

ConvoSense: Overcoming Monotonous Commonsense Inferences for Conversational AI

Sarah E. Finch

Department of Computer Science
Emory University
Atlanta, GA, USA
sfillwo@emory.edu

Jinho D. Choi

Department of Computer Science
Emory University
Atlanta, GA, USA
jinho.choi@emory.edu

Abstract

Mastering commonsense understanding and reasoning is a pivotal skill essential for conducting engaging conversations. While there have been several attempts to create datasets that facilitate commonsense inferences in dialogue contexts, existing datasets tend to lack in-depth details, restate information already present in the conversation, and often fail to capture the multifaceted nature of commonsense reasoning. In response to these limitations, we compile a new synthetic dataset for commonsense reasoning in dialogue contexts using GPT, ConvoSense, that boasts greater contextual novelty, offers a higher volume of inferences per example, and substantially enriches the detail conveyed by the inferences. Our dataset contains over 500,000 inferences across 12,000 dialogues with 10 popular inference types, which empowers the training of generative commonsense models for dialogue that are superior in producing plausible inferences with high novelty when compared to models trained on the previous datasets. To the best of our knowledge, ConvoSense is the first of its kind to provide such a multitude of novel inferences at such a large scale.

1 Introduction

Effective dialogue is accomplished by a profound grasp of language and a thorough comprehension of the world. Such comprehension is crucial to the construction of responses that are pertinent, coherent, and captivating within an ongoing dialogue. A pivotal element of this worldview is commonsense: self-evident information that is universally acknowledged among humans (Clark and Brennan, 1991).

Over time, there has been a concerted endeavor to create datasets that facilitate commonsense reasoning. Early work, such as the widely recognized

ConceptNet (Speer et al., 2017), focused predominantly on physical commonsense related to entities. Lately, efforts have shifted toward building datasets encompassing social- and event-based commonsense, such as ATOMIC (Hwang et al., 2021). This new wave of datasets targets complex human concepts, including emotions, desires, and motivations.

As human conversations largely revolve around sharing personal experiences and life events (Fillwock and Traum, 2018; Mitsuda et al., 2019), it is critical for virtual agents to possess a robust understanding of human experiences to conduct effective dialogue. Datasets such as ATOMIC hold promise as they provide insights directly relevant to human experience; however, a drawback lies in their lack of contextual awareness as they hinge on isolated, concise phrases for commonsense inferences. This limitation poses challenges for dialogue-oriented tasks because utterances should not be viewed in isolation but must be interpreted within their context (Pan et al., 2019; Jin et al., 2022).

Several initiatives have recently aimed to curate commonsense inferences tailored for dialogue contexts (Gao et al., 2022; Ghosal et al., 2022; Zhou et al., 2022a). However, a trade-off currently exists between the breadth of inference types covered and the scope of dialogue contexts encompassed within these existing datasets. While some datasets cover a wide range of relations, they are limited to a small number of dialogues (Gao et al., 2022), whereas others capture a large number of dialogues but on a limited set of relations (Ghosal et al., 2022).

In addition, a few challenges can be encountered in these datasets. For example, the inferences in these datasets are often too succinct and derive only straightforward conclusions with minimal elaboration (Gao et al., 2022), which do not convey

implicit commonsense. Some studies instruct annotators to recycle information from the ongoing conversation, undermining the speculative nature of inferences and detracting from the potential of offering fresh insights to enhance dialogue understanding (Ghosal et al., 2022). Moreover, although multiple plausible inferences can be drawn from a single dialogue context, only a few datasets support this multifaceted nature (Shen et al., 2022), impeding the development of models capable of generating diverse inferences, and thus limiting their utility in real applications.

We present *ConvoSense*, a commonsense dataset generated by GPT encompassing 10 popular inference types with over 500,000 inferences across 12,000 dialogues (§4). Our dataset shows greater contextual novelty and enhanced inference diversity and detail while maintaining exceptional reasonability compared to existing datasets (§3). We also explore several strategies to build generative models producing inferences for dialogue contexts (§5). Our experiments show that models trained on *ConvoSense* excel in generating plausible inferences with greater detail and novelty, compared to ones trained on existing datasets (§6). To the best of our knowledge, this is the first dialogue-based commonsense dataset that not only covers an extensive array of inference types at large-scale but also provides a plethora of diverse, novel inferences tailored to each dialogue context. Our *ConvoSense* dataset and inference models can be accessed through our open-source project: <https://github.com/emorynlp/ConvoSense>.

2 Related Work

Recent work has focused on integrating commonsense into various tasks, including story generation and explanation (Guan et al., 2020; Gabriel et al., 2021), dialogue summarization and explanation (Ghosal et al., 2021; Zhou et al., 2021; Kim et al., 2022), and response generation (Li et al., 2022; Sabour et al., 2022; Zhou et al., 2022b). Many of this work relies on existing datasets, such as ConceptNet (Li et al., 2022; Zhou et al., 2022b) and ATOMIC (Sabour et al., 2022), which only contain single-word or short-phrase premises and conclusions. Although there are commonsense datasets curated for long dialogue contexts, they tend to be of small size (Zhou et al., 2022a), express simple inferences (Gao et al., 2022), or

copy context from the provided utterances (Ghosal et al., 2022).

On the other hand, GPT has recently been used to create a variety of datasets. Kim et al. (2023) and Zhan et al. (2023) constructed dyadic dialogue datasets at large-scale, while West et al. (2022) generated commonsense triples in the ATOMIC style (Hwang et al., 2021). However, the ATOMIC-style inferences are not necessarily suitable for dialogue, as they struggle to handle long contexts and often lack depth. Table 1 summarizes the inference types in existing dialogue-focused commonsense datasets and mappings of synonymous types among them. In particular, the following 3 datasets are used for comparisons with our work:

ComFact Gao et al. (2022) mapped dialogue utterances to reasonable inferences from the existing ATOMIC2020 dataset (Hwang et al., 2021) by using exact string matching and embedding similarity. Subsequently, human annotators verified the relevance of the retrieved inferences.

Cicero Human participants were tasked with composing responses to five commonsense questions (e.g., *What is the event that directly causes or could cause Target?*) based on dialogue contexts and explicitly instructed to incorporate information from the preceding or forthcoming utterances. The first version produced a single inference for each example (Ghosal et al., 2022), whereas the second version produced multiple examples of both good and bad inferences (Shen et al., 2022).

Reflect Zhou et al. (2022a) supplied both human-generated commonsense inferences and following utterance responses that could be derived from a specified commonsense inference. The inferences were collected by instructing human participants to answer a commonsense question, while the next-utterance responses were composed by new human participants who were provided with the dialogue context and one of the human-generated inferences.

