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

1 Introduction

Mastering commonsense understanding and reasoning is a pivotal skill essential for conducting engaging conversations. While previous efforts have led to the creation of datasets aimed at facilitating commonsense inferences in dialogue settings, these datasets often suffer from limitations such as small size (Zhou et al., 2022), lack of detailed information (Gao et al., 2022), redundancy with existing conversation content (Ghosal et al., 2022), and inadequate representation of the multifaceted nature of commonsense reasoning (Shen et al., 2022). In response to these challenges, we present ℂonvo𝕊ense: a novel synthetic dialogue commonsense dataset created using GPT that boasts greater contextual novelty, offers a higher volume of inferences per example, and substantially enriches the detail conveyed by the inferences.[1] Our experimental results demonstrate that ℂonvo𝕊ense enables the training of generative commonsense models for dialogue that surpass the performance of models trained on previous datasets in terms of inference reasonability, novelty, and detail.

2 Data Collection: ConvoSense

We construct a zero-shot GPT prompting strategy to collect commonsense inferences of a particular type for a given dialogue context. GPT (gpt-turbo-3.5-301) is instructed to output a list of likely inferences given a dialogue context, a target terminal utterance, and a commonsense-focused question and answer prefix. Each inference type covered in this work has a tailored question and answer prefix to guide inference generation. We cover 10 popular inference types synthesized from previous works on dialogue commonsense (Gao et al., 2022; Ghosal et al., 2022; Shen et al., 2022; Zhou et al., 2022), including subsequent events, causes, prerequisites, motivations, attributes, emotional reactions, desires, and event constituents. Appendix A showcases the prompt design and the question and answer prefixes used for each commonsense type.

To construct our ℂonvo𝕊ense dataset, we utilize dialogue contexts from the high-quality SODA dialogue dataset (Kim et al., 2022), organizing them into topical groups using BERTopic (Grootendorst, 2022). One dialogue is selected from n groups to assemble the data splits, with n set to $[10K, 1K, 1K]$ for training, validation, and test splits, respectively. Each dialogue is trimmed to end with the most topically salient utterance to its group based on cosine similarity to the BERTopic group topic string. We then employ GPT to generate inferences for the commonsense types for these trimmed dialogues. The resulting ℂonvo𝕊ense dataset consists of over 500,000 inferences across 12,000 dialogues, with an average of 5.1 inferences per type per dialogue (examples in Appx. B).

Data Evaluation We compare our GPT-generated inferences to those written by human annotators from 3 existing dialogue commonsense datasets: 𝓒omFact (Gao et al., 2022), 𝓒icero (Ghosal et al., 2022; Shen et al., 2022), and 𝓡eflect (Zhou et al., 2022). 300 examples are sampled from each dataset and GPT inferences are generated following the procedure described in Section 2 (examples in Appx. B). Crowdsourced human annotators from SurgeAI are provided both a human-written and GPT-generated inference for the same example. Annotators categorize inferences into levels of **reasonability**: *always/likely* (+), *sometimes/possible* (+), *never/farfetched* (-), or *invalid/nonsense* (-) and into levels of **novelty**: *new & detailed* (+), *new & simple* (+), and *purely repetitive* (-). Following Hwang et al. (2021), the two metrics are converted into binary representations when analyzing the final outcomes, with stratifications into the positive/negative binary representation indicated by (+) and (-) previously.

[1] https://github.com/emorynlp/ConvoSense

Table 1 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 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, we observe that GPT inferences achieve higher detail than that observed from human-generated inferences when comparing the percentage of *new & detailed* inferences out of all positive novelty inferences in Figure 1.

	𝒞omFact		𝒞icero		ℛeflect	
	R	**N**	**R**	**N**	**R**	**N**
GPT	93	91	93	80	89	86
Human	81	73	88	70	91	82

Table 1: The % of total samples (#) labeled as reasonable (**R**) and novel (**N**). Underline denotes statistical significance against human-written inferences for the indicated dataset (McNemar's test, $\alpha = 0.05$).

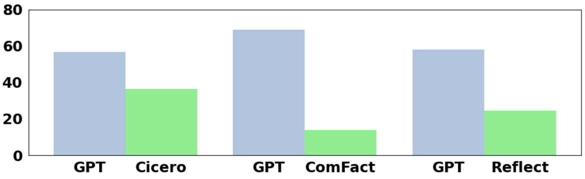


Figure 1: Average % of *new & detailed* inferences out of all positive novelty inferences for each data source.

3 Dialogue Commonsense Models

Given the success of GPT-generated inferences observed in Section 2, we train a T5-3b (Raffel et al., 2020) commonsense generation model (**ConvoSenseM**) on our collected ℂonvo𝕊ense dataset. This model takes as input a dialogue context, the current response for which inferences are to be generated, and a commonsense question. The expected output is an inference that is applicable to the provided input. Based on the observed diversity of inferences per example in ℂonvo𝕊ense, Hamming-distance diverse beam search decoding (Vijayakumar et al., 2016) is used to generate k diverse outputs from the trained model. We also train a comparable model (**HumanGenM**) using the existing human-written dialogue commonsense datasets indicated in Section 2.

Model Evaluation We evaluate the performance of each model on the reasonability and novelty metrics defined in Section 2 using an external conversational AI expert who is unaffiliated with this study. The annotator evaluates both best-case performance (Top-1 output) and multi-inference performance (Top-5 outputs), where the top one or five beams are taken as the outputs for each setting, respectively. For the multi-inference setting, the annotator also evaluates the ability of the model to output diverse inferences by clustering the outputted inferences into semantic groups. Evaluation is performed on 60 examples (300 inferences) per model in a blinded design.

	Top-1		**Top-5**		
	R	**N**	**R**	**N**	**Clusters**
ConvoSenseM	90	98	93	98	3.42 (68%)
HumanGenM	75	70	81	70	3.17 (63%)

Table 2: Percentage of reasonable (**R**) and novel (**N**) inferences from each model. Underline denotes a statistically significant result (chi-square proportions test, $\alpha = 0.05$). **Clusters** shows the average number of inference clusters and % of unique inferences per example.

Table 2 demonstrates ConvoSenseM's superior performance compared to the HumanGen model. ConvoSenseM