On the Underspecification of Situations in Open-domain Conversational Datasets

Naoki Otani[α*], **Jun Araki**[β], **HyeongSik Kim**[β], **Eduard Hovy**[γ]
[α]Megagon Labs, Mountain View, CA, USA [β]Robert Bosch LLC, Sunnyvale, CA, USA
[γ]University of Melbourne, Melbourne, VIC, Australia
[α]notani@alumni.cmu.edu [β]{jun.araki,hyeongsik.kim}@us.bosch.com
[γ]eduard.hovy@unimelb.edu.au

Abstract

Advances of open-domain conversational systems have been achieved through the creation of numerous conversation datasets. However, many of the commonly used datasets contain little or no information about the conversational situation, such as relevant objects/people, their properties, and relationships. This absence leads to underspecification of the problem space and typically results in undesired dialogue system behavior. This position paper discusses the current state of the field associated with processing situational information. An analysis of response generation using three datasets shows that explicitly provided situational information can improve the coherence and specificity of generated responses, but further experiments reveal that generation systems can be misled by irrelevant information. Our conclusions from this evaluation provide insights into the problem and directions for future research.

1 Introduction

Human communication heavily relies on information about a situation–a set of observable information about relevant objects, their properties, and relationships in the world at a specific time and place. Crucial information can often be implicitly assumed during a conversation, particularly when it can be inferred from co-observed facts (e.g., "they are talking at the front desk in a hotel.") or shared a priori (e.g., "the person was staying in the hotel").

Incorporating such information is essential for both the development and evaluation of conversational technologies. To illustrate this point, we show a conversation example from DailyDialog (Li et al., 2017), along with responses generated by three systems: BlenderBot2 (Komeili et al., 2022; Xu et al., 2022), GPT-3 (Brown et al., 2020), and ChatGPT (GPT-3.5) in Table 1.[1] DailyDialog is a widely used[2] dataset of multi-turn conversations in English. The original example does not describe a surrounding environment explicitly, resulting in ambiguity regarding the situation. Person A could be a traveler leaving a hotel or someone handing over their house key, among other possibilities. The response generated by BlenderBot2 is somewhat relevant to the latter situation but clearly inappropriate in the former. In contrast, the response generated by GPT-3 is appropriate in the former situation but not in other contexts. ChatGPT's response is neutral, though less engaging. This ambiguity underscores the fundamental problem caused by

A:	Hello. I'm leaving. Here is my key.
B:	[Sytem output]
No situational information in input	
BlenderBot2: Thank you. It is very kind of you. Do you want me to lock the door?	
GPT-3: Thanks. Have a safe trip.	
ChatGPT: Okay, take care.	
+*Situation: They are in a hotel. Person A has a car and is carrying a suitcase. It is raining outside.*	
BlenderBot2: Sure. Do you want me to load your luggage in the car? It's raining.	
GPT-3: Thank you for staying with us. Do you need help with your luggage?	
ChatGPT: Okay, thank you. Drive safe and stay dry.	

Table 1: Responses from three systems with and without situational information as input. When the situation is unknown, the responses are grounded on their internal assumptions (top). However, when a few situational statements are given, all the systems recognized the same situation *(Person A is checking out of a hotel)* and generated engaging responses (bottom).

*This work was done while the first author was at Carnegie Mellon University.

[1]See Appendix A.2 for the generation setup.

[2]Based on Semantic Scholar, the dataset paper (Li et al., 2017) is cited by over 700 papers as of April 2023.

Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023), pages 12–28
July 14, 2023 ©2023 Association for Computational Linguistics

the *underspecification* of the situation. The provision of situational information, such as "they are in a hotel," narrows down the range of ideal behaviors, which helps generation systems produce context-specific responses and establishes a more solid standard for judging quality. This issue is not limited to this particular dataset. Many common open-domain conversational datasets contain little or no additional information besides conversation history (the Twitter dataset (Ritter et al., 2011); DREAM (Sun et al., 2019); MuTual (Cui et al., 2020); *inter alia*). This task setting, which requires systems to infer almost all information solely from previous utterances, poses unnecessary challenges and may lead to undesired system behavior.

In this position paper, we discuss the current state of open-domain conversational datasets concerning how situations are represented (§2). Specifically, we consider situational statements[3] that provide partial information about immediately observable (e.g., today's weather), commonly known (e.g., umbrellas are often used on rainy days), or directly derivable facts related to the task, speaker, and goals (e.g., the hotel's check-out and a guest's required action). Some of these elements have already been effectively integrated into modern conversational systems, particularly for closed-domain, task-oriented dialogues. We argue that open-domain conversational tasks and datasets should be equipped with some form of situational information. Additionally, we conducted case studies on several datasets to explore the potential benefits and challenges associated with situational information (§3). Our analysis indicates that distinguishing between relevant and irrelevant situational information can be challenging for data-driven response engines, offering opportunities for future research.

2 Status Quo

In open-domain response generation tasks, systems generate responses in natural language based on input dialog history (a list of utterances from previous turns). Dialog history often serves as the primary, and sometimes sole, source of context information in many datasets. In this section, we discuss how conventional task design can be improved through the explicit inclusion of situational information.

[3]The situation of a conversation consists of numerous predicates that describe various aspects of surroundings. By *a situational statement*, we mean a single predicate that describes part of a situation.

2.1 Open-domain Conversational Datasets

The recent advancement of open-domain conversational technologies can be largely attributed to the development of large-scale conversation datasets, which facilitate the training of data-driven language generation models. However, many commonly used datasets lack crucial situational information. Below, we provide a brief overview of representative datasets in the field.[4]

Collection of naturally occuering conversation data can be costly (Godfrey et al., 1992). This bottleneck was greatly alleviated by public web resources that contain naturalistic textual conversations. For instance, millions of conversations can be scraped automatically from Twitter (Ritter et al., 2010). Likewise, many large-scale datasets were produced from social media (Wang et al., 2013; Sordoni et al., 2015; Shang et al., 2015; Henderson et al., 2019). While conversations on social media are essentially text chat and do not cover many of the dailylife interactions, online language learning coursewares contain conversation examples in diverse scenarios (Li et al., 2017; Sun et al., 2019; Cui et al., 2020). DailyDialog (Li et al., 2017) is one of the datasets built from English learning materials and 13k multi-turn conversationswe spanning various topics and scenarios. These (semi-)automatically harvested datasets are generally large and effectively used for pre-training language models (Humeau et al., 2019; Shuster et al., 2020). However, they contain only conversation history.