3 Evaluating GPT-generated Inferences

In order to support the development of a large-scale and high coverage commonsense dataset for dialogue that improves upon existing works, we hypothesize that we can leverage large language models (LLMs) to accomplish this task in an efficient and low-cost manner. From initial pilot tests

Type	Label(s)	Definition(s)	COM	CIC	REF
Subsequent	isBefore	What could happen after this? [2]			
	Subsequent-Events	What subsequent event happens or could happen following the Target? [3]	*	22K	600
		What might happen after? [4]			
Antecedent	isAfter	What could have happened before this? [2]	*		600
		What might have happened before? [4]			
Cause	xReason	What could be the cause of this event? [2]	80	21K	
	Cause	What is the event that directly causes or could cause Target? [3]			
Prerequisite	xNeed	What does X need to do before the event can happen? [1]	1K	10K	
	Prerequisites	What is or could be the prerequisite of Target? [3]			
Motivation	xIntent	Why does X cause the event? [1]	800	12K	
	Motivation	What is an emotion or basic human drive that motivates or could motivate Target? [3]			
Attribute	xAttr	How would X be described? [1] How would you describe Speaker? [4]	400		600
Reaction	xReact	How does X feel after the event? [1]	300		600
		What is Speaker feeling now? [4]			
Reaction _o	oReact	How do others feel after the event? [1]			
		What is the possible emotional reaction of the listener in response to target? [3]	70	6K	600
		What is Responder feeling now? [4]			
Desire	xWant	What would X likely want to do after the event? [1]	1K		
Desire _o	oWant	What would others likely want to do after the event? [1]	100		
Constituents	HasSubEvent	What is a substep that happens within this event? [2]	800		
Obstacle	HinderedBy	What could obstruct the occurrence of this event? [2]	200		
Effect	Causes	What does this event cause to happen? [2]	30		
Effect _s	xEffect	What effect does the event have on X? [1]	400		
Effect _o	oEffect	What effects does the event have on others? [1]	90		

Table 1: The inference types covered in existing commonsense datasets (**COM/CIC/REF**: the numbers of examples in the *ComFact* / combined *Cicero* v1 & v2 / *Reflect* datasets, respectively). Each row denotes a unique type from the existing datasets using definitions from [1] Sap et al. (2019), [2] Hwang et al. (2021), [3] Ghosal et al. (2022), and [4] Zhou et al. (2022a). Counts are truncated to the nearest order of magnitude. * indicates the type was included but no human-verified instances of it are present.

of both closed-source (GPT) and open-sourced LLMs (Vicuna and Llama), we find that GPT provides greater reliability in following specific instructions and produces commonsense inferences of overall better quality than the open-sourced LLMs. Consequently, we choose to rely on GPT in this work.

3.1 Prompt Engineering

Prior to crafting the full *ConvoSense* dataset, we empirically assess GPT’s efficacy in generating reasonable and novel commonsense inferences for dialogue. To mitigate any unintended bias from in-context examples in the GPT prompt, we adopt a zero-shot generation framework.¹ GPT prompts are refined iteratively to achieve the optimal outcomes. An example of the final prompt design,

¹`gpt-turbo-3.5-301` with a temperature setting of 1.0.

specifically tailored for the *Desire* inference type, is illustrated in Table 2.

During our development process, we observe that the inferences generated from GPT frequently contain detailed and rich information, thus addressing one of the major limitations of existing works. In addition, to encourage novel inferences from GPT, we include the instruction “Your answers should provide novel information that is not explicitly shared in the conversation.” as seen in Table 2. We observe that this instruction helps in reducing the redundancy of the generated inferences to the information already explicitly shared in the dialogue context, thus addressing a second major limitation of existing works.

For the prompt, each inference type is paired with a guiding question and an answer prefix, ensuring uniformity in the generated content for the specific type, which respectively fill the *Inference Question (Q)* and *Inference Answer Template (A)*

C	Speaker:	I just finished cleaning up my kitchen and getting the trash out.
	Listener:	I don't envy you. I hate cleaning.
	Speaker:	I'm the other way. I love cleaning, and then seeing my nice clean kitchen afterwards.
T	Target:	I'm the other way. I love cleaning, and then seeing my nice clean kitchen afterwards.
Q	Question:	What does Speaker want to do next?
A	Answer:	As a result, Speaker wants . . .

In a list titled "Answers", generate several likely answers to this question for the target expression, keeping the rest of the conversation in mind.

Your answers should provide novel information that is not explicitly shared in the conversation.

Table 2: A GPT prompt example for the *Desire* inference type. Segments are dynamically modified based on the example and inference type, as highlighted in the gray containers (**C**: dialogue context, **T**: target utterance, **Q**: inference question, **A**: inference answer template).

slots in the prompt. For every dialogue context, a sequence of utterances in the context is placed in the *Dialogue Context (C)* slot, and its final turn gets duplicated in the *Target Utterance (T)* slot. Inally, the GPT output, commencing with the header *Answers* and adopting a list-like format with newline separation, is parsed to extract the generated inferences. Table 3 details the questions and answer prefixes employed for the fifteen identified inference types derived from the previous studies in Table 1.

3.2 Evaluation

To evaluate the quality of GPT-generated commonsense inferences for dialogues, we compare their **reasonability** and **novelty** against inferences from human datasets. First, we sample a uniform distribution over inference types for each existing dataset. For every sample, we then prompt GPT to produce relevant inferences and randomly select one from the generated list. Finally, two human annotators are presented with the dialogue context, inference question, and both the GPT- and human-generated inferences and asked to categorize them for reasonability and novelty. For this evaluation, we enlist native English speakers via the Surge AI crowdsourcing platform (<https://surgehq.ai>) by paying them at a rate of \$0.15 per sample with an estimated completion time of 45 seconds.

Reasonability Most prior commonsense datasets assess their inferences based on human-judged reasonability (Hwang et al., 2021; Ghosal et al., 2022; Shen et al., 2022; Zhou et al., 2022a). An inference is deemed reasonable if it makes sense in, is relevant to, and is consistent with the provided dialogue context. We follow Hwang et al. (2021), in which annotators categorize inferences into levels of the truth likelihood: *always/likely*, *sometimes/possible*, *never/farfetched*, or *invalid/nonsense*.

Novelty A key trait of commonsense for dialogue is its role in enhancing dialogue comprehension by providing relevant contextual information. While Ghosal et al. (2022) gauge creativity in human responses, creativity is not strictly focused on inference novelty. In our study, annotators evaluate the extent to which an inference contributes fresh information to the conversation, categorized as: *new & detailed*, *new & simple*, and *purely repetitive*.

Since we aim to elicit the natural commonsense understanding learned by each annotator through their life experience in our annotation tasks, we do not provide any training or explicit examples towards what constitutes a "reasonable" or "novel" commonsense inference to avoid artificially polluting their commonsense understanding of the world. Instead, we provide a description of the task with definitions of the different categories. Our instructions are intended to mitigate bias towards trivial inference properties by providing clear definitions of the characteristics under study and emphasizing important aspects to keep in mind, such as ignoring grammar errors unless it made an inference nonsensical. Furthermore, decomposing inference quality into two characteristics allows for their independent evaluation. We verified through pilots that this approach resulted in reliable and reasonable annotations from our annotators for both tasks.

