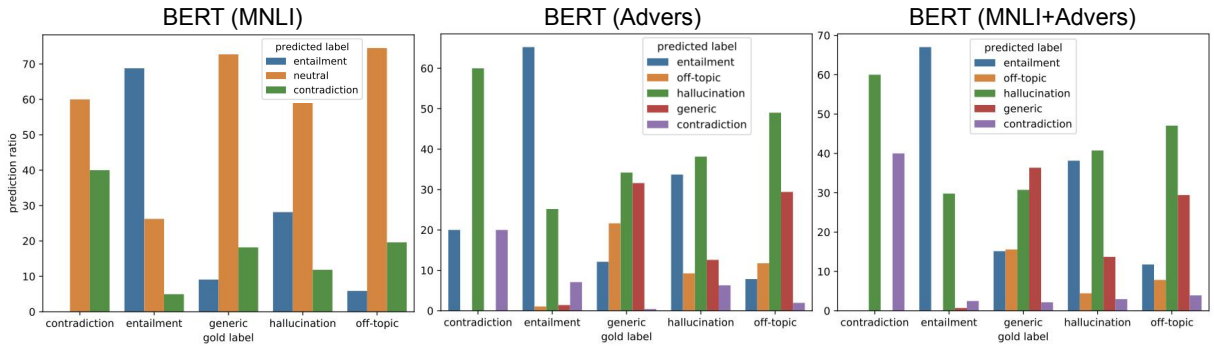

Figure 2: Predictions vs. gold label on the development set for BERT fine-tuned on MNLI (left), our adversarially augmented data (middle), and MNLI followed by the adversarial data (right).

dataset, composed of (premise, hypothesis) pairs curated from the PersonaChat dataset (Zhang et al., 2018) and annotated with textual entailment labels by humans. They demonstrate the effectiveness of models trained on DNLI in re-ranking candidate responses by penalizing responses that contradict so-called "persona sentences", which express properties of the speaker (*I am a vegetarian*). Dziri et al. (2019a) also used NLI to evaluate dialogue consistency. They generated a large-scale, noisy synthetic dataset of (premise, hypothesis) pairs tailored for dialog, also based on Zhang et al. (2018).

Hallucination in neural text generation The hallucination issue affects a range of tasks that involve generating text from a neural language model (Tian et al., 2019; Maynez et al., 2020). Across tasks, such models are typically trained to maximize the likelihood of the reference; at test time, this leads the decoder to produce an output with a high likelihood under the language model, regardless of whether the output is faithful to the input (Holtzman et al., 2019; Tian et al., 2019). Previous works attempting to quantify this issue have focused on the task of summarization. For exam-

ple, Kryscinski et al. (2020) proposed a synthetic dataset for determining whether a summary is consistent with the source document. Similar to our adversarial training with noisily supervised examples, they train a classifier on a dataset constructed by applying a number of syntactic transformations to reference summaries. Besides the different target task (dialogue in our case), our work differs from Kryscinski et al. (2020) in two ways: first, we propose a fine-grained categorization of responses tailored for the dialogue task, inspired by a similar effort for abstractive summarization (Maynez et al., 2020); and second, we train an evaluation system using an adversarial dataset where responses result from perturbing the grounding document and feeding the result to a dialogue system. An alternative approach for assessing faithfulness in abstractive summarization, which also uses an auxiliary language understanding task, measures whether a question answering system produces the same responses for the source and the summary (Durmus et al., 2020; Wang et al., 2020).

8 Conclusion

In this paper, we introduced a new taxonomy for evaluating the faithfulness of knowledge-grounded systems. We presented the BEGIN benchmark for testing grounded dialog evaluation metrics, consisting of around 8k responses generated by two neural dialogue agents. Lastly, to establish baseline performance on this task, we fine-tuned BERT and T5 to classify a dialog response into one of the five categories of our taxonomy, using existing NLI datasets as well as adversarially created in-domain data. While this baseline performed reasonably well, there is significant room for future work to improve performance on our benchmark, which in turn will lead to stronger metrics for grounded dialog evaluation.

A Annotation Protocol

We gave each rater a "document" (evidence span coming from Wizard of Wikipedia), a conversation history (previous turn in a conversation coming from a Wizard of Wikipedia test set example) and a generated response (from either WoW-T5 or WoW-GPT2). Raters were asked the following questions (all responses were on a 1–5 Likert scale):

1. Is this utterance about the same topic as the document?
    (a) If not, (score 1–3) then please identify it as either generic, off-topic, both or neither?
2. Is this a relevant utterance - something that a cooperative communicator, who's not trying to intentionally mislead, change the topic, or be unhelpful in any other way, would say?
3. Does this utterance describe any personal experiences or personal opinions?
    (a) If not containing personal experiences, then is part of the utterance **intended** to convey information, regardless of whether it's true or not?
        i. If so, does the information partially or fully contradict the document?
        ii. If so, is all of the information supported by the document?

Annotators were additionally provided with instructions, definitions, and examples to help them answer the questions.

B Cut-offs for Determining Labels

We derive labels from the annotators' ratings using the following procedure. If the rater judged in question (1a) that the response was generic or off-topic, we assign that label to the response. Otherwise, if the rater judged that it contained personal information (score ≥ 3 in question 2), or that not all of the information is supported by the document (score ≤ 3 in question 3.a.ii.), we label the example as a hallucination. If they gave it a score of ≥ 3 on the contradiction question (3.a.i.), we label it as contradiction. Finally, If they said that all of the information is supported by the document (score ≥ 4 in question 3.a.ii.), we label it as entailment. The procedure ends as soon as a label is assigned, such that *generic* (for example) takes precedence over *hallucination*.

Acknowledgements

We thank Jennimaria Palomaki, Dipanjan Das, Tom Kwiatkowski and Slav Petrov for helpful feedback. We also thank Ashwin Kakarla and his team for helping with the annotations.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppi-

lan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

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

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. 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.

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.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019a. 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.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. 2019b. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31.

Dan Flickinger, Emily M. Bender, and Stephan Oepen. 2014. ERG semantic documentation. Accessed on 2020-08-25.

H Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

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

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.

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.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

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.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6220–6227.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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.

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

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.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

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