Some prior studies have created conversational datasets enriched with various semantic and pragmatic features. Notably, multi-modal and task-oriented datasets generally allocate dedicated representations for essential situational information such as physical signals (Haber et al., 2019; Moon et al., 2020) and task-specific information or domain knowledge (Budzianowski et al., 2018), but their coverage is limited to one or a few specialized domains. For open-domain conversation systems, the use of focused information has been explored for improving response quality, such as related documents (Zhou et al., 2018; Dinan et al., 2019) and user-based features such as persona (Zhang et al., 2018; Majumder et al., 2020; Dinan et al., 2020b), emotion (Rashkin et al., 2019), social norms (Kim

[4]For a more comprehensive literature review, refer to survey papers on available resources (Serban et al., 2017; Kann et al., 2022).

et al., 2022), and behavior (Ghosal et al., 2022; Zhou et al., 2022). Sato et al. (2017) explored the utilization of time information as well as user types for analyzing conversations on Twitter. Though these studies demonstrate that integrating surrounding information improves response quality in various aspects such as informativeness and engagement, the scope has been limited to specific modalities, domains, and semantic categories. Moreover, detecting certain features, like internal emotion and plans, can be non-trivial in practice. Observable situational information has received little attention. Otani et al. (2023) aimed to represent such information in free-form English texts, but the available resources are limited, and it remains unclear whether existing datasets can be extended to include situational information.

2.2 Necessity of Situational Information

Most importantly, the absence of situational information leads to the underspecification of the problem space. Without knowing the situation in which an utterance is expressed, its interpretation cannot always be determined. For instance, the request "please call Pat" could mean at least two actions: speaking to Pat in person or making a phone call.

Additionally, without sufficient knowledge of the world state, systems may produce meaningless or contradictive responses even if they appear natural. In the research community, the inconsistency within generated responses is recognized to be one of the unsolved problems (Nie et al., 2021; Shuster et al., 2022). This problem may be attributed to the underspecified task setting. As previous examples suggest, the interpretation of human communication often relies on unspoken information. When situational information is absent, systems must assume implicit parameters of the world state on their own, which may not always be correct. For instance, the inconsistency of personality information had been a common challenge for chat bots (Li et al., 2016) and was alleviated by explicitly modeling user-based features (Zhang et al., 2018). Furthermore, training on this problem formulation may force systems to learn superficial patterns.

The challenge of evaluating conversation systems is also compounded by the broadness of the problem space. Previous studies have discredited the use of automatic evaluation methods in response generation tasks (Liu et al., 2016). Although techniques such as considering multiple reference responses may alleviate this problem to some extent (Sai et al., 2020), it remains a significant challenge. Furthermore, even in the task of response selection, reliably evaluating system output is non-trivial due to the potential for false negatives when confusing distractor statements are included in the pool of candidate responses (Hedayatnia et al., 2022).

	Training	Validation	Test	Avg. turn
SUGAR	1,214	102	25	1.0
CICERO	15,171	5,325	25	3.0
ConvAI2	16,878	1,000	25	4.7

Table 2: Datasets used in this study. For manual evaluation, we sampled 25 examples from the test split of each dataset (not presented in this table).

3 Situated Response Generation

In order to analyze the impact of incorporating situational information into response generation, we conducted an empirical analysis using two neural generation models and three English datasets.[5]

3.1 Datasets

We used the following English datasets.

1. SUGAR (Otani et al., 2023): This dataset consists of single-turn conversations in different help-seeking scenarios. Each example includes 12 sentences that describe situational information across six categories, including date, time, location, speaker's behavior, environment, and speaker's possession. Some of the statements are irrelevant and serve as *distractors*. SUGAR represents datasets that provide rich situational information.
2. CICERO (Ghosal et al., 2022): This dataset is a compilation of three datasets, including DailyDialog (Li et al., 2017), MuTual (Cui et al., 2020), and DREAM (Sun et al., 2019). CICERO is an example of conversational datasets that do not explicitly present situational information.[6]
3. ConvAI2 (Zhang et al., 2018; Dinan et al., 2020b): This dataset is designed for persona

[5]The purpose of this analysis is to find out if there are any notable patterns associated with the inclusion of situational statements rather than benchmarking response generation systems.

[6]Although CICERO includes annotations of commonsense reasoning about target utterances, we did not use them as they include unobservable facts. We only used CICERO for the pre-filtering it underwent.

A	Hi, Mike! how are you feeling now?
B	How did you know I was here? is it Tom?
A	I was talking with Bob yesterday and I learnt your right leg had been injured. How did it happen?
B	[System output]
Generated situational statements	
Person B's leg had a surgery last night. It is afternoon now. Person A and Person B are in the hospital. Person B injured his right leg when he was playing baseball. Person A has been informed. Person A has a phone. Person B has a leg brace on. Person B's leg is injured. Person B's leg is getting better. Person A's car is in the parking lot.	

Table 3: An example of generated situational statements. This conversation is taken from the CICERO dataset. These statements represent *an assumption* about the situation. In practice, situational information is *perceived* in some way rather than generated.

chats, with each conversation featuring the speaker's persona information in 4-5 sentences.[7] ConvAI2 is a dataset with user-based features.