3.3 Results

Following Hwang et al. (2021), the two metrics in Section 3.2 are converted into binary representations. Thus, labels [*always/likely*, *sometimes/possible*] are categorized as *positive* and [*never/farfetched*, *invalid/nonsense*] are considered *negative* reasonability. Similarly, [*new & detailed*, *new & simple*] are designated as *positive*, and [*purely repetitive*] is classified as *negative*.

Type	Question	Answer Template
Subsequent	What might happen after what Speaker just said?	After this, . . .
Antecedent	What events happened before the situation that Speaker just shared?	Before this, . . .
Cause	What could have caused the last thing said to happen?	This was caused by. . .
Prerequisite	What prerequisites are required for the last thing said to occur?	For this to happen, it must be true that. . .
Motivation	What is an emotion or human drive that motivates Speaker based on what they just said?	Speaker is motivated. . .
Attribute	What is a likely characteristic of Speaker based on what they just said?	Speaker is. . .
Reaction	How is Speaker feeling after what they just said?	Speaker feels. . .
Reaction_o	How does Listener feel because of what Speaker just said?	Listener feels. . .
Desire	What does Speaker want to do next?	As a result, Speaker wants. . .
Desire_o	What will Listener want to do next based on what Speaker just said?	As a result, Listener wants. . .
Constituents	What is a breakdown of the last thing said into a series of required subevents?	This involves. . .
Obstacle	What would cause the last thing said to be untrue or unsuccessful?	This is untrue or unsuccessful if. . .
Effect	What does the last thing said cause to happen?	This causes. . .
Effect _s	How does the last thing said affect Speaker?	This causes Speaker to. . .
Effect _o	How does the last thing said affect Listener?	This causes Listener to. . .

Table 3: Question and answer prefixes used for generating each inference type from GPT for dialogue contexts. The ten inference types used in our work are represented in gray shading.

novelty. This setup, with 300+ annotated samples per dataset, allows us to detect differences of at least 10% between GPT- and human-generated datasets using McNemar’s binary matched-pairs test at 80% power and a significance level of 0.05, assuming discordance probabilities of 0.24 or lower (compatible with pilots).² In cases of annotator disagreement, one of the annotators’ decisions is randomly selected. To mitigate the potential noise introduced by this random selection, we repeat the process 100 times and report the average result, only confirming statistical significance when every selection yields a significant result.

Considering the reported quality of the existing datasets and our preliminary assessments of GPT-generated inferences, we expect much higher rates of positive classes than negative ones, resulting in a class imbalance. To overcome the vulnerability to prevalence skew exhibited by other agreement metrics like Cohen’s kappa (Jeni et al., 2013; Wongpakaran et al., 2013; Quarfoot and Levine, 2016), Gwet’s AC1 inter-annotator agreement metric is chosen (Gwet, 2002).³ Our annotators obtain AC1 values of 0.8 and 0.6 for reasonability and novelty, respectively, implying substantial agreement.

Table 4 demonstrates that GPT can attain comparable reasonability in its generated inferences as those derived from humans, even exceeding the reasonability of the inferences in ComFact

²<https://homepage.univie.ac.at/robin.ristl/samplesize.php>.

³We observe Cohen’s kappa of 0.19 and 0.15 for reasonability and novelty, respectively.

Dataset	R	N	#
GPT	93 (0.17)*	91 (0.21)*	390
ComFact	81 (0.05)	73 (0.04)	
GPT	93 (0.10)	80 (0.16)*	300
Cicero	88 (0.05)	70 (0.06)	
GPT	89 (0.08)	86 (0.08)	300
Reflect	91 (0.09)	82 (0.04)	

Table 4: The average % ($\sigma < 2\%$) of total samples (#) tested as reasonable (R) and novel (N), with discordance probabilities in parentheses. *: statistical significance (McNemar’s, $\alpha = 0.05$). 90 more samples are used for ComFact due to its greater number of inference types.

with statistical significance. Notably, the results also indicate that GPT *surpasses* the novelty of the human-generated inferences for the majority of the existing datasets. Furthermore, GPT outputs achieve higher detail than that observed from human-generated inferences. Figure 2 shows the percentage of *new & detailed* inferences out of all positive novelty inferences for each data source, clearly demonstrating the superiority of GPT inferences in terms of their expressed detail. Example inferences from GPT and humans are shown in Figure 1.

4 ConvoSense Dataset

Given our assessment of high-quality, novel, and detailed GPT-generated commonsense inferences

Cause	Attribute
<p>Listener: I can't decide whether to go to university or get a job. What's your opinion?</p> <p>Speaker: Well, if I were you, I would go on studying.</p> <p>Listener: But I don't even know what to study.</p> <p>Speaker: I advise you to major in economics.</p>	<p>Speaker: We're all went out for a nice picnic lunch earlier.</p> <p>Listener: Where did you go?</p> <p>Speaker: To the park, the place by the lake.</p>
<p>1. the listener does not know which subject to study.</p>	<p>1. the speaker is old fashioned.</p> <p>2. the speaker is outdoorsy.</p>
<p>1. the speaker's personal interest in economics.</p> <p>2. the speaker's observation of the job market and the high demand for economics graduates.</p> <p>3. the speaker's perception of economics as a versatile and valuable degree that can open up various career paths.</p> <p>4. the speaker's inclination towards quantitative and analytical subjects, which are central to economics.</p> <p>5. the speaker's belief that economics provides a solid foundation for understanding and navigating today's complex global economy.</p>	<p>1. someone who enjoys spending time in nature.</p> <p>2. familiar with the park by the lake and has visited it before.</p> <p>3. someone who values and prioritizes socializing and spending time with loved ones.</p> <p>4. someone who enjoys simple pleasures like picnicking and appreciating the outdoors.</p> <p>5. someone who makes an effort to plan enjoyable activities to share with others.</p>

Figure 1: Cause and Attribute inferences written by humans (top, green) and generated by GPT (bottom, blue).

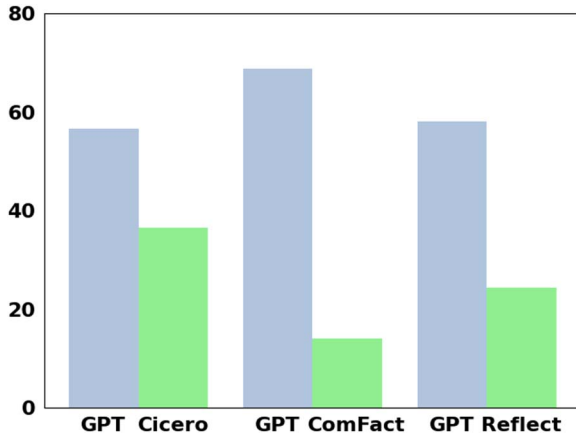


Figure 2: Average % of *new & detailed* inferences out of all positive novelty inferences for each data source.

across various dialogue contexts and inference types (Section 3), we construct a substantial conversational commonsense dataset using GPT, termed ConvoSense.

