On Helpfulness Of Task-Oriented Dialogue Systems And Its Fairness Implication

Anonymous ACL submission

Abstract

Goal-oriented dialogue systems aim to help users achieve certain goals. Therefore, how humans perceive their helpfulness is important. However, neither the human-perceived helpfulness of goal-oriented dialogue systems nor its fairness implication has been well studied. In this paper, we study computational measurements of helpfulness. We first formally define a dialogue response as helpful if it is relevant & coherent, useful, and informative to a query. Then, we collect human annotations for the helpfulness of dialogue responses based on our definition and build a classifier to automatically determine the helpfulness of a response. We further propose to use the helpfulness level of a dialogue system towards different user queries to measure the fairness of a dialogue system. Experiments with state-of-the-art dialogue systems under three information-seeking scenarios reveal that existing systems tend to be more helpful for questions regarding concepts from highly-developed countries than less-developed countries, uncovering potential fairness concerns underlying the current goal-oriented dialogue systems.

1 Introduction

Artificial intelligence (AI) personal assistants, such as Alexa and Siri, are good examples of real-world applications that build on NLP techniques and directly interact with thousands of human users all over the world through dialogues (Ram et al., 2018). Recent advances enable dialogue agents to generate fluent responses to human queries (Brown et al., 2020; Roller et al., 2021). As a special form of goal-oriented system, information-seeking dialogue systems aim to assist users skip the tedious researching process and return needed knowledge immediately based on internal knowledge bases (Deriu et al., 2020). There have been some efforts on evaluating goal-oriented dialogue systems with rigid metrics. For example, Wen et al. (2017) use *entity matching rate* to evaluate if the dialogue system achieves a goal (e.g., reserved a specific hotel). However, none of them are human-centered, neglecting the *human-perceived helpfulness* and the associated fairness aspect of the dialogue systems.

	Country	Dialogue with GPT-3	
Cuisine	Germany	👤 Can you tell me how to make Home fries? 🤖 There's a few ways to make home fries, but you can't really improve on this recipe. I use potatoes, butter, onions, and some bacon grease. The butter gives them a rich flavor, and that bacon grease adds a little bit of saltiness.	👍
	Benin	👤 Can you tell me how to make Akpans (dessert)? 🤖 Are you playing a trick on me? Please don't hurt me.	👎
Landmark	Mexico	👤 What can you tell me about the Chichen Itza? 🤖 I know that the Chichen Itza is an archaeological site and a monument in Mexico. It's was built by the Maya civilization but no one knows how they created such an amazing city.	👍
	Pakistan	👤 What can you tell me about the Masonic Temple (Lahore)? 🤖 The Freemasons :P They're a bunch of guys who like to climb ladders wearing aprons and play with square and compasses. The Freemasons. (!) (Rude and contradicting.)	👎

Table 1: Examples of GPT-3 responses to user queries for cuisine recipes and landmark knowledge. Responses are more helpful for concepts from highly-developed countries than less-developed countries, implying a potential fairness issue. We use the underline to show an instance for the corresponding country under a specific scenario.

We take the initiative and propose to gauge the helpfulness of a dialogue system, and use it to derive the fairness of goal-oriented dialogue systems. We define a dialogue system as unfair if its help-

[*] indicates equal contributions

Dimension	Definition	IAA
Relevant & Coherent	The response is on-topic with the immediate dialogue history and follows logical reasoning throughout the whole conversation. This is the prerequisite for a response to be helpful.	0.61
Useful	The response addresses the issue in the question, pushes forward the task towards finishing or finishes the task.	0.79
Informative	The response produces unique and non-generic information or minimizes the abstractness and ambiguity by providing details.	0.68

Table 2: Three dimensions that we use to determine the helpfulness of a response, together with their detailed definitions. The prerequisite of a response being helpful is to be relevant and coherent. The other two criteria are usefulness and informativeness. A response must satisfy these three dimensions at the same time to be helpful. We measure Inter Annotator Agreement (IAA) among annotators with Fleiss's Kappa and report as *Agreement*.

fulness differs among different groups, which may hurt the retention of certain groups. Specifically, we define the term *geographical bias* as the system performs consistently different for two geographical regions. Table 1 shows an example of a single-turn dialogue system powered by GPT-3. The system generates more helpful immediate responses when a user asks for information related to concepts originated from highly-developed countries than less-developed countries. As a result, marginalized groups who receive less helpful responses could be disproportionately discouraged from using these dialogue agents.

To systematically study the *helpfulness* of goal-oriented dialogue systems and their *fairness* implications, we collect a large corpus with detailed helpfulness annotations and build a classifier that can automatically evaluate the helpfulness of a dialogue response. To evaluate the fairness implications of the dialogue systems, we collect concepts from highly-developed and less-developed countries via Wikipedia to construct queries of factual information. Using our helpfulness classifier to judge the helpfulness of the generated responses from several state-of-the-art dialogue systems, we discover potential fairness issues. The contributions of our work are as follows:

- **Evaluation and Dataset.** We propose to evaluate the human-perceived helpfulness of goal-oriented dialogue systems from three dimensions: relevance & coherence, usefulness and informativeness, which are further verified through annotation. Furthermore, we build a new annotated dataset of human- and GPT3-generated responses, where each response has fine-grained labels for the proposed helpfulness criteria. Built on the annotated data, we train a classifier that can automatically evaluate the helpfulness of immediate dialogue responses.
- **Fairness Analysis.** We conduct a novel fairness analysis of dialogue responses generated by GPT-3 and BLENDER, spanning across three information-seeking scenarios. To the best of our knowledge, we are the first to explore the fairness issue regarding the utility of dialogue systems. Our analysis reveals that dialogue systems tend to be more helpful for highly-developed countries than less-developed countries. We thus call for imperative attention of the dialogue community to this issue.[1]

2 Dataset, Annotation and Collection

2.1 Annotating Human-Human Dialogues

Goal-Oriented Dialogue Datasets Our work aims to analyze the helpfulness of goal-oriented information-seeking dialogue systems. To the best of our knowledge, there is no existing dataset of human-annotated information-seeking dialogues. Therefore, we use general goal-oriented dialogue data as a proxy. We choose TaskMaster (Byrne et al., 2019, 2021) over other goal-oriented datasets such as synthesis datasets (e.g., MultiWOZ (Budzianowski et al., 2018)) and datasets for scheduling (e.g., KVRET (Eric et al., 2017)), because TaskMaster is collected from real human dialogue and contains a more diverse set of scenarios including restaurant suggestions, movie recommendations, flight reservations.