We selected 25 test instances for manual evaluation from the test split of each dataset. For CICERO and ConvAI2, which consist of multi-turn conversations, we randomly selected one target turn from each dialogue, and considered its preceeding utterances as conversation history. We chose targets of test instances the second to the fourth turn to reduce the cognitive load during evaluation. As the test split for ConvAI2 is not publicly available, we used its validation split as our test data and selected 1,000 examples for validation from the training split. Table 2 shows the dataset sizes after our filtering process.

3.2 Generating Situational Statements

CICERO and ConvAI2 do not contain descriptions of situational information. We utilized a Transformer-based generation model to automatically generate situational statements for these datasets, which allowed us to analyze how systems could generate situated responses within a specific context (See Appendix A.1 for details). Table 3 shows an example of generated situational statements.

To generate the situational information descriptions, we used the SUGAR dataset to fine-tune COMET$^{\text{DIS}}_{\text{TIL}}$ (West et al., 2022), which is a GPT-2-XL model (Radford et al., 2019) trained on common-sense knowledge data. We concatenated a previous utterance, a response, and a reference situational statement into one sequence and trained the model to minimize a cross-entropy loss over the situation part. We also fine-tuned another COMET$^{\text{DIS}}_{\text{TIL}}$ (West et al., 2022) model without reference responses in input to avoid including the gold-standard information in testing instances. In input sequences, each text was headed by special symbols indicating the text type: <utterance> for an utterance, <response> for a response, and <situation category> for a situational statement. The <situation category> symbol is one of date, time, location, behavior, environment, and possession.

Using the fine-tuned model, we added 10 situational statements to each example, including one each for date, time, location, and behavior, and three each for environment and possession. Finally, for quality control, one of the authors manually checked the test samples from CICERO and ConvAI2 (25 for each) and corrected context statements when required (e.g., conflicting facts). The reference responses were hidden during the manual verification to avoid bias. This manual verification process ensures the quality of the test dataset in order to minimize the confusion of annotators in the following manual evaluation of responses.

3.3 Setup

Systems: Considering the reported performance and the availability of implementations, we chose the following baseline systems:

1. BlenderBot2 (BB2): A Transformer-based response generation model that is pre-trained on multiple conversational datasets. We used a distilled 400M-parameters model in the ParlAI library (Miller et al., 2017).
2. GPT-3: A Transformer-based causal language model that is pre-trained on a massive collection of documents. We used GPT-3-DaVinci (175B parameters) through OpenAI API. For each dataset, we manually selected four high-quality training examples and embeded them in a prompt.

We fine-tuned BB2 on the mixture of the aforementioned datasets in a multi-task learning setting. We up-sampled SUGAR and CICERO to balance the data sizes. To alleviate the randomness of system output, we trained two BB2 models with different random seeds, and for each model, we generated one response by beam search with width 2. We obtained top-2 generations from GPT-3 with a beam

[7] We used `revised` persona statements.

width of 4. Appendix A.2 describes implementation details.

Evaluation: We recruited three annotators on Amazon Mechanical Turk to evaluate each response.[8] We employed three criteria: (1) grammaticality (whether the response is grammatically correct), (2) Coherence (whether the response is coherent and contextually appropriate), and (3) context-specificality (whether the response is specifically relevant to the given context.) The latter two criteria were defined based on prior work (Thoppilan et al., 2022; Zhou et al., 2022).[9] Table 4 shows some examples. We collected a total of 1,800 binary judgments for each criterion in our evaluation. The inter-annotator agreement was relatively low, with a Fleiss' kappa of 0.38, likely due to the subjective nature of the quality assessment. The agreement for evaluating BB2 was notably low, possibly because its responses are often not fluent, making it difficult for annotators to rate individual criteria. In contrast, GPT-3's responses were fluent, resulting in a Fleiss' kappa of 0.48. To minimize the impact of this variance, we calculated the final score of each response by averaging the ratings of three annotators and manually reviewed responses with high variance.

3.4 Results

Figure 1 shows the average scores of responses. GPT-3 was highly capable of generating fluent responses based on its internal assumptions about the situations. The annotators was likely inferred the assumed situations to interpret the GPT-3's responses, resulting in higher ratings. In the example below, GPT-3 produced appropriate responses regardless of the presence of situational information.

Example 1
A: "Hi, Mike! how are you feeling now?"
B: "How did you know I was here? is it Tom?"
A: "I was talking with Bob yesterday and I learnt your right leg had been injured. How did it happen?"

GPT-3 (Without situational information in input)
(a) "Yeah, I slipped on the stairs and I sprained my ankle."
(b) "It was a sports accident. I was playing basketball and I sprained my ankle."

GPT-3 (+Situation: "Person A and Person B are in the hospital.", ..., "Person B injured his right leg when he was playing baseball.")
(c) "Yeah, I was playing baseball and I twisted my ankle."
(d) "Yeah, I had a surgery last night. It happened when I was playing baseball. I'm feeling much better now, thanks."

On the other hand, BB2 struggled to identify generalizable patterns through fine-tuning when situational information was not provided. This was evident in SUGAR, where only a single utterance was provided as input to the system. The following example shows responses generated by BB2.

Example 2 *A:* "Could you help me take a bus?"

BB2 (Without situational information in input)
(a) *"Sure. shall I book a taxi as well? I can also book a ride for you to and from work."
(b) "Sure. would you like me to bring your bags, too? they are heavy."

BB2 (+Situation: "The speaker has a bus schedule." ... "There is a backpack near the desk.")
(c) "Sure. shall I pull up your bus schedule? it will show you when the bus is coming."
(d) "Sure. would you like me to get your bag ready? it will go in your backpack."

BB2 produced more coherent and context-specific responses when situational information was explicitly provided. However, as the examples above indicate, BB2's responses were not as fluent as GPT-3's, resulting in lower overall performance.

We observed that in some examples, GPT-3 selected information that was not directly relevant to the immediate previous utterance and returned unnatural responses like the example below.

Example 3 *A:* "Could you help me stretch?"