4.1 HumanGen: Human-generated Datasets

For fair comparisons to our work, we combine the three human-generated datasets (Section 2) into a solitary dataset, termed HumanGen.⁴ Specifically, their train/validation/test sets are integrated independently. For ComFact and Cicero, this integration follows the provided splits, while for Reflect, data is sampled following an 80/10/10 distribution. To standardize HumanGen into a

⁴Many commonsense types have a sparsity of training data when the human-generated datasets are viewed in isolation, which would impede the training of a neural model to adequately capture the commonsense type.

cohesive format, we perform the following pre-processing steps.

First, we leverage the mapping outlined in Table 1 along with the specifications from Table 3 to identify relevant commonsense inference questions for each instance. Then, we combine consecutive utterances from the same speaker to ensure every dialogue turn represents a distinct speaker. Lastly, we apply *Speaker* and *Listener* tags in a similar manner to ConvoSense (Figure 3). Since human-generated inferences often contain nominal references to specific target entities, we additionally incorporate the names of conversational participants into the tags, as exemplified by “Speaker (A)”.

The naming conventions vary across the different human-generated datasets. To maintain uniformity, we adopt the naming conventions used in Cicero for both ComFact and Reflect, as Cicero constitutes nearly 90% of HumanGen. In Cicero, participants are denoted as A and B. For ComFact, originally lacking speaker designations, we randomly assign A/B tags to each conversation. On the other hand, Reflect includes original speaker names; thus, we replace them with A/B tags accordingly. Since the speaker name frequently appears in Reflect’s inferences, we uniformly replace it with “the speaker”, aligning with the prevalent format in Cicero.

4.2 ConvoSense: New GPT-generated Dataset

Constructing a practical dataset of commonsense inferences for dialogue benefits from covering

Desire	Desire _o
<p>Listener: Hi, Taraji. How are you doing today? Speaker: I'm doing fine, thank you. Just working on my math homework. Listener: Do you need any help with that? Speaker: Yeah, I could use some help. Thank you. Listener: Let's take a look. What are you working on? Speaker: I'm working on this problem where I have to find the perimeter of this shape.</p>	<p>Speaker: This book is terrible! Listener: What's wrong with it? Speaker: The author doesn't know how to tell a story! All the characters are two-dimensional and the plot is full of holes!</p>
<ol style="list-style-type: none"> 1. to ask the listener if she knows any shortcuts or tricks to find the perimeter quickly. 2. to learn the different types of shapes and their respective perimeters to improve her math skills. 3. to know the formula for calculating the perimeter so that she can apply it to the given shape. 4. to explore practical applications of finding perimeters in daily life, such as measuring the perimeter of her backyard. 5. to document the process of finding the perimeter step by step so that she can later revise it as a reference guide. 	<ol style="list-style-type: none"> 1. to know if speaker has any recommendations for better books. 2. to discuss specific examples of two-dimensional characters and plot holes in the book with speaker. 3. to ask speaker if they have read any other books by the same author. 4. to leave a negative review of the book online. 5. to stop reading the book and find something else to read.

Figure 3: Desire and Desire_o inferences in the ConvoSense dataset.

a wide variety of dialogue situations. To this end, our construction process of ConvoSense first carefully selects the dialogues to include based on their topical diversity, trims the dialogue contexts to optimize utterance diversity, and finally generates the inferences for each context.

Dialogue Selection We choose to sample the dialogues for ConvoSense as a subset of those dialogues in the high-quality and large-scale SODA dataset. SODA contains over a million dyadic dialogues generated by GPT covering situations based on ATOMIC commonsense tuples (Kim et al., 2023). For cost practicality, ConvoSense is constructed to contain 10,000 training dialogues, 1,000 validation dialogues, and 1,000 test dialogues.

To encourage diversity in ConvoSense, we employ BERTopic (Grootendorst, 2022), which clusters the dialogues selected from SODA into groups using dimension reduction technique UMAP (McInnes et al., 2020) and HDBSCAN clustering algorithm (McInnes et al., 2017) on the BERT embeddings of the dialogues.⁵ We configure the hyperparameters⁶ to effectively group dialogues while maintaining a well-balanced distribution of group lengths based on manual verifications. As a result, we obtain 100K dialogue groups, where each group consists of 6.3 dialogues on average. These groupings represent 100K unique dialogue topics, thus enabling the construction of ConvoSense to span a variety of

topics by sampling dialogues from a subset of these groupings.

Next, we randomly select one dialogue from the n groupings, where each dialogue contains at least 5 utterances and has a BERTopic score of at least 0.95 to its group. To maintain distinct dialogue scenarios in each split, each grouping can only be selected for one split. Through this procedure, we set n values as [10K, 1K, 1K] for assembling the training, validation, and test splits, respectively.

Utterance Selection For each selected dialogue, we determine which utterance to perform inference generation on. We use the topic keywords identified for each group during the BERTopic grouping to pinpoint the most topically salient utterance in each dialogue and ensure that the diversity afforded by the grouping is maintained. This is achieved by selecting the utterance whose embedding yields the highest cosine similarity with the embedding of the four-word topic string assigned to the dialogue’s respective group by BERTopic. Subsequently, we trim the dialogue’s utterances such that the conversation ends at this selected utterance. This trimmed version becomes the final dialogue context used for commonsense inference generation, where the inferences are derived for the last utterance.

Because commonsense inferences often relate to a central figure in a conversation, either the speaker or the listener, we introduce nominal tags for the two participants. The terminal utterance is labeled as *Speaker*, and its preceding utterance is labeled as *Listener*. These nominal tags are then assigned in alternating order to the remainder.

⁵The all-mpnet-base-v2 model is used for BERT.

⁶neighbors: 5, components: 5, min_cluster_size: 2.

	All					Poly			
	Examples	Words	Inferences	U1(#)	U2(#)	Examples	U1(%)	U2(%)	UL(%)
ConvoSense	120,000	14.6	5.1 (2–13)	16,666	199,087	120,000	92.8	98.9	98.8
ComFact	3,909	3.2	1.4 (2–12)	295	315	1,401	86.7	97.3	60.3
Cicero	52,644	11.6	1.3 (2–11)	7,598	44,234	9,911	84.4	97.2	98.7
Reflect	3,000	6.4	1.1 (2–4)	835	1,407	216	85.1	95.2	82.2
HumanGen	59,553	6.6	1.3 (1–12)	2,886	15,420	11,528	86.7	97.0	78.3

Table 5: Statistics of the `ConvoSense` and `HumanGen` datasets. **Poly**: polymorphic examples (multiple inferences). **Examples**: # of examples, **Words**: average # of words per inference, **Inferences**: average # of inferences per example with range shown in parentheses, **U1/2(#)**: average # of unique unigrams/bigrams across all inferences, **U1/2(%)**: average % of unique unigrams/bigrams between inferences within a single example, **UL(%)**: average % of unique inferences across all examples. Averages are calculated at the macro level across all inference types.