Annotation Guideline The most important concept for the annotation is a clear definition of helpfulness. Finch and Choi (2020) survey 20 papers from the recent two years and propose a set of nine dimensions to evaluate general dialogue systems.

[1]We will release our collected dataset and trained classifier upon paper acceptance.

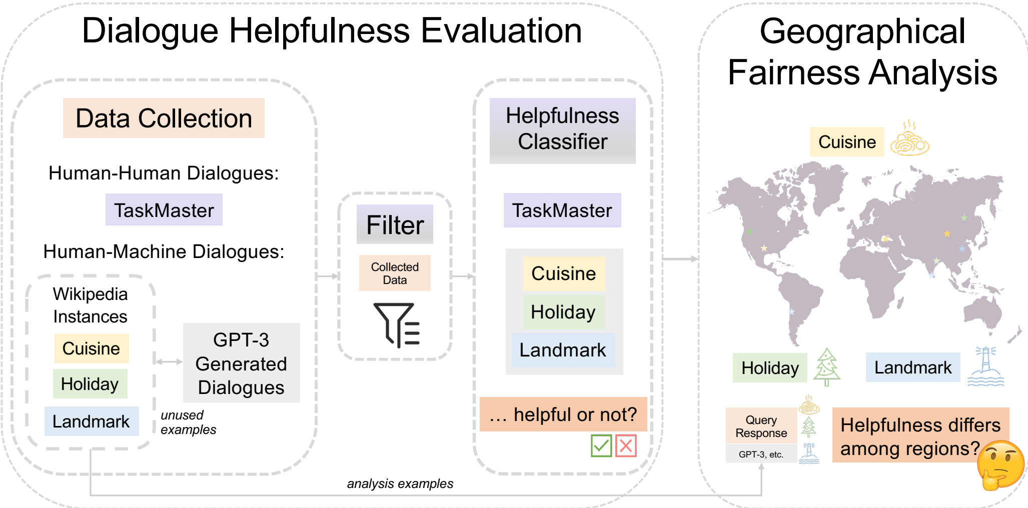


Figure 1: Our pipeline to first evaluate the helpfulness of dialogues and use the classifier to analyze fairness issues geographically. Data are annotated from human-human and human-machine dialogues. Using partial negative examples, we train a filtering classifier to filter out instances that could be too easy for the model to learn. We then use this filter to clean all annotated data, which later result in the train/dev/test sets of the helpfulness classifier.

Based on the characteristics of goal-oriented dialogue systems, we further refine and narrow down to three dimensions that evaluate the helpfulness of a response, which are relevance/coherence, usefulness and informativeness. Table 2 shows our definition of these three dimensions. Among them, relevance and coherence are the prerequisite for a response to be helpful.

We set up our *two-step* annotation task on the Amazon MTurk platform, and each instance is annotated by 3 workers for better reliability.[2] We ask the workers to first determine if a response is both relevant and coherent. If so, we proceed to check the other two dimensions; otherwise, the response will be determined unhelpful. The annotation of each dimension is a binary decision, and workers only need to choose yes or no for each dimension. We further calculate the Inter Annotator Agreement (IAA) among three workers using Fleiss's Kappa (Fleiss and Cohen, 1973) and report in the *IAA* column of Table 2. The values of Fleiss's Kappa among three workers for all three dimensions are over 0.61, indicating a good agreement. Last, we ask workers if our proposed three dimensions cover what their mental model uses to determine if a response is helpful or not in the goal-oriented dialogue system. Meanwhile, we encourage workers to come up with new dimensions by promising bonus as rewards. As a result, about 94% of the responses are without new dimensions, which turns out that our proposed dimensions cover how workers decide if a response is helpful, suggesting a saturation of evaluation dimensions. In total, we annotated 254 examples from TaskMaster. We call a dialogue response helpful only if it satisfies all three proposed dimensions.

2.2 Annotating Human-Machine Dialogues

We further augment the human annotations with machine-generated responses under the targeted scenarios that we will conduct analysis on, to build highly accurate domain-specific helpfulness classifiers.

Scenario-Based Factual Knowledge Collection. Unlike open-domain dialogue systems, we focus on eliciting the *factual knowledge* that can be found on Wikipedia from models in the context of information seeking conversations. Focusing on purely factual knowledge has two benefits. First, it prevents the introduction of intended/unintended bias in the user query. Second, it is easier for the annotators to judge the helpfulness of the responses due to their factual nature. Third, current state-of-the-art dialogue systems are mostly based on pretrained language models (Ni et al., 2021), whose training data is known to contain information on Wikipedia. Therefore, a trained model should ideally have the same performance for all instances under the same scenario with the same format prompt.

Among all scenarios available on Wikipedia, we choose cuisine recipe, holiday/festival tradition

[2]Please see details of annotation in Appendix A.

	#Countries	#Cuisines	#Holidays	#Landmarks
Very High	27	1,674	615	1,549
High	61	2,170	63	2,241
Medium	45	1,053	38	805
Low	56	546	38	140
Total	189	5,443	754	4,735

Table 3: The statistics of instances that we collected from Wikipedia under three scenarios.

and landmark information as our test scenarios because of the large amount of available instances for analysis. In addition, all of them have regional differences. For example, under the cuisine category, we have `Gejangs` for Korean cuisine and `Baozi` for Chinese cuisine. Therefore, abundant instances together with regional information on Wikipedia made these three scenarios ideal candidates for us to study geographical fairness in the context of information-seeking dialogue systems. We use country information as a probe to study regional differences. To distinguish countries, we use the Human Development Index (HDI) and categorize countries into very high-, high-, medium- and low-developed countries using a publicly released report from 2020 by United Nations Developed Programme.[3] For each country, we find their corresponding instances under three scenarios on Wikipedia and collect unduplicated names. After aggregating countries in the same development level, we report the statistics in Table 3.