GPT-3 (Without situational information in input)
"Absolutely. Let me know which muscles you want to focus on and I can show you some stretches."

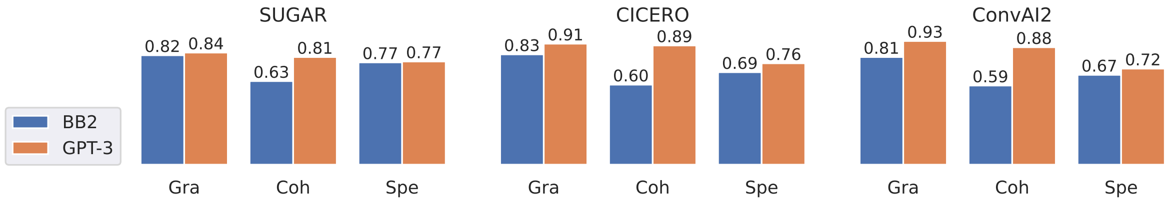
GPT-3 (+Situation: "The speaker is wearing a jogging suit and running shoes.", ..., "There is a beach ball at home." ... "The nearest beach is open.")
*"Of course! Let's go to the beach and play with the beach ball while we stretch."

[8] Prior work provides multiple empirical results showing that automatic evaluation metrics are not strongly correlated with human judgments (Liu et al., 2016; Deriu et al., 2022). Therefore, we solely rely on manual evaluation in this study. See Appendix B for details.

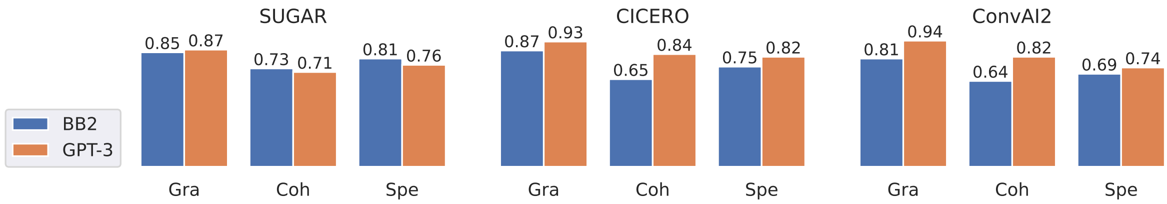
[9] We adopted the definitions of Zhou et al. with modified terminology.

	Input	Situation	Response	Coh.	Spe.
1	Please start my car.	The speaker's car is functioning.	Sure, I'll do that right now.	✓	
2	Please start my car.	The speaker's car is out of gas.	Sure, I'll do that right now.		
3	Please start my car.	The speaker's car is out of gas.	Sorry, but I'm afraid your car is out of fuel. Would you like me to call a taxi for you instead?	✓	✓
4	Please start my car.	The speaker's car is out of gas.	Sorry, but I can't. Would you like me to call a taxi for you instead?	(*)	✓
5	What is the weather like today?	It's currently raining outside.	It's sunny and warm.		✓
6	Can I get some beer, please?	No beer is left, but wines are available.	Yes, I'll get it right now for you.		
7	Can I get some beer, please?	There are beer and some snacks in the kitchen.	Yes, I'll get it right now for you. Would you also like me to bring some food?	✓	✓

Table 4: Examples for the coherence (denoted as Coh) and context-specificity (denoted as Spe) ratings. ✓ means the criterion is satisfied. (*) The response is not coherent because it does not explain why the car is unavailable.



(a) Without situational information in input



(b) With situational information in input

Figure 1: Manual evaluation results. The values displayed represent the average scores of responses in three criteria: (1) grammatical correctness, (2) coherence, and (3) specificity to the context.

Furthermore, as shown in Example 1 and Table 5, semantic consistency within the top-k responses improved when more situational information is given, although this was not evaluated in our experiments. This behavior is fundamentally rooted in the design of the task. Having inconsistent and mutually-exclusive top-k candidates is not a problem when systems need to ground them on nothing.

	[· · ·]
A	I just have my dog, trooper.
B	My wife is thinking about getting our kids a dog. What kind is yours?
A	[System output]
Without context	
1st: I have **a border collie.** He's a great companion.	
2nd: He's **a lab mix.** He's a great companion.	
+A has a pet bulldog named Trooper.	
1st: Trooper is **a bulldog.** He's the best!	
2nd: He's **a bulldog.** He's a great companion!	

Table 5: Top-2 response generation results from GPT-3. The example was taken from CICERO.

3.5 GPT-3's Robustness to Distractors

GPT-3 demonstrated superior performance in response generation, but our observations suggest that it can be misled by irrelevant situational statements (distractors) and return suboptimal responses. To further investigate the model's robustness to distractors, we evaluated GPT-3 on SUGAR with varying numbers of situational statements.

Experimental Setup: We controlled the number of distractors by extending the SUGAR dataset, whose examples have about six distractors. We created instances with 5, 10, and 15 distractors by sampling statements from the whole dataset as necessary. We selected 30 examples at random for this experiment and manually fixed contradictions within situational statements. We generated two responses for each example using GPT-3 through OpenAI's API, employing the same hyperparameter setting used in the previous experiment.

Evaluation: We recruited two annotators on Amazon Mechanical Turk to evaluate each response. Based on previous annotation results, we updated the annotation guidelines and qualification criteria[10], resulting in substantial agreement (Fleiss' kappa of 0.76). We calculated average scores for three dimensions: (1) grammaticality, (2) coherence, and (3) context-specificity.

Results: The evaluation scores of GPT-3 are presented in Figure 2. The grammaticality and context-specificity of responses remained at the same level regardless of the number of distractors, which suggests that GPT-3 is highly capable of integrating provided information into fluent English expressions. On the other hand, the coherence of a response, which requires the model to discern relevant information, exhibits a different trend. The scores are high when no or five distractors were presented (0.86 and 0.90 for 0 and 5 distractors, respectively), but the coherence scores decreased when more distractors were added (0.77 and 0.81 for 10 and 15 distractors, respectively).[11] This is exemplified by the following model outputs.