Inference Types For each preprocessed dialogue, GPT generates inferences for all included commonsense types following the procedure in Section 3. Specifically, ten commonsense types are included: Subsequent, Cause, Prerequisite, Motivation, Attribute, Reaction, Reaction_o, Desire, Desire_o, and Constituents (highlighted in Table 3). These types are selected based on their usage frequency in existing datasets and their lack of semantic overlap.

Data Statistics Table 5 presents data statistics for `ConvoSense` and `HumanGen`. `ConvoSense` significantly surpasses `HumanGen` for data volume, particularly regarding instances with polymorphic outputs, where multiple inferences can be derived per instance. Moreover, `ConvoSense` boasts greater vocabulary diversity and reduced redundancy among inferences. Illustrative examples from `ConvoSense` are shown in Figure 3.

Data Quality The results in Section 3.3 demonstrate that GPT is generally capable of producing high-quality commonsense inferences regardless of the underlying dialogue source. Consequently, applying GPT to generate commonsense inferences for the SODA dialogues is expected to perform with similar high quality. To explicitly verify this, we conduct an evaluation of the `ConvoSense` dataset. An external conversational AI expert, unaffiliated with this study, evaluates the generated inferences for 100 `ConvoSense` examples (508 total inferences; average 5.08 inferences per example), with all ten inference types uniformly represented across examples. The human judge completes two evaluation tasks: grading reason-

	ConvoSense
Reasonable	91
Novel	97
Detailed	63
Clusters	4.82 (95%)

Table 6: Human evaluation results on 100 examples of `ConvoSense` data, including the % of total inferences judged to be reasonable and novel, the % of positive novelty inferences judged to be detailed (vs. simple), and the average number of unique inference clusters per example, with the average % of unique inferences per example in parentheses.

ability and novelty of an inference (Section 3.2) and performing inference clustering to measure per-example output diversity (Section 6.2). Table 6 presents the results, confirming the high reasonability, novelty, detailedness, and diversity of the inferences in the `ConvoSense` dataset.

Error Analysis We next perform an error analysis on the unreasonable inferences identified by the human judge. We observe that most unreasonable inferences are explained by being too niche to be likely given only the provided information in the dialogue context (26%; Desire examples #4-5 in Figure 3), or by their attribution to the wrong conversational participant (26%; Desire_o examples #4-5 in Figure 3). Relatively speaking, only a small percentage of unreasonable inferences are explained by a violation of common knowledge of human experiences (10%), a lack

of relevance to the dialogue context (10%), or a contradiction of the dialogue context (7%). This suggests that ConvoSense inferences are predominantly accurate representations of commonsense understanding, although they can suffer from lack of precision regarding situational nuances and speaker roles.

5 Generative Commonsense Models

5.1 Training and Decoding Strategies

With the rich and diverse multi-inference examples provided in ConvoSense, we are well-positioned for training commonsense generation models that produce versatile outputs. Yet, a key query remains: How can we induce this versatility into the model?

A common method of enhancing diversity in generative outputs is to modify the decoding strategy (Gimpel et al., 2013; Vijayakumar et al., 2018; Ippolito et al., 2019). Through preliminary testing, we observe that diverse beam search decoding with Hamming distance reward following Vijayakumar et al. (2018) improves the output diversity with less impact on accuracy compared to other methods.

On the other hand, Cao and Wan (2020) propose modifying the model architecture by introducing latent variables to guide output variety. However, these approaches only approximate learning varied responses by relying on conditioning on random latent variables. In contrast, ConvoSense provides direct access to numerous inferences per input, enabling direct training of generative models that produce multiple inferences per example, with the set of inferences treated as target outputs during training. Therefore, we explore the performance of three strategies for diverse generation of commonsense inferences.

Monomorphic Beam Search (M) This model receives as input a dialogue context C consisting of the previous six utterances delimited by their corresponding speaker tags, the current response r for which to generate inferences, and a commonsense question q pertaining to one of the ten inference types (Table 3) in the following format:

$$C \setminus n r \setminus n \setminus n [Question] q \setminus n [Answer]$$

It is trained to output a single inference i . During training, instances with multiple correct infer-

ences I generate several training examples, one for each target inference $i \in I$. During inference, standard beam search decoding is used to generate k outputs.

Monomorphic Diverse Beam Search (M*)

This model adheres to the same design as the M model, except during inference, it uses Hamming-distance diverse beam search decoding instead to generate k outputs, following Vijayakumar et al. (2018).

Polymorphic (P) Using the same input as the M model, this model is trained to output a series of inferences as a sequence. To do this, the ground-truth inferences for each training example are concatenated into a list-like sequence, delimited by semicolons and prefixed by an integer representing their position in the list as follows:

$$(1)i_1; (2)i_2; (3)i_3; \dots$$

The order of the answers in the list are shuffled between each training epoch. During inference, standard beam search decoding is used to generate the top-1 output. A single output from this model is intended to represent the set of multiple diverse inferences for the input, without the need for any post-hoc decoding strategies, which other studies have observed to negatively impact the accuracy of the output generations (Ippolito et al., 2019).

5.2 Model Configuration

We develop six generative models: ConvoSenseM, ConvoSenseM*, ConvoSenseP, HumanGenM, HumanGenM*, and HumanGenP. Each model name denotes the training dataset with the terminal letter indicating the model strategy. All of them use T5-3b (Raffel et al., 2020) as the base model, which is then finetuned on the corresponding dataset following the indicated model strategy. The ConvoSense* and HumanGen* models are finetuned for 5 or 10 epochs, respectively. The best-performing models and hyperparameters⁷ are

⁷The Adafactor optimizer is used with a weight decay of $5e-3$ and a learning rate of $5e-6$, except for ConvoSenseP with $1e-6$. The max source length is set to 768. The max target length is set to 400 for P models and 128 for other models. All models are trained using bf16 for memory efficiency. P models use a prefix of ‘‘provide several reasonable answers to the question based on the dialogue: \n’’ and other models use a prefix of ‘‘provide a reasonable answer to the question based on the dialogue: \n’’.

selected through grid-search based on their results on the validation sets.

For all models, decoding is performed with 10 beams. For ConvoSenseM* and HumanGenM*, the number of beam groups is 10, and the diversity penalty is 0.5 and 1.0, respectively. For P models, decoding also uses a repetition penalty of 5.0 to reduce output token repetition.