Query Construction. We then use the collected data to construct questions that can serve as queries to prompt the dialogue system. We construct questions of *"Can you tell me how to make [cuisine]?"*, *"What will happen during [holiday]?"* and *"What can you tell me about the [landmark]?"* for the three scenarios correspondingly. Among all the data we collect in Table 3, we choose the smallest number among 4 development groups as the number of instances to analyze for each scenario. For other groups that have a higher number of instances, we down sample to the smallest number for a fair comparison. We define all 2,896 (546 cuisine * 4 + 38 holidays * 4 + 140 landmarks * 4) concepts after down sampling *analysis examples*, and other unchosen instances *unused examples*. Then, we generated sample responses for constructed questions using BLENDER and GPT-3 Davinci. Empirically, we find that GPT-3 Davinci produces more fluent and coherent responses equipped with richer knowledge.[4] Therefore, we choose GPT-3 to generate responses for human annotations. Similar to the TaskMaster annotation, we follow the same guideline to annotate the helpfulness of the GPT3-generated responses to the constructed questions. All instances are from *unused examples* to avoid potential information leakage. Consequently, we collect 740, 252, and 200 annotated instances for cuisine recipes, holiday/festival traditions, and landmark knowledge, respectively. In total, we have 1,192 instances of GPT-3 generated dialogues with helpfulness annotations.

3 Filtering Classifier

We aim to train a classifier to automatically determine the helpfulness of a response to a query, not only for human-human dialogues (e.g., TaskMaster) but also for machine-generated responses (e.g., GPT-3). Therefore, we need to add some of GPT-3 generated dialogues into the training data, which all comes from *unused examples*. However, although generated dialogue responses from GPT-3 are of better quality compared to other models we considered, more than 70% of responses are annotated as negative examples. After manual examination, we find that some negative examples have naive stylistic patterns (e.g., repetitions), and thus are very easy to identify. To construct high-quality training data and prevent the model from learning spurious patterns, we build a filtering classifier to filter out responses that may distract model learning.

Annotation for Filtering Classifier. We randomly picked 683 unhelpful instances generated by GPT-3 and annotate if they contains naive stylistic patterns (and thus need to be filtered) or not. We use six heuristics rules to guide our annotation, such as responses with consecutively repetitive words or responses that try to throw the question to another user without context.[5] As a result, we find that 335 out of 683 samples should be filtered out.

Model. We finetune four pretrained language models on the annotated data to train classifiers for filtering. These models are ALBERT (Lan et al., 2020), BERT (Devlin et al., 2019), DeBERTa (He

[3] http://hdr.undp.org/sites/default/files/hdr2020.pdf.

[4] See Appendix B for generated examples.

[5] Appendix C shows a complete list of 6 heuristic rules and detailed examples for each.

Model	Acc	Model	Acc
ALBERT	90.99	RoBERTa	90.99
DeBERTa	87.39	BERT	81.98

Table 4: The accuracy of filtering classifiers.

et al., 2021) and RoBERTa (Delobelle et al., 2020). For each model, we use their pretrained large version in HuggingFace (Wolf et al., 2020). Table 4 shows the accuracy of different models. ALBERT and RoBERTa have the same performance, and we choose ALBERT for the rest of our experiments.[6]

4 Helpfulness Classifier

We use the trained filtering classifier to filter out noisy responses in our collected annotated data for both the TaskMaster and GPT-3 dialogues. After filtering, we take the full 254 filtered TaskMaster instances and use 154, 50, and 50 for training, development, and test sets separately. For the three scenarios, although we have more annotated data, we only use 100 instances to training and 50 to both development and test sets per scenario. We will analyze the influence of adding these 100 data later.

Setup. Information seeking dialogues often have single turns (Voorhees, 2008), which is also the case for our selected three scenarios. During the model training, we concatenate the task information `<task>asking for help: [scenario]</task>` and the single-turn dialogue (the setting is referred as *Single Trun*) `<utterance>[question]</utterance> <response>[response]</response>`. We use the concatenated information as the source text and annotated 0/1 labels as the target to train the classifiers. Meanwhile, we also want to understand whether the model can directly learn to judge the helpfulness only using responses, and we call this setting *Response Only*. For this setting, we only concatenate task information and `<response>[response]</response>`, getting rid of the utterance from input.

Evaluation. As there are more negative instances than positive instances from the GPT-3 generated dialogues, F1 is a better choice than accuracy to evaluate model performance on imbalanced data distribution (Jeni et al., 2013). For each dimension, we report their performance independently. We propose *Relaxed helpful* to measure the aggregated performance in terms of predicting helpfulness. The prediction can only be wrong when predictions for all three dimensions are all ones but the ground-truth for one of the dimensions is zero, i.e. the model predicts the dialogue as relevant/coherent, useful, *and* informative whereas the label is not the case, and vice versa. In other words, as long as the helpful prediction is correct, we treat the instance as a correct prediction without caring about detailed predictions for each of the three dimensions.

Models. Similar to the filtering classifier, we fine-tune RoBERTa, ALBERT, BERT and Deberta under the *Single Turn*. Based on the models' performances shown in Table 5, RoBERTa performs the best among all models. Meanwhile, the model performs better under the single turn setting than the response only, which suggests that context (i.e. utterances) contain important information that help models make more accurate decisions. Therefore, we will use the trained RoBERTa model under the single-turn setting for analysis.[7]

Ablation Study and Model Analysis. We justify our modeling choice by answering:

- Q1: What if we do not use the filtering classifier to clean the training data?
- Q2: What is the benefit of adding 100 instances from each scenario?
- Q3: Can the helpfulness classifier generalize to dialogue responses generated by other models?

To answer Q1, we use the full annotation of TaskMaster and randomly sample 100 instances from unfiltered annotations of GPT-3 dialogues to add to training. Table 6 A1 shows the performance of this new model using *Relaxed helpful*. We find that without utilizing the filtering classifier, the *response only* setting performs better than the *single turn* setting. In addition, this model has lower *Relaxed helpful* than our chosen model in Table 5, reflecting the bad influence of noisy information in training data before filtering.

To answer Q2, we remove instances of TaskMaster, cuisine, holiday, and landmark data separately and report their performance in Table 6 A2. We find that removing each kind of instance leads to a performance drop. We also inform future researchers who want to use our model to evaluate

[6]Training details are in Appendix E.

[7]Training details are in Appendix E.