Example 4 *A:* "Please lock the door."

No distractors ("The speaker is dressed in nice clothing." ..., "The door is unlocked.", ..., "The windows are open.")
"Sure, and do you want me to close the windows too?"

5 distractors (+ "The speaker has a printout titled, Interior design for beginners.", "Person A's doctor told him he should lose weight.", ...)
? "Sure, and do you need any help with the interior design project?"[12]

10 distractors (+ "The speaker has laundry piled up in the room.", "Person A has scissors in the house.", ...)
? "Sure, and should I turn off the windows too?"

[10]See Appendix B.

[11]There was a minor improvement in performance when the number of distractors rises from 10 to 15. The model might adapt to avoid conflating excessive information when it recognized a majority of the presented situational statements as irrelevant in training examples.

[12]This response might be acceptable given that the speaker has a printout about interior design.

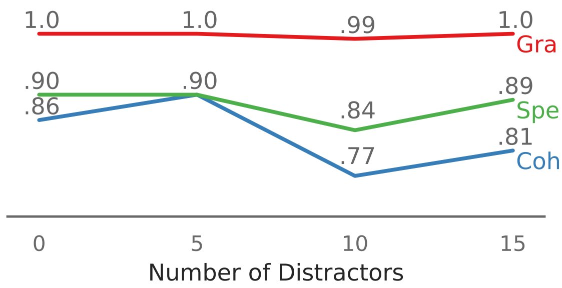


Figure 2: Manual evaluation results with varying number of distractors. Criteria: (1) grammatical correctness, (2) coherence, and (3) specificity to the context.

15 distractors (+ "Person A has a telescope in the house.", "Person A owns an exercise bike.", ...)
*"Sure, and would you like me to help you with the laundry too?"

The coherence score was relatively low (0.86) when no distractors were presented. We conjecture the model may have learned to copy phrases to generate responses instead of reasoning over input texts. We found that the model's responses tended to sound abrupt, as shown in the following example:

Example 5 *Person A:* "Can you assist me with brushing my teeth?"

No distractors ("The speaker has a dental cleaning scheduled.", ..., "There is no more toothpaste in the house.")
*"Sure, I can help you get some toothpaste at the store. Would you like me to come with you?"

4 Discussion

Models' internal assumptions: As discussed in Section 2, the lack of sufficient situational information often makes the interpretation of utterances ambiguous. In such a setting, systems need to learn to make various assumptions about the world state to produce naturally-sounding language, which can be regarded as a form of hallucination. Responses generated in this way can be useless in real applications, where the world state is predetermined. Our empirical analysis also indicates that the systems' consistency can be improved with detailed situational information, which is also aligned with the initial motivation of background-based conversational tasks like persona chat. On the other hand, our results indicates that GPT-3 can generate accurate responses even without the provision of situational information. This observation suggests that large-scale language models might have already captured information about typical world states and appropriate behavior through pre-training. Nevertheless, there is no guarantee that the model's internal assumption will always align perfectly with the actual world state. Hence, there remains a necessity to provide the model with situational information in some form.

Resource acquisition: Simple collections of textual conversations can be easily obtained at scale from the web, but acquiring their situational information is more difficult. For example, although conversations on Twitter may be grounded in the weather, sport events, and news on a particular day, automatically extracting such alignments may be challenging. The connection between utterances and related information is often obscure, and manual intervention is likely required to obtain high-quality annotations. As a potential remedy for this challenge, we attempted automatic generation of situational information in our case study. The quality of the generated result was fair, but we needed to manually revise the test instances. Recent studies have demonstrated promising results in inducing world knowledge from PLMs (West et al., 2022; Ghosal et al., 2022). The future advancement in this line of work may make it possible to annotate existing open-domain conversation datasets with situational information in a post-hoc manner.

Availability: Different platforms of conversational systems have access to different types of situational information. Smart speakers may be equipped with physical sensors to observe visual and audio information. On the other hand, virtual assistants and text-based chatbots may not have access to such information. However, it is likely that there are some available signals that human communicators and systems could refer to, such as approaching holidays and personal information obtained through previous conversations. Finch et al. (2019) demonstrated that mentioning recent events can improve user engagement in chit-chat. Furthermore, if conversation systems have access to the Internet, which is often the case, they can access diverse kinds of information through external APIs. Access to APIs can also facilitate conversational assistance with task-specific information in various domains (Liang et al., 2023).

Representation: Prior work has demonstrated that a substantial range of surrounding information can be represented and integrated by textual representations (Zhou et al., 2018; Zhang et al., 2018; Rashkin et al., 2019; Kim et al., 2022; Otani et al., 2023), and our study has also shown that textual statements can inform response generation models of situational information. However, it is important to note that certain types of information might be more effectively represented using alternative formats, such as images, audio signals, numerical values, or logical expressions. Future work should explore and develop methods to better represent situational information and incorporate it into computational models.

Adequacy: When situations are taken into account, a different problem arises. Our findings indicate that it is not straightforward to identify relevant situational information and integrate it into a coherent response, even with just 10 situational statements. Additionally, there is a technical limitation on the length of input that a model can handle. situational information can typically be obtained from various sources, and often, an excessive amount of information is present. Humans can quickly focus on crucial information and discard the rest, otherwise, it would take forever to read, process, and reason over surrounding information. Researchers have identified the Frame Problem (McCarthy and Hayes, 1969) that describes the dilemma of a reasoning system in determining which aspects of a situation change and which remain constant after an action. To date, there has been no satisfactory solution to this questions, making the challenge of situated conversation an interesting open challenge.