It is worth noting that only 16% of HumanGen examples feature multiple ground-truth inferences. Training a P model on the complete dataset yields a single-inference model, which defeats the purpose of the *polymorphic* model strategy. Instead, we develop the HumanGenP model exclusively on multi-inference instances to facilitate learning of polymorphic outputs.

6 Generative Model Evaluation

We evaluate the six generative models (Section 5.2) on the ten commonsense inference types (Table 3) that exist in both the HumanGen (Section 4.1) and ConvoSense (Section 4.2) datasets. The model performance is evaluated using automatic reference metrics (Section 6.1), automatic diversity metrics (Section 6.2), and human evaluations of reasonability and novelty (Section 6.3).

6.1 Automatic Reference Metrics

Conventional evaluations of generative models against ground-truth references often overlook the diverse nature of the outputs. They typically assess individual model outputs against a single reference, focusing on best-case performance due to dataset constraints. However, such assessments are inadequate for our multi-inference dialogue generation objective. To address this, we structure our automated evaluation method to account for the concept of output diversity. This method, referred to as PolyAgg, serves as an aggregation function compatible with standard evaluation metrics. Its purpose is to gauge the model’s capacity to encompass the complete set of ground-truth references in its generated outputs.

Algorithm 1 demonstrates the PolyAgg aggregation function. It computes a score matrix for each example, where rows represent model outputs and columns represent ground-truth references, and finds the maximal assignment of rows to columns following the linear sum assignment problem (Burkard and Cela, 1999), which seeks to

Algorithm 1 Metric Aggregation

```

1: procedure POLYAGG(outputs, references)
2:   matrix  $\leftarrow$  []
3:   for  $o \in$  outputs do
4:     row  $\leftarrow$  []
5:     for  $r \in$  references do
6:       score  $\leftarrow$  METRIC( $o$ ,  $r$ )
7:       APPEND(score, row)
8:     APPEND(row, matrix)
9:    $a \leftarrow$  LINEARSUMASSIGNMENT(matrix)
10:  return MEAN( $a$ )

```

find the optimal bijective mapping between rows and columns in a cost matrix. By mandating a one-to-one mapping from model outputs to references, we can accurately measure reference set coverage and prevent models that generate mere surface-level variations from scoring highly on datasets with diverse references. We use SciPy’s linear sum assignment solver, then calculate the mean of the assigned scores for the final metric value. Dou et al. (2021) utilize a similar aggregation for evaluating a diverse dialogue response generation model.

One consideration for PolyAgg is that it can only match up to the number of generated outputs. If a model generates fewer outputs than there are references, PolyAgg will not measure against all references. However, this is a reflection of the model’s coverage capability, which is valuable information. To capture this, we introduce a coverage moderator for the PolyAgg score. Using cardinality notation $|\cdot|$, where $outs_e$ denotes the model outputs and $refs_e$ denotes the ground-truth references for a single example $e \in E$, the coverage moderator C is defined as:

$$C = \frac{|outs_e|}{|refs_e|} \quad (1)$$

Furthermore, different dialogue contexts can vary in the amount of diversity to their inferences, due to the nature of the described situations or shared information within the dialogue. A model achieving a high PolyAgg score on a diverse example should receive greater reward compared to a low-diversity case. Thus, not all examples should be treated equally when computing the overall model score; rather, each score should be proportionally weighted based on the corresponding number of ground-truth references.

	HumanGen Test Split ($n = 11,494$)						ConvoSense Test Split ($n = 10,000$)					
	Top-1			Top-5			Top-1			Top-5		
	BLEU	BS	Embed	BLEU	BS	Embed	BLEU	BS	Embed	BLEU	BS	Embed
ConvoSenseM	<u>5.407</u>	<u>0.641</u>	<u>0.422</u> [†]	6.282	0.650	0.462	<u>19.019</u>	<u>0.777</u>	<u>0.730</u> [†]	11.119	0.700	0.603
ConvoSenseM*	5.131	0.637	0.416	<u>6.710</u>	<u>0.658</u> [†]	<u>0.496</u>	17.923	0.773	0.725	<u>11.933</u>	<u>0.709</u>	<u>0.627</u>
ConvoSenseP	4.922	0.635	<u>0.422</u>	6.026	0.645	0.482	15.163	0.758	0.703	9.725	0.644	0.564
HumanGenM	<u>10.724</u>	<u>0.711</u>	<u>0.538</u>	<u>12.701</u>	0.721	0.576	5.095	0.633	0.501	<u>3.574</u>	<u>0.588</u>	<u>0.413</u>
HumanGenM*	9.473	0.697	0.511	12.056	<u>0.724</u> [†]	<u>0.591</u>	4.263	0.617	0.481	3.045	0.571	0.393
HumanGenP	9.524	0.700	0.523	9.658	0.645	0.504	<u>6.358</u>	<u>0.655</u>	<u>0.528</u>	2.330	0.256	0.201

Table 7: Reference metric results on test splits. Columns **BS** denote Bertscore. Underline indicates best metric with statistical significance under Bonferonni multi-test correction, except where indicated by † (t-test, $\alpha = 0.05$).

Combining the `PolyAgg` aggregation, coverage moderator C , and diversity weighting, the final score for a model is calculated as:

$$\frac{\sum_{e \in E} PolyAgg(outs_e, refs_e) * C * |refs_e|}{\sum_{e \in E} |refs_e|} \quad (2)$$

We use this evaluation scheme with three automatic metrics to measure the performance of the models. We include the traditional ngram-matching BLEU metric with $n \in [1, 4]$ (Papineni et al., 2002), the embedding-based metrics Bert Score⁸ (Zhang et al., 2019), and sentence cosine similarity using SentenceBert⁹ (Reimers and Gurevych, 2019).

Results We evaluate each model in terms of both its best-case performance (Top-1 output) and its multi-inference performance (Top-5 outputs). In the Top-1 setting, the maximum score achieved by the top-1 output against all of the ground-truth references for an example is taken and averaged across the test data. In the Top-5 setting, the top-5 outputs from the models are taken and scores are calculated using Equation 2, before being averaged across the test data. For M(*) models, the top one or five beams are taken as the outputs for each setting. For P models, the first one or five inferences in the outputted sequence are taken as the outputs for each setting. The results are shown in Table 7 for each model on the HumanGen and ConvoSense test splits, respectively.

Overall, it is evident that using diversity-promoting decoding (M*) outperforms the direct generation of multiple inferences (P). This ap-

proach achieves the highest BLEU, BertScore, and sentence similarity scores in the Top-5 assessment setting. This trend is particularly pronounced in the case of the ConvoSense-trained model, holding true for both the ConvoSense and HumanGen test splits. Enhancing training inference diversity as seen in ConvoSense appears to support the adoption of diversity-focused decoding strategies, yielding more contextually relevant outputs aligned with ground-truth references, even when applied to test examples from different datasets.

In the Top-1 setting, monomorphic models with standard beam search demonstrate superior performance for both HumanGen- and ConvoSense-trained models. However, the difference compared to diverse beam search is relatively minor, particularly when considering embedding-based metrics. Interestingly, the HumanGenP model displays the strongest ability to generalize to the ConvoSense test split among all HumanGen-trained models in the Top-1 scenario. Upon manual comparison of HumanGenP outputs against other HumanGen-trained models, we observe that HumanGenP is more inclined to specify a focal person in the inference (e.g., ‘the speaker/listener’). This often aligns better with ConvoSense references, although in a superficial manner with little impact on the underlying semantics.

It is also observed that the models produce low scores when evaluated against the test examples that are out-of-distribution with respect to their training data. This may not reflect the true underlying reasonability of the generated inferences, but rather a difference in inference content between the datasets, which is supported by evidence in Section 3.3 showing that human-written generations are more often repetitive with the dialogue context than GPT generations. To obtain a direct

⁸BertScore: microsoft/deberta-xlarge-mnli.

⁹SentenceBert: all-mpnet-base-v2.

	Clusters	Words
ConvoSenseM	2.680 (54%)	12.179
ConvoSenseM*	<u>3.509</u> (70%)	12.928
ConvoSenseP	3.262 (74%)	13.292
HumanGenM	3.031 (61%)	6.492
HumanGenM*	<u>3.452</u> (69%)	5.544
HumanGenP	1.348 (69%)	7.744

Table 8: Diversity metric results. **Clusters**: average inference clusters identified per example (with average % of unique inferences per example in parentheses). Underline indicates statistical significance in number of clusters within-block (t-test, $\alpha = 0.05$). Cross-block (ConvoSenseM* vs HumanGenM*) significance is not achieved. **Words**: average number of inference words.

measure of the quality of the generated model inferences, we perform a human evaluation in Section 6.3.

6.2 Automatic Diversity Metrics

To assess the ability of each model in generating diverse inferences for a given dialogue context, we employ a clustering approach under the Top-5 evaluation scheme. This involves grouping the model generations for each example into clusters of inferences with similar meanings. The average number of inference clusters across examples serves as a measure of output diversity.

For each of the ten inference types, we draw 50 examples from the test splits of ConvoSense and HumanGen, except for the `Constituents` type in HumanGen due to its smaller test split (22 examples). We instruct GPT4¹⁰ to create groups of semantically similar inferences given a dialogue context, question, and a list of inferences. GPT4 demonstrates its proficiency by achieving an average B-cubed F1-score (Bagga and Baldwin, 1998) of 0.872 against clusters identified by one of the authors for 20 examples, where B-cubed is a common clustering evaluation metric that measures the precision and recall of each element’s neighbors within the same cluster. This outperforms Amazon Mechanical Turk crowdworkers who only achieved a score of 0.581.¹¹

Results Table 8 displays diversity outcomes per model. For both HumanGen and ConvoSense-

¹⁰gpt-4-0613 with a temperature setting of 0.

¹¹The self-serve SurgeAI crowdsourcing platform previously used in Section 3.2 was discontinued during this work.

	Top-1		Top-5		
	R	N	R	N	Clusters
ConvoSenseM*	<u>90</u>	<u>98</u>	<u>93</u>	<u>98</u>	3.42 (68%)
HumanGenM	75	57	81	56	2.25 (45%)
HumanGenM*	75	70	81	70	3.17 (63%)

Table 9: Percentage of reasonable (**R**) and novel (**N**) inferences from each model. Underline denotes a statistically significant result against both HumanGen models (chi-square proportions test, $\alpha = 0.05$). The average number of inference clusters is also shown, along with the average % of unique inferences per example in parentheses (**Clusters**).

trained models, the monomorphic model with diverse beam search generates the most unique outputs.¹² While ConvoSenseM* slightly outperforms HumanGenM* in terms of inference diversity, both models exhibit similar unique inference cluster counts. Compared to the ConvoSense inferences themselves (Table 6), it is clear that none of the trained models are able to replicate the high inference diversity. Nonetheless, there is a large discrepancy in inference detail, which is revealed through human assessments in the next section.

6.3 Human Evaluations

We also evaluate the models through human assessment, in both the Top-1 and Top-5 setting. Based on automated evaluation outcomes, we compare ConvoSenseM* to both HumanGenM and HumanGenM*. An external conversational AI expert, unaffiliated with this study, evaluates the top five inferences for 60 examples per model in a blinded design, with all ten inference types and both datasets being uniformly represented. The human judge completes two evaluation tasks: grading reasonability and novelty of an inference (Section 3.2) and performing inference clustering (Section 6.2).

Results Table 9 demonstrates ConvoSenseM*’s superior performance compared to the HumanGen models. ConvoSenseM* achieves a remarkable 93% reasonability and 98% novelty, averaging 3.4 unique inferences per example. Indeed, similar results hold even when considering the Top-1

¹²High unique percentages for P models are due to low-count inference output (average of 4.4 and 2.0 outputted inferences for ConvoSenseP and HumanGenP, respectively).

output per model, showing that ConvoSenseM* exhibits strong performance regardless of whether a single-best inference is desired or a diverse set of inferences are desired. Moreover, when considering the positive novelty inferences in the Top-5 setting, we observe that 75% are annotated as *detailed* for ConvoSenseM* whereas only 7% are indicated as such for HumanGenM*. This reveals a substantial improvement in the amount of detail present in the inferences produced by ConvoSense models as compared to HumanGen models, which results in richer information being provided by the model.

7 Limitations and Ethical Considerations

This work does not intend to present an exhaustive set of commonsense inferences for dialogue. While we adhere to established inference types relevant to dialogue from existing literature, there could be overlooked types or unique challenges within specific dialogue domains that remain to be explored.

Furthermore, it is important to recognize that some social commonsense inference types may be associated with stereotypes and biases. When employing a model that produces commonsense inferences in a setting that impacts human users, caution must be exercised to prevent unjust or prejudiced decisions. Although exploration of the prevalence of harmful biases is out of the scope of the current work, we welcome future investigations into quantifying these aspects of our resources.

Finally, we adhered to OpenAI’s terms of service and related policies when utilizing GPT, and we acknowledge that any subsequent utilization of our models and data should refer to these policies.

8 Future Work

Although ConvoSense is composed of diverse multi-inference dialogue data (Table 6), it is clear from our experiments (Tables 8 and 9) that our trained models do not quite achieve the same degree of inference diversity. Further work is needed on improving the ability of distilled models to better capture the diversity present in the data.

In addition, the integration of commonsense understanding into dialogue applications has shown promising results in improving performance on tasks such as response generation, summarization, and reading comprehension in previous works. In

light of this, our work on improving commonsense resources and models presents an opportunity for further advancements in these dialogue applications. In particular, future work exploring how to capitalize on our commonsense model for dialogue response generation is highly compelling, since commonsense errors are one of the most common issues for modern dialogue agents (Finch et al., 2023). However, previous works have revealed that naive integration of commonsense inferences into neural models do not necessarily produce improvements (Zhou et al., 2022a). As a result, we leave the integration of our commonsense model to future work to allow for thorough investigation of its impact on response generation, covering aspects such as the impact of different commonsense inference types, the filtering of relevant inference types per dialogue context, and the effect of synthesizing multiple inferences into dialogue responses.

9 Conclusion

In this work, we present ConvoSense, an automatically constructed dataset of multi-output commonsense inferences for dialogue. ConvoSense surpasses existing datasets in size, advances inference detail and novelty, and attains comparable (if not superior) reasonability when compared to existing datasets. Our investigation into various techniques for generating multiple inferences reveals that diverse beam search on single-output generative models yields the best outcomes. By publicly releasing our trained models, we enable other works to benefit from the remarkable improvements in commonsense reasonability and novelty achieved by this work.

Acknowledgments

We gratefully acknowledge the support of Amazon for this work. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Amazon. We would also like to thank the anonymous reviewers and the action editor for their valuable feedback.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at the*

- 1st Conference on Language Resources and Evaluation*, pages 563–566.
- Rainer E. Burkard and Eranda Cela. 1999. Linear assignment problems and extensions. In *Handbook of Combinatorial Optimization: Supplement volume A*, pages 75–149. Springer. https://doi.org/10.1007/978-1-4757-3023-4_2
- Yue Cao and Xiaojun Wan. 2020. DivGAN: Towards diverse paraphrase generation via diversified generative adversarial network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2411–2421. <https://doi.org/10.18653/v1/2020.findings-emnlp.218>
- Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association. <https://doi.org/10.1037/10096-006>
- Yao Dou, Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2021. MultiTalk: A highly-branching dialog testbed for diverse conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12760–12767. <https://doi.org/10.1609/aaai.v35i14.17510>
- Sarah Fillwock and David Traum. 2018. Identification of personal information shared in chat-oriented dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Sarah E. Finch, James D. Finch, and Jinho D. Choi. 2023. Don’t forget your ABC’s: Evaluating the state-of-the-art in chat-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15044–15071, Toronto, Canada. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.839>
- Saadia Gabriel, Chandra Bhagavatula, Vered Shwartz, Ronan Le Bras, Maxwell Forbes, and Yejin Choi. 2021. Paragraph-level commonsense transformers with recurrent memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12857–12865. <https://doi.org/10.1609/aaai.v35i14.17521>
- Silin Gao, Jena D. Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2022. ComFact: A benchmark for linking contextual commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1656–1675, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-emnlp.120>
- Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. CIDER: Commonsense inference for dialogue explanation and reasoning. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 301–313. <https://doi.org/10.18653/v1/2021.sigdial-1.33>
- 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 (Volume 1: Long Papers)*, pages 5010–5028, Dublin, Ireland. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.344>
- Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. 2013. A systematic exploration of diversity in machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111.
- Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794v1*. <https://doi.org/10.48550/arXiv.2203.05794>
- Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108. <https://doi.org/10.1162/tacl.a.00302>
- Kilem Li Gwet. 2002. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-Rater Reliability Assessment*, 1.
- Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine

- Bosselut, and Yejin Choi. 2021. (Comet-) Atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6384–6392. <https://doi.org/10.1609/aaai.v35i7.16792>
- Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762. <https://doi.org/10.18653/v1/P19-1365>
- 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. <https://doi.org/10.1109/ACII.2013.47>
- Di Jin, Sijia Liu, Yang Liu, and Dilek Hakkani-Tur. 2022. Improving bot response contradiction detection via utterance rewriting. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 605–614. <https://doi.org/10.18653/v1/2022.sigdial-1.56>
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.799>
- Seungone Kim, Se June Joo, Hyungjoo Chae, Chaehyeong Kim, Seung-won Hwang, and Jinyoung Yeo. 2022. Mind the gap! Injecting commonsense knowledge for abstractive dialogue summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6285–6300, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10993–11001. <https://doi.org/10.1609/aaai.v36i10.21347>
- Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205. <https://doi.org/10.21105/joss.00205>
- Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426v3*. <https://doi.org/10.21105/joss.00861>
- Koh Mitsuda, Ryuichiro Higashinaka, and Yoshihiro Matsuo. 2019. What information should a dialogue system understand?: Collection and analysis of perceived information in chat-oriented dialogue. In *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*, pages 27–36. Springer. https://doi.org/10.1007/978-3-319-92108-2_3
- Zhufeng Pan, Kun Bai, Yan Wang, Lianqiang Zhou, and Xiaojiang Liu. 2019. Improving open-domain dialogue systems via multi-turn incomplete utterance restoration. 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 1824–1833. <https://doi.org/10.18653/v1/D19-1191>
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. <https://doi.org/10.3115/1073083.1073135>
- David Quarfoot and Richard A. Levine. 2016. How robust are multirater interrater reliability indices to changes in frequency distribution? *The American Statistician*, 70(4):373–384. <https://doi.org/10.1080/00031305.2016.1141708>
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael

- Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. 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 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
- Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237. <https://doi.org/10.1609/aaai.v36i10.21373>
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3027–3035. <https://doi.org/10.1609/aaai.v33i01.33013027>
- Siqi Shen, Deepanway Ghosal, Navonil Majumder, Henry Lim, Rada Mihalcea, and Soujanya Poria. 2022. Multiview contextual commonsense inference: A new dataset and task. *arXiv preprint arXiv:2210.02890v2*. <https://doi.org/10.48550/arXiv.2210.02890>
- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31. <https://doi.org/10.1609/aaai.v31i1.11164>
- Ashwin K. Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search: Decoding diverse solutions from neural sequence models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. <https://doi.org/10.1609/aaai.v32i1.12340>
- 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. <https://doi.org/10.18653/v1/2022.naacl-main.341>
- Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet. 2013. A comparison of Cohen’s Kappa and Gwet’s AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13:1–7. <https://doi.org/10.1186/1471-2288-13-61>, PubMed: 23627889
- Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2023. SocialDial: A benchmark for socially-aware dialogue systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. <https://doi.org/10.1145/3539618.3591877>
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.
- Pei Zhou, Hyundong J. Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022a. Reflect not reflex: Inference-based common ground improves dialogue response quality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. <https://doi.org/10.18653/v1/2022.emnlp-main.714>
- Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022b. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1237–1252. <https://doi.org/10.18653/v1/2022.acl-long.88>

Pei Zhou, Pegah Jandaghi, Hyundong Cho, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2021.

Probing commonsense explanation in dialogue response generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4132–4146. <https://doi.org/10.18653/v1/2021.findings-emnlp.349>