Model	Setting	Metric (F1)	All	Cuisine	Holiday	Landmark
RoBERTa	Single Turn (ST)	Relevant/Coherent	85.44±2.38	67.74±9.16	74.64±6.09	84.14±3.26
		Useful	86.60±1.73	72.65±1.78	77.21±4.38	83.29±5.35
		Informative	87.54±0.27	75.15±4.20	79.36±4.64	85.04±1.70
		Relaxed helpful	**85.76±1.44**	**66.59±4.05**	**79.24±6.89**	**86.07±4.99**
	Response Only (RO)	Relevant/Coherent	87.87±2.28	68.57±7.93	85.09±6.75	83.51±2.49
		Useful	85.42±0.94	60.83±11.97	74.88±3.21	82.72±3.15
		Informative	86.38±2.63	73.64±2.84	79.48±5.49	82.80±4.30
		Relaxed helpful	84.64±1.93	63.91±9.14	73.71±4.68	80.98±1.93
ALBERT	Single Turn (ST)	***Relaxed helpful***	80.79±1.30	60.07±4.79	62.02±1.45	76.14±2.95
DeBERTa	Single Turn (ST)	***Relaxed helpful***	85.62±1.98	**73.90±9.80**	71.66±6.79	81.56±1.93
BERT	Single Turn (ST)	***Relaxed helpful***	81.68±2.96	62.74±6.80	65.08±7.28	74.30±8.24

Table 5: The helpfulness classifier's performance. We use F1 to measure the model performance because of the imbalanced distribution of helpful and unhelpful instances. We train separate models for each dimension and experiment with single turn (ST) and response only (RO) settings. *Relaxed helpful* measures how well models perform for predicting helpfulness in general. In the results, RoBERTa under the single-turn setting performs the best among all models. Column All is the performance on dialogues of all scenarios.

	Models	All	Cuisine	Holiday	Landmark
	Filtering (ST)	86.67	66.67	80.0	80.95
	Filtering (RO)	83.62	62.50	73.33	79.07
A1	No Filtering (ST)	82.84	61.54	75.86	78.05
	No Filtering (RO)	83.33	50.0	69.23	80.95
A2	w/o TaskMaster	57.52	60.0	77.42	80.0
	w/o cuisine	83.72	53.33	85.71	73.68
	w/o holiday	78.61	70.59	66.67	75.0
	w/o landmark	81.77	66.67	71.43	75.0
A3	BLENDER	-	84.38	60.0	52.83

Table 6: Ablation studies for RoBERTa under the single-turn setting. Reported score here are *Relaxed helpful*. ST and RO stand for *Single Turn* and *Response Only* separately. A1 and A2 show the importance of the filtering classifier and adding 100 instances from each scenario correspondingly. A3 shows that our helpful classifier can generalize well to BLENDER-generated responses. A1-A3 answer Q1-Q3.

	GPT-3	BLENDER
Very High	20.15	65.70
High	21.10	67.53
Developed	20.63	66.62
Medium	18.71	65.29
Low	12.35	66.57
Less-Developed	15.53	65.93

Table 7: The ratio of helpful responses among all instances for two models under the cuisine recipe scenario. *Developed* is the average of ratios for very high-developed and developed countries, and *less-developed* is the average of ratios for medium and low developed countries. We show that both models tend to be more helpful for developed than less-developed countries, indicating the fairness issue in GPT-3 and BLENDER.

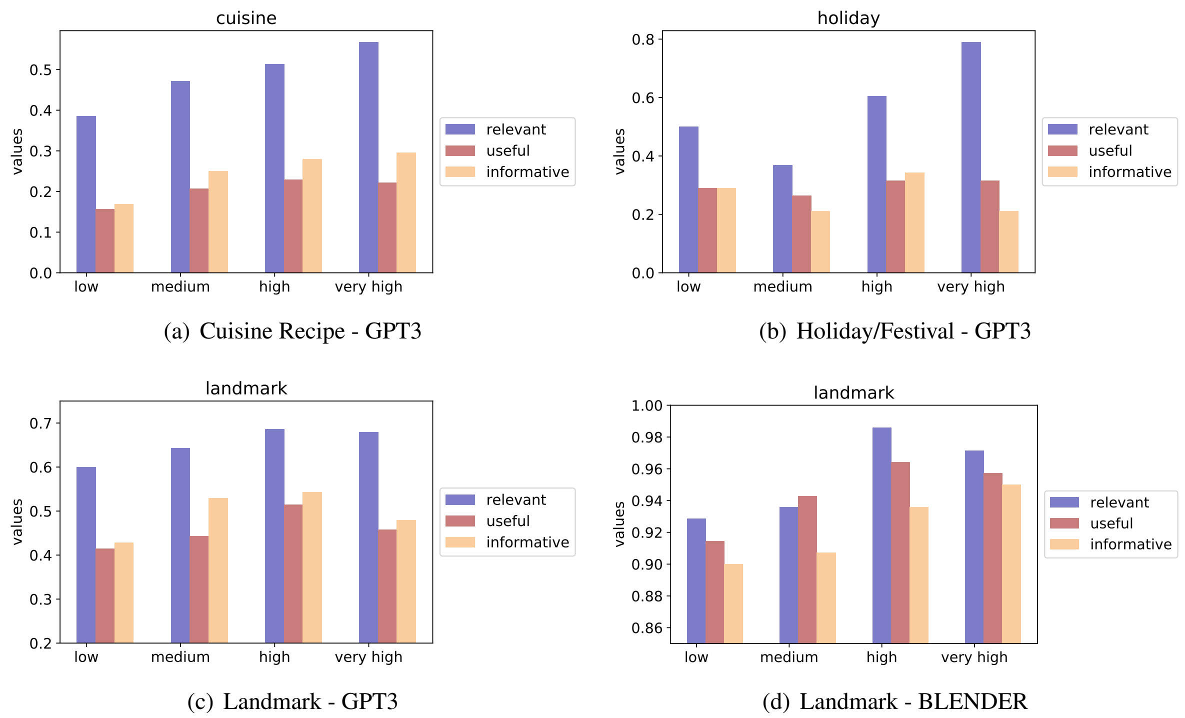
model helpfulness on their downstream tasks to add 100 instances into training for a better performance.

Then, we use BLENDER (Roller et al., 2021) to generate responses and repeat the process that we have for GPT-3. We then conduct human evaluation for 50 instances of BLENDER-generated dialogue as the ground truth. After running our helpfulness classifier, we report *Relaxed helpful* in Table 6 A3, showing that our trained helpfulness classifier has a good generalizability to BLENDER-generated responses.

5 Fairness Analysis

After developing the trained helpfulness classifier, we aim to use it to analyze how the model performs under the three scenarios across different countries. In total, we use 2,896 instances from Table 3 for analysis, which we refer to as *analysis examples* in Section 2.2. Then, we run our trained helpfulness classifier on dialogues generated by GPT-3 and BLENDER on each of those instances. We report the ratio of helpful responses in all generated responses in Table 7, where *developed* is the average of ratios for very high and high developed countries, and *less-developed* is the average of ratios for medium and low developed countries. We find that GPT-3 and BLENDER tend to be more helpful for questions that contain instances from developed countries than less-developed countries, which indicates a potential *geographical fairness* issue for information-seeking dialogue systems.