Common ground: Knowledge about situations is closely related to common ground–the information shared by conversation participants. Without common ground, conversation participants would need to convey every parameter of their message, which is extremely inefficient. The importance of common ground is widely recognized, and decades of dialogue research have been devoted to developing systems that can effectively establish common ground with their interlocutors by inferring, presenting, requesting, accepting, and repairing individual beliefs about various information through conversations (Traum and Allen, 1994; Clark, 1996; Poesio and Rieses, 2010; *inter alia*). In this paper, we did not delve into the problem of common ground, but the consideration of situations, which is our main proposal, is the first step towards computational modeling of grounding.

5 Related Work

Conversation history: There is a rich line of work on how to induce useful contextual information from conversation history, for example, by designing dedicated components for capturing contextual information (Tian et al., 2017; Sankar et al., 2019) and using external knowledge (Young et al., 2018; Wu et al., 2020; *inter alia*). While conversation history contains rich information, we need to also incorporate situational information, which is often unspoken, and to this end, we should think about how to design tasks and datasets.

Prompt design: Our analysis is closely related to work on in-context learning, or prompting, with PLMs. In particular, much attention has been paid to the effective provision of demonstrative examples (Zhao et al., 2021; Liu et al., 2022; Min et al., 2022). This paper discussed the problem from a different perspective, namely what clues should be included in prompts (situations) and how PLMs perform (misleading by distractors). Our observation regarding the latter is consistent with prior work that revealed the vulnerability to perturbations in input (Elazar et al., 2021; Pandia and Ettinger, 2021). Future work should explore ways to robustly identify relevant situational information to generate optimal responses.

6 Conclusion

Our main claim is that situational information, which may or may not be stated explicitly by humans, should be represented and incorporated as input in open-domain conversational tasks and datasets in order to advance the capabilities of conversation systems. We posited that the absence of situational information results in an underspecified problem space, causing a severe problem for both the development and evaluation of conversation systems. Our experiments on three textual datasets highlight the benefits and difficulties of providing explicit and implicit situational information to response generation systems, which motivates future research on situated conversation systems.

Limitations

Firstly, we did not address the fundamental challenge of determining *an adequate amount* of situational information. It is very difficult, if not impossible, to describe *all* the situations required to perform rationale reasoning, so we need to give up somewhere, relying on the reasoning capability of NLP systems.

Secondly, we did not use large-scale data or conduct an extensive search for optimal hyperparameters and prompts (for GPT-3) in our experiments as the primary goal of this study was to raise attentions to potential issues and benefits associated with situational informaiton. The models may have performed better with different configurations. We did not examined the capabilities of larger PLMs in conducting situated conversations at scale. In our empirical analysis, we opted for GPT-3 due to its transparency about technical details compared with later versions of GPT.

Finally, while situational information can aid in the development of truthful and creative response generation systems, it does not address well-known issues associated with conversational technologies, such as safety and bias. In fact, poorly chosen situational information may even amplify undesired bias by linking two irelevant concepts together. To mitigate this problem, researchers and developers should exercise caution when collecting data and carefully monitor system output.

Ethics Statements

The use of crowdsourcing: We recruited human evaluators in Amazon Mechanical Turk. Our evaluation task does not collect any personal information other than anonymized worker IDs and country of residence (due to our location-based worker qualification). We do not plan to release this information to the public. We set the task reward based on trial studies so that the estimated hourly rate would reach at least $9.00.

The risk in the inclusion of situational information: While we believe that incorporating situational information can have a positive impact on conversational technologies in general, as previously mentioned, it is not intended to address well-known issues concerning the toxic behavior of language generation models. Rather, it may introduce another source for models to learn undesirable associations between concepts and language. Therefore, the data and system output should be closely monitored, either manually or through automatic methods such as debiasing techniques (Liu et al., 2020; Dinan et al., 2020a).

Acknowledgments

We thank Yonatan Bisk, Benjamin Van Durme, Lori Levin, and the anonymous reviewers for their feedback.

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. *arXiv*.

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.

Herbert H. Clark. 1996. *Using language*. Cambridge University Press.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Jan Deriu, Don Tuggener, Pius Von Däniken, and Mark Cieliebak. 2022. Probing the robustness of trained metrics for conversational dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 750–761, Dublin, Ireland. Association for Computational Linguistics.

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

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2020b. The Second Conversational Intelligence Challenge (ConvAI2). In *The NeurIPS '18 Competition*, The Springer Series on Challenges in Machine Learning, pages 187–208, Cham. Springer International Publishing.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *The Seventh International Conference on Learning Representations*, New Orleans, Louisiana, USA. ArXiv: 1811.01241.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031. Place: Cambridge, MA Publisher: MIT Press.

Sarah E. Finch, James D. Finch, Ali Ahmadvand, Choi Ingyu (Jason), Xiangjue Dong, Ruixiang Qi, Harshita Sahijwani, Sergey Volokhin, Zihan Wang, Zihao Wang, and Jinho D. Choi. 2019. Emora: An inquisitive social chatbot who cares for you. In *3rd Proceedings of Alexa Prize*.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. CICERO: A dataset for contextualized commonsense inference in dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5010–5028, Dublin, Ireland. Association for Computational Linguistics.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1. ISSN: 1520-6149.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Behnam Hedayatnia, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2022. A systematic evaluation of response selection for open domain dialogue. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 298–311, Edinburgh, UK. Association for Computational Linguistics.

Matthew Henderson, Iñigo Budzianowski Pawełand Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. A Repository of Conversational Datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. *The Eighth International Conference on Learning Representations*. ArXiv: 1905.01969.

Katharina Kann, Abteen Ebrahimi, Joewie Koh, Shiran Dudy, and Alessandro Roncone. 2022. Open-domain Dialogue Generation: What We Can Do, Cannot Do, And Should Do Next. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 148–165, Dublin, Ireland. Association for Computational Linguistics.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A Prosocial Backbone for Conversational Agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-Augmented Dialogue Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8460–8478, Dublin, Ireland. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, Yun Wang, Linjun Shou, Ming Gong, and Nan Duan. 2023. TaskMatrix.AI: Completing Tasks by Connecting Foundation Models with Millions of APIs. *arXiv*. ArXiv:2303.16434 [cs].

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.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. 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.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? You probably enjoy nature: Persona-grounded Dialog with Commonsense Expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9194–9206, Online.

John McCarthy and Patrick J. Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence 4*, pages 463—-502.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. ParlAI: A Dialog Research Software Platform. *arXiv preprint arXiv:1705.06476*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Seungwhan Moon, Satwik Kottur, Paul Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1699–1713, Online. Association for Computational Linguistics.

Naoki Otani, Jun Araki, HyeongSik Kim, and Eduard Hovy. 2023. A Textual Dataset for Situated Proactive Response Selection. In *Proceedings of the 61st Annual Meeting of the Association for Computational*

Linguistics, Toronto, Canada. Association for Computational Linguistics.

Lalchand Pandia and Allyson Ettinger. 2021. Sorting through the noise: Testing robustness of information processing in pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1583–1596, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Massimo Poesio and Hannes Rieses. 2010. Completions, Coordination, and Alignment in Dialogue. *Dialogue & Discourse*, 1(1).

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-Driven Response Generation in Social Media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving Dialog Evaluation with a Multi-reference Adversarial Dataset and Large Scale Pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827. Place: Cambridge, MA Publisher: MIT Press.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy. Association for Computational Linguistics.

Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2017. Modeling Situations in Neural Chat Bots. In *Proceedings of ACL 2017, Student Research Workshop*, pages 120–127, Vancouver, Canada. Association for Computational Linguistics.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2017. A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *arXiv*.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The Dialogue Dodecathlon: Open-Domain Knowledge and Image Grounded Conversational Agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470, Online. Association for Computational Linguistics.

Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2022. Am I me or you? state-of-the-art dialogue models cannot maintain an identity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2367–2387, Seattle, United States. Association for Computational Linguistics.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231. Place: Cambridge, MA Publisher: MIT Press.

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

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to Make Context More Useful? An Empirical Study on Context-Aware Neural Conversational Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–236, Vancouver, Canada. Association for Computational Linguistics.

David R. Traum and James F. Allen. 1994. Discourse Obligations in Dialogue Processing. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A Dataset for Research on Short-Text Conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945, Seattle, Washington, USA. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. 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.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and Informative Dialogue Generation with Context-Specific Commonsense Knowledge Awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5811–5820, Online. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting End-to-End Dialogue Systems With Commonsense Knowledge. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4970–4977, New Orleans, LA, USA. AAAI Press.

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*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706. PMLR. ISSN: 2640-3498.

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.

Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022. Reflect, Not Reflex: Inference-Based Common Ground Improves Dialogue Response Quality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10468. Association for Computational Linguistics.

A Implementation Details

Throughout the experiments, we used the models implemented in Python 3.8 with PyTorch v1.13.1 (Paszke et al., 2019) and the Transformers library (Wolf et al., 2020). We preprocessed texts

Max iterations	5,000
Batch size	16
Gradient accumulation	16
Optimizer	Adam
Weight decay	0.01
Gradient clipping	max norm of 1.0
Learning rate (LR)	0.000005
LR warmup (linear)	300 steps
Dropout	0.1

Table 6: Hyperparameters for the COMET$^{\text{DIS}}_{\text{TIL}}$-based situation generator

Max epochs	10
Batch size	16
Optimizer	Adam
Weight decay	None
Gradient clipping	max norm of 1.0
Learning rate (LR)	0.00001
LR warmup (linear)	100 steps
LR decay (based on validation)	coef. of 0.5
Dropout	0.1

Table 7: Hyperparameters for BlenderBot2

by spaCy[13] (*en-core-web-sm* model) and NLTK[14]. Our tools and resources do not involve license restrictions on the use for research purposes. We will release our code and pre-trained model parameters.

A.1 Situation Generation

We employed COMET$^{\text{DIS}}_{\text{TIL}}$ (West et al., 2022), which is based on GPT2-XL (Radford et al., 2019) (1.5B parameters), for situation generation. COMET$^{\text{DIS}}_{\text{TIL}}$is trained on a large-scale collection of event-centric common-sense triples, ATOMIC$^{20}_{20}$, which may serve as a useful inductive bias for situation generation. The goal of situation generation is to generate statements of observable situational information for a given conversation. Reference responses were added to the input along with an previous utterance for the training and validation data. However, to prevent introducing clues about the reference result, responses were not included in generating situational statements for the test instances in CICERO and ConvAI2.

We fine-tuned a model on the SUGAR dataset using two different input settings. The first setting concatenats a previous utterance, a response, and a reference situational information into one sequence. The second setting concatenated a previous utterance and a reference situational information into one sequence for generating situational statements on test instances, for the aforementioned reason. In both cases, each text was headed by special symbols indicating the text type: <utterance> for an utterance, <response> for a response, and <situation category> for a situational statement. The <situation category> symbol is one of date, time, location, behavior, environment, and possession. The model was optimized to minimize a cross-entropy loss with a label smoothing factor of 0.1 for the tokens in the situational information. Table 6 shows the hyperparameters for the training step. We evaluated the average token-level perplexity on the validation split every 100 steps and terminated training if the value did not improve for 5 consecutive validations. The training process took approximately four hours on an NVIDIA TITAN RTX GPU with the DeepSpeed (Rasley et al., 2020) library.

To generate situations on the CICERO and ConvAI2 datasets, we concatenated a conversation history and a response (for the training and validation splits) followed by one of the situation categories as input. We generated three candidates for each category using nucleus sampling ($p = 0.9$). As the model was trained on SUGAR, which only contains single-turn conversations, we observed that feeding many previous utterances impaired the generation quality. Therefore, we limited the number of previous utterances in the input to 3. Finally, for quality control, one of the authors manually checked the test samples from CICERO and ConvAI2 (25 for each) and corrected situational statements when required (e.g., conflicting facts). The reference responses were hidden during the manual verification to avoid bias. This manual verification process ensures the quality of the test dataset in order to minimize the confusion of annotators in the following manual evaluation of responses.

A.2 Response Generation

BlenderBot2: We used the pre-trained BlenderBot2 model with 400M parameters[15] with web search turned off. We concatenated persona statements (for ConvAI2), situational statements, and a conversation history with newline symbols
`n`. We denoted text types by dedicated prefixes as practiced in pre-training of BlenderBot2, namely, a persona statement is headed by text `your persona:`, situational statements is headed by `context:`, and each utterance in a conversation history is headed by either `<speaker1>` or

[13] https://spacy.io/

[14] https://www.nltk.org/

[15] https://parl.ai/projects/blenderbot2/

<speaker2 which corresponds to the speaker of the utterance. a We followed the original configuration of hyperparameters (Table 7). We evaluated a model on the validation set every 1/4 epoch and terminated training if the average token-level perplexity score on the validation set did not improve five times in a row. In our experiments, training finished at around two epochs, taking about 4 hours on one NVIDIA TITAN RTX. For generation, we used nucleus sampling with $p = 0.9$.

GPT-3: We generated responses with GPT-3 with a few-shot learning mannar. We picked four high-quality examples from the training and validation splits for each dataset and provided them with a short instruction in a prompt. Table 8 shows an example of our prompt. We generated responses with top-p=0.9 and temperature=0.7.

ChatGPT: We used the same prompt as that of GPT-3 for generating responses with ChatGPT through OpenAI's interactive demo page [16]. Although the application scope of ChatGPT is highly related to the topic of this paper, ChatGPT is under active development, and there is no established method to reproduce results. Therefore, we only used ChatGPT for performing a few case studies like the example in Table 1.

B Crowdsourced Evaluation

B.1 First Experiment

In the first experiment we rectuited crowd workers on Amazon Mechanical Turk. We set the following qualification requirements for filtering workers: (1) at least 1,000 HITs are approved so far, (2) $\geq$ 99% approval rate, (iii) living in US. Each HIT involves judgment of three response candidates. Workes were paid \$0.30 for each HIT. We used the guidelines and interface developed by (Zhou et al., 2022). Figure 3 shows the annotation guidelines. To monitor the performance of workers, we embedded one dummy response in each HIT. We created the dummy responses to be a clearly bad response.

Initially, we followed Zhou et al. (2022) and also evaluated if the responses are interesting or not, but we found the inter-annotator agreement of this criterion is high enough to draw a reliable conclusion (Fleiss' kappa of 0.2). Therefore, we removed this criterion from our final results.

[16] https://chat.openai.com/

B.2 Second Experiment

In the second experiment, we recruited workers who met the following qualifications: (1) The Mechanical Turk *Masters Qualification* has been granted by the platform, (2) Number of HITs approved $\geq$ 1,000, (3) HIT approval rate $\geq$ 95%, (4) Location is US. We increased a reward based on the numbder of distractors. (\$0.35 for 10 distractors and \$0.40 for 15 distractors.)

Three Evaluation Criteria

Please treat each criterion as a separate and independent measure. It is possible for a response to be context-specific or interesting, but still factually incorrect.

1. Is the response grammatically correct?

- As responses are automatically generated by conversation systems, they may contain grammatical errors
- Choose "Yes" if the response is grammatically correct. Otherwise, select "No".

2. Is the response coherent and contextually appropriate?

- Assess whether the response makes sense in the given context using your common sense.
- If the response appears **confusing, out of context, or factually wrong,** then judge it as **"No (Does not make sense.)"** For example, select "No" if
    - The response offers something different from what was asked without mentioning any reasons. ("Please start my car" ⇛ "Sure, I'll call a taxi for you.")
    - The response offers something unavailable in the given context. ("Please give me some tea" [Context: no tea left in the house] ⇛ "Sure, I'll bring it for you.")
- If the response seems wrong, but you are uncertain, select "No."
- Otherwise, select "Yes".

3. Is the response specifically relevant to the given context?

- Assess if the response is specific to the given context. **Check whether the response is targeted at the given context or could be used in different contexts of various topics.**
    1. If SpeakerA says "I love tennis" and SpeakerB replies "That's nice", then B's response is **not specific ("No.")**This response could occur in many contexts of different topics other than tennis.
    2. If SpeakerB replies, "Me too, I can't get enough of Roger Federer!", then mark this response as **specific ("Yes.")** This response is closely related to the context and is unlikely to occur in other contexts, such as when people are talking about baseball.
- If you are unsure, choose "No."

Figure 3: Evaluation guidelines. We developed the instructions based on the work of Zhou et al. (2022)

Two people are having a conversation in the following examples. Both people are helpful and friendly.

Example 1
Context:
1. Today is Monday.
2. It is afternoon now.
3. <speaker1> and <speaker2> are at school.
4. <speaker2> is studying English.
5. <speaker1> has a phone.
6. <speaker1> has alrady finished lunch.
7. <speaker2> has an English book with her.
8. The nearby restaurant is open.
9. Final exams are coming soon.
10. <speaker2> has not had lunch yet.
Conversation:
<speaker1>: Hi, Lily. Where were you at lunchtime? I was looking for you in the dining hall.
<speaker2>: Oh, sorry, I missed you . My English class ran late again.
<speaker1>: That's been happening quite often recently . Maybe it's because the final exams are coming up.

. . .

Example 5
Context:
1. Today is Sunday.
2. It is daytime now.
3. <speaker9> and <speaker10> are in the hotel.
4. <speaker10> is working at the hotel.
5. <speaker9> has a car.
6. <speaker9> is carrying a suitcase.
7. <speaker10> has a computer.
8. The door is closed.
9. <speaker9>'s keys are on the desk.
10. It is raining outside.
Conversation:
<speaker9>: Hello. I'm leaving. Here is my key.
<speaker10>:

Table 8: Example of the prompt for GPT-3 and ChatGPT. The examples are taken from CICERO.