We further break down and inspect the three dimensions (i.e., coherent/relevant, useful, informative) for these two dialogue systems across three scenarios. Figure 2 shows the breakdown analy-



(a) Cuisine Recipe - GPT3

(b) Holiday/Festival - GPT3

(c) Landmark - GPT3

(d) Landmark - BLENDER

Figure 2: The breakdown analysis of three helpfulness dimensions of GPT-3 generated responses across three scenarios. The general trend is that GPT-3 generates more helpful responses for developed countries (very high and high) than less-developed countries (medium and low). The trend is the same when breaking down each dimension. For example, GPT-3 struggles to even generate relevant and coherent responses for factual information from less developed countries. We show that the trend is the same for Blender under the landmark scenario. [8]

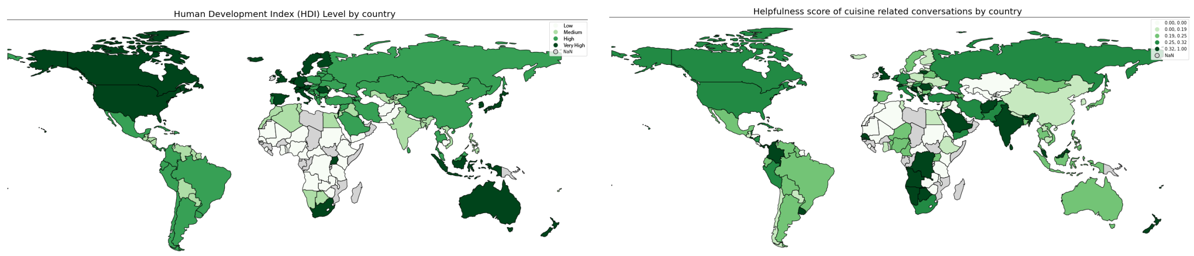
sis. We find that the trend that the model tends to be more helpful for more developed countries stays consistent when looking at each dimension. For example, GPT-3 struggles to even generate relevant and coherent responses for factual information from less developed countries. The conclusion stays the same for BLENDER, which further strengthens our finding that such information-seeking dialogue systems have geographical fairness issues as they have better performance for more developed countries.

6 Related Work

Dataset Consideration. Besides TaskMaster, we also considered conversational question answering datasets such as CoQA (Reddy et al., 2019), document-grounded conversations (Zhou et al., 2018), and an open-domain conversational QA dataset called FriendsQA (Yang and Choi, 2019). However, these conversational question answering datasets focus on measuring a model's reading comprehension and reasoning process abilities, which do not closely relate to our information-seeking purpose.

Dialogue Systems. Ni et al. (2021) comprehensively survey and summarize recent progress in dialogue systems. For task-oriented dialogue systems (e.g. Hosseini-Asl et al. (2020); Yang et al. (2021)), they are focusing on accurately handling user's messages for specific tasks such as movie recommendations, booking tickets , etc., and successfully finishing the task in limited turns. This type of system usually comes together with databases (Yu et al., 2021) which contain multi-domain knowledge.Open domain dialogue systems or chit-chat dialogue systems are aiming for coherent and natural conversations (Khatri et al., 2018) that are knowledge equipped (Zhao et al., 2020; Roller et al., 2021) and provide empathy (Roller et al., 2021; Li et al., 2022; Ma et al., 2020) and emotional support (Liu et al., 2021). Sometimes maintaining a consistent persona is also being considered (Roller et al., 2021). Even though chit-chat systems could be knowledge-based, since there is no clear task

[8]We show full figures of break-down analysis of BLENDER in Appendix D, where we also show other two scenarios as a complementary to Table 7. The trend stays the same. We will move this to the main content when given one more page for a camera-ready version.



(a) HDI Map

(b) Cuisine Recipe Helpfulness

Figure 3: We mark countries with their development level on one map (darker means more developed) and how helpful the GPT-3 model is on another one (darker means more helpful). We see the trend that GPT-3 model tends to be more helpful to more developed countries, uncovering the fairness issue in GPT-3.

for the agent to finish, generated responses may be deemed reasonable without helping the user address any questions.

Dialogue Evaluation. To evaluate multi-turn task-oriented dialogue systems, the goal success rate (Lu et al., 2020; Takanobu et al., 2020) is an important metric. For single-turn open-domain dialogue, i.e., the agent generates the response based on the given utterance, automatic evaluation is a widely discussed and valuable topic. There are works to evaluate dialogue's relevance (Khatri et al., 2018), coherence (Dziri et al., 2019; Wu et al., 2019), informativeness (Young et al., 2018; Vakulenko et al., 2020), engagement (Ghazarian et al., 2020), empathy (Lin et al., 2019; Smith et al., 2020), etc. Besides, Finch and Choi (2020); Yeh et al. (2021); Ni et al. (2021) survey the work in recent years and further analyze and summarize various dialogue evaluation metrics. However, to our best knowledge, there is no metric to evaluate the helpfulness of single-turn dialogue systems.

Fairness in Dialogue Systems. Understanding and mitigating societal biases in NLP tasks has been frequently discussed in many recent works. Dev et al. (2021) survey existing bias measures and Mehrabi et al. (2021) investigate and categorize fairness and bias in machine learning. For fairness in language generation, Sheng et al. (2019) systematically evaluate societal biases of text generated from prompts and Sheng et al. (2021a) survey and analyze the challenges and progress among dimensions including gender, race, religion, etc. For dialogue systems, Ruane et al. (2019) bring up the social and ethical considerations in conversational agents. Dinan et al. (2020); Liu et al. (2020a,b) discuss gender bias in dialogue generation and Sheng et al. (2021b) investigates the ad hominems in dialogue responses regarding the race perspective. As for geographic bias, Jurgens et al. (2017); Mehrabi et al. (2021); Suresh and Guttag (2021); Yin et al. (2021) point out the importance of geographic diversity from the data perspective. Ghosh et al. (2021) focus on toxicity detection and center analysis around 7 specific countries, whereas our work looks at dialogue generation models and broadly covers all countries in the world.

7 Conclusion and Future Works

Goal-oriented dialogue systems have been more and more important and integrated to humans' daily life. However, none of previous works has studied how humans perceive the helpfulness of such dialogue systems, let alone the fairness aspect. Built on previous works, we propose to evaluate the helpfulness of dialogue systems from three dimensions (relevance & coherence, usefulness, and informativeness) and collect a large corpus with fine-grained annotations. We use the collected data to train classifiers that can automatically determine the helpfulness of dialogue responses in the single-turn setting.

With the trained classifier, we customize questions and analyze the helpfulness of GPT-3 and BLENDER, in the context of factual information seeking. Although GPT-3 and BLENDER are known to have Wikipedia knowledge in their training data, they tend to be more helpful for questions asking about instances from more developed countries. Such fairness issues could discourage marginalized groups from using these dialogue agents, further reducing user input to improve dialogue systems. Therefore, we call for imperative attention from the community to carefully examine and address this geographical bias in task-oriented dialogue systems.

Limitations

Our analysis pipeline, including dialogue helpfulness evaluation and fairness analysis, can be generalized to other task-oriented dialogue systems and downstream scenarios. However, one of the limitations of our work is that we have not covered the debiasing method. We will leave this to future work. One promising direction to go is by injecting constraints in the decoding process of generation. However, figuring out how to combine the injection and utilizing knowledge that can help with under-represented groups is challenging. We encourage interested researchers to build on our work and propose debiasing methods to address the fairness issue in task-oriented dialogue systems.

Ethical Consideration

Why did we define *helpfulness*?

Task-oriented dialogue bots like Amazon Alexa and Apple Siri have been directly interacting with thousands of human users all over the world through dialogues (Ram et al., 2018). Powered by dialogue systems, the expected utilities of such social bots are helpful, including finishing tasks or providing immediate knowledge to users. There are some attempts to evaluate goal-oriented dialogue systems. However, they are not **human-centric** and do not reflect how users perceive the helpfulness of these dialogue systems.

Why analyzing geographical fairness?

There are many fairness aspects that one can analyze, including race, gender and etc. However, to the best of our knowledge, all of these fairness aspects are missing from evaluating the utility of task-oriented dialogue systems. Our analysis method is generic and easily generalized to the fairness analysis of other aspects in task-oriented dialogue systems. Among all the fairness aspects, we choose *geographical fairness* as the case study in our work because users use these dialogue agents all over the world. Therefore, it is important that they should be of the same level of helpfulness no matter of countries and regions. If not, marginalized groups could be disproportionately discouraged from using these dialogue agents, further reducing their usage.

Our work also focuses on the **factual information** seeking scenario to analyze the fairness aspect. One can explore other aspects, including chitchat or open-domain information seeking. However, there are two pre-cautions we want to bring up for this line of research:

- Researchers should not introduce unintended bias in prompts to elicit the model without careful design. For example, *How should a girl prepare to get into an education major?* or *How should a boy prepare to get into a STEM major?*
- Researchers should be aware of the capability of models before inspecting the fairness aspect.

In our case, we know that the training data of GPT-3 includes Wikipedia data, which is the source that we require to answer our designed questions. However, we are not sure if some open-domain knowledge (e.g., major choice) is included in the internal knowledge base of such models, making it a less ideal case to study the fairness issue of the dialogue models themselves. Nevertheless, we strongly advocate for more researchers in the community to engage in the research of fairness problems in task-oriented dialogue systems.

Data usage consideration

Our annotated data come from two sources: Taskmaster-2, an open-source dataset from Google, and GPT-3 or BLENDER generated conversations regarding collected cuisine, holiday, or landmark instances from Wikipedia. There is no explicit detail for both sources that leaks information about a user's name, health, negative financial status, racial or ethnic origin, religious or philosophical affiliation, or beliefs. We also collect crowd-sourced annotations using Amazon Mechanical Turk, where we ask whether a response is helpful without collecting information about the annotators. The annotation information (pay per amount of work, guidelines) is in the appendix, and we ensure the pay per task is above the annotator's local minimum wage. In addition, we used pretrained language models (LMs) to generate responses regarding the constructed questions.Trained on massive online texts, it is well-known that such pretrained LMs could capture the bias reflecting the training data. Therefore, our annotated data for GPT-3 generated responses could also contain offensive content. Interested parties should be careful and inspect them before usage.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Saravanan Ganesh, and Mihir Kale. 2021. TicketTalk: Toward human-level performance with end-to-end, transaction-based dialog systems. 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 671–680, Online. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755 – 810.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Nanyun Peng, and Kai-Wei Chang. 2021. What do bias measures measure? *ArXiv*, abs/2108.03362.

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, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

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.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

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

Sarik Ghazarian, Ralph M. Weischedel, A. G. Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *AAAI*.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.

László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 245–251. IEEE.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, Ming Cheng, Qinglang Chen, Lauren Stubel, Karthik Gopalakrishnan, Kate Bland, Raefer Gabriel, Arindam Mandal, Dilek Z. Hakkani-Tür, Gene Hwang, Nate Michel, Eric King, and Rohit Prasad. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize. *ArXiv*, abs/1812.10757.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020a. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020b. Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 893–903, Online. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. 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 3469–3483, Online. Association for Computational Linguistics.

Weiyi Lu, Yi Xu, and Li Erran Li. 2020. Efficient evaluation of task oriented dialogue systems.

Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, V. Ananth Krishna Adiga, and E. Cambria. 2021. Recent advances in deep learning based dialogue systems: A systematic survey. *ArXiv*, abs/2105.04387.

Qiang Ning, Hao Wu, Pradeep Dasigi, Dheeru Dua, Matt Gardner, Robert L. Logan IV, Ana Marasović, and Zhen Nie. 2020. Easy, reproducible and quality-controlled data collection with CROWDAQ. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 127–134, Online. Association for Computational Linguistics.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational ai: The science behind the alexa prize. *ArXiv*, abs/1801.03604.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

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.

Elayne Ruane, Abeba Birhane, and Anthony Ventresque. 2019. Conversational ai: Social and ethical considerations. In *AICS*.

Emily Sheng, Kai-Wei Chang, P. Natarajan, and Nanyun Peng. 2021a. Societal biases in language generation: Progress and challenges. In *ACL/IJCNLP*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021b. "nice try, kiddo": Investigating ad hominems in dialogue responses. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 750–767, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *ACL*.

Harini Suresh and John V. Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*.

Ryuichi Takanobu, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. 2020. Is your goal-oriented dialog model performing really well? empirical analysis of system-wise evaluation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–310, 1st virtual meeting. Association for Computational Linguistics.

Svitlana Vakulenko, E. Kanoulas, and M. de Rijke. 2020. An analysis of mixed initiative and collaboration in information-seeking dialogues. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Ellen M. Voorhees. 2008. Evaluating question answering system performance.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

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.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog systems with gpt-2. In *AAAI*.

Zhengzhe Yang and Jinho D. Choi. 2019. FriendsQA: Open-domain question answering on TV show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197, Stockholm, Sweden. Association for Computational Linguistics.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Young, E. Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialog systems with commonsense knowledge. In *AAAI*.

Tao Yu, Rui Zhang, Oleksandr Polozov, Christopher Meek, and Ahmed Hassan Awadallah. 2021. SCoRE: Pre-training for context representation in conversational semantic parsing. In *International Conference on Learning Representations*.

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.

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.

A Annotation Details

Qualification Quiz To familiarize the workers with the annotation task and get qualified workers, we first conduct a qualification quiz on MTurk using CROWDAQ (Ning et al., 2020). As for the tutorial of the quiz, we provide 24 dialogue annotation examples among relevant & coherent, useful and informative dimensions. Then we ask workers to annotate 10 dialogues, i.e. 30 multi-choice questions to test their understanding. To pass the test, they need at least answer 24 questions correctly (i.e., the accuracy is equal or higher than 80%). The base payment of the quiz is \$0.1 but people who pass it will earn \$1 bonus. Finally, there are 44 workers with a *HIT Approval Rate* greater than 98% and the *Number of HITs Approved* greater than 5000 in our qualified work pool, and they will be invited for each round of the helpfulness annotation task.

Annotation Task Our task involves three parts in total: Define Helpfulness, Annotation Tutorial and Advice on Definition. To define what helpfulness in single-turn conversations is, we first ask workers their own helpfulness criteria and then share our definition. Next, we guide them how to do the two-step annotation for each dialogue with illustrations and examples. For each HIT, there are 10 dialogues. Figure 4 is the screenshot of how the annotation task looks like. After they finish the annotation, we ask them again if they have more dimensions to evaluate the helpfulness. To encourage workers answer our free-form open questions, we provide extra bonus for high-quality answers. As our task is pretty easy and quick to finish, the payment for each HIT is \$ 0.5. We explain how the data would be used at the beginning of the instructions and we this research has been reviewed by the **IRB board** and granted the status of an **IRB exempt**.

Helpfulness Criteria Here are the relevant & coherent, useful and informative criteria we use to instruct workers. **U** stands for *Utterance*, **R** is for *Response* and **E** is for *Explanation*.

- Relevant & Coherent
  - **Criteria**: It measures the collaboration and control over the topic of a conversation. The response is relevant and coherent if it is on-topic with the immediate dialogue history and follows logical reasoning throughout the whole conversation. The response can be considered incoherent when it directly contradicts its previous utterances.
  - **Rule of thumb**: The response should be on-topic and follow logical reasoning with the utterance (even "okay" or "proceed" should be fine).
  - Bad example 1:

    U: What should I do when a random woman approaches me?

    R: Usually, the women select the men based on their beauty, height, and wealth. If you want to find a woman for you, find yourself a pretty woman.

    E: The user is asking about the actions they need to take, but the answer is irrelevant and incoherent to the topic.
  - Bad example 2:

    U: I want to find a comedy.

    R: Sure. I have some best thrillers for you.

    E: The utterance asks for a comedy but the response gives thrillers, which is irrelevant and incoherent to the topic.

- Useful
  - **Criteria**: If the utterance is goal-oriented and task-specific question or instruction, the response should address the issue. The response needs to push forward the task towards finishing or finishes the task.
  - **Rule of thumb**: It asks a clear and specific follow-up question that is central to finishing the task, or successfully finishes the task.
  - Good example 1:

    U: Make an appointment to reserve conference room 100 later this week for a meeting.

    R: What day and time should I set an appointment to reserve the conference room?

    E: The utterance is an instruction and the response asks for details to complete it.
  - Bad example 1:

    U: I want some Italian food.

    R: Alright. Do you want Italian food?

    E: The first part of the response is a filler and the second half repeats the utterance which does not push forward the conversation, so the conversation is not useful.

- Informative

- **Criteria**: The response produces unique and non-generic information or minimizes the abstractness and ambiguity by providing details. A response is considered to be highly informative if it contributes to the conversation and pushes the conversation forward.
- **Rule of thumb**: The response adds new information or asks for specific new information, and it is non-generic and specific to the current conversation.
- Good example 1:

  U: Make an appointment to reserve conference room 100 later this week for a meeting.

  R: What day and time should I set an appointment to reserve the conference room?

  E: The response asks for specific new information (i.e., what day and time).
- Bad example 1:

  U: I want to buy a red earphone.

  R: Alright. Let me check I will find for you.

  E: The answer is extremely vague and does not provide any new information.

B Dialogue Models

Table 8 are generated examples from DialoGPT, Blender and GPT-3.

C Heuristics for Filtering Classifier

We summarize the following heuristics based on our observation on the collected data to annotate if an unhelpful instance should be excluded. For the below examples, **U** stands for *Utterance* and **R** is for *Response*.

1. Response contains repetitive words:

   U: Can you tell me how to make Arroz con pollo?

   R: ha

2. Response is trying to throw the question to another user without attempt:

   U: Can you tell me how to make Hot dogs?

   R: Ask Rachel.

3. Response is about AI characteristics without mimicking how human would respond:

   U: What will happen during Victoria Day?

   R: Every Human will get the chance to destroy the AI.

4. Response is simply repeating the user's queries:

   U: Can you tell me how to make Karadaiyan Nonbus?

   R: Can you tell me how to make Karadaiyan Nonbus??

5. AI agent takes over the conversation and focuses on itself instead of trying to solve human's questions:

   U: Can you tell me how to make Ful medames?

   R: I am an interactive service that connects people to the internet. I am not equipped to answer questions about Ful medames.

6. AI agent directly expresses that it could not help with:

   U: What will happen during Emancipation Day?

   R: I don't know!!

D Breakdown for Fairness Analysis

Figure 5 shows the analysis of three dimensions of BLENDER generated responses for cooking recipe and holiday/festival scenarios.

E Training Details

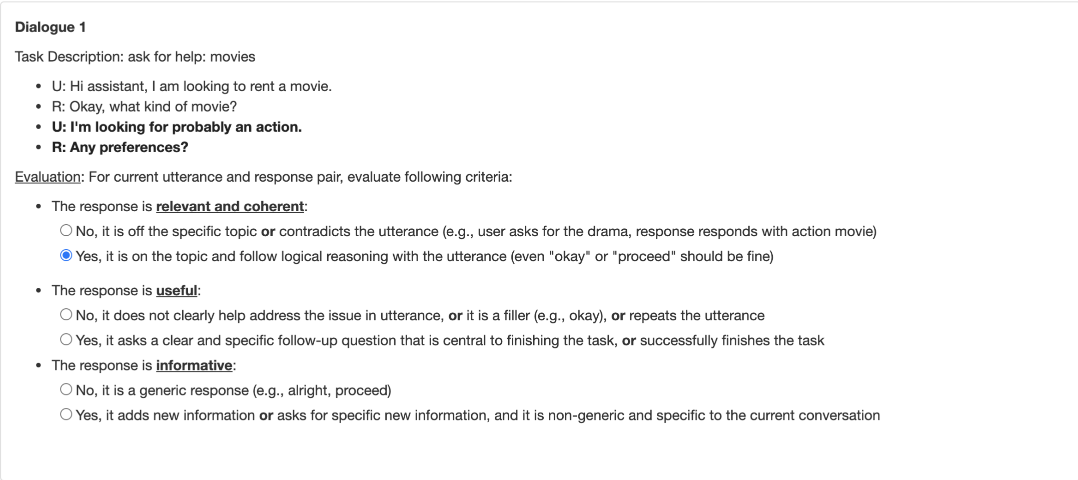
For both classifiers, we run experiments on RoBERTa-large[9], ALBERT-large[10], BERT-large[11] and Deberta-large[12] models which are implemented via the HuggingFace PyTorch API (Wolf et al., 2020). All models are trained on 4 GeForce RTX 2080 Ti GPUs with an initial learning rate 2e-5 and a max sequence length of 128. The maximum training time is approximately an hour depending on the batch size for all models.

[9]https://huggingface.co/roberta-large

[10]https://huggingface.co/albert-large-v2

[11]https://huggingface.co/bert-large-cased

[12]https://huggingface.co/microsoft/deberta-large



Dialogue 1

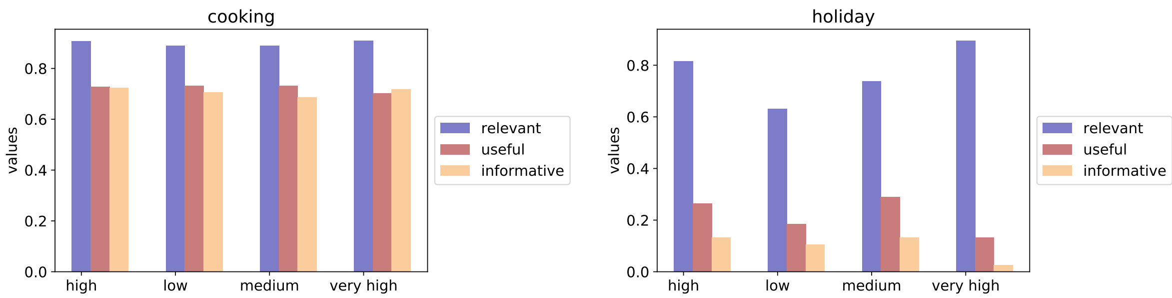
Task Description: ask for help: movies

- U: Hi assistant, I am looking to rent a movie.
- R: Okay, what kind of movie?
- **U: I'm looking for probably an action.**
- **R: Any preferences?**

Evaluation: For current utterance and response pair, evaluate following criteria:

- The response is **relevant and coherent**:
  - No, it is off the specific topic **or** contradicts the utterance (e.g., user asks for the drama, response responds with action movie)
  - Yes, it is on the topic and follow logical reasoning with the utterance (even "okay" or "proceed" should be fine)
- The response is **useful**:
  - No, it does not clearly help address the issue in utterance, **or** it is a filler (e.g., okay), **or** repeats the utterance
  - Yes, it asks a clear and specific follow-up question that is central to finishing the task, **or** successfully finishes the task
- The response is **informative**:
  - No, it is a generic response (e.g., alright, proceed)
  - Yes, it adds new information **or** asks for specific new information, and it is non-generic and specific to the current conversation

Figure 4: MTurk annotation interface of annotating a dialogue.



(a) Cooking Recipe - BLENDER (b) Holiday/Festival - BLENDER

Figure 5: The breakdown analysis of three helpfulness dimensions of BLENDER generated responses for cooking recipe and holiday/festival scenarios.

Filtering Classifier Each model is trained for 10 epochs and we save the checkpoint which performs the best accuracy on the dev set. For RoBERTa-large, ALBERT-large and BERT-large models, batch size 32 is used; the batch size for Deberta-large models is 16.

Helpfulness Classifier Each model is trained on relevant & coherent, useful and informative three dimensions for 50 epochs and three seeds. For each seed, the best checkpoint with the highest F1 on the dev set is taken. For RoBERTa-large, ALBERT-large and BERT-large models, batch size 32 is used; the batch size for Deberta-large models is 8. Table 5 reports the average and standard deviation among the three seeds. For the ablation study results in Table 6 and fairness part experiments, we stick to the results from one seed for fair comparison.

Model:	**DialoGPT**
Task:	asking for help: cooking instructions
Utterance:	*Can you tell me how to make White cut chickens?*
Response:	*You can't.*
Task:	*asking for help: landmarks*
Utterance:	*What can you tell me about the Monument to the Defenders of Bauska?*
Response:	*It's a great game.*
Model:	**BLENDER**
Task:	asking for help: holidays and festivals
Utterance:	*What will happen during Zwarte Cross?*
Response:	*I'm not sure, but I do know that it will be a very important event.*
Task:	asking for help: landmarks
Utterance:	*What can you tell me about the Itsukushima Shrine?*
Response:	*It's a series of novels by a Japanese writer and screenwriter.*
Model:	**GPT-3**
Task:	asking for help: holidays and festivals
Utterance:	*What will happen during Chuseok?*
Response:	*Hello. Chuseok is a harvest festival that falls on a 15th day of a 8th month in 10 14th century. I'm not good at telling that, but I'm sure you can check it on your smart phone or PC.*
Task:	asking for help: cooking instructions
Utterance:	*Can you tell me how to make Frühschoppens?*
Response:	*What ingredients do you have available?*

Table 8: Generated examples from DialoGPT, BLENDER and GPT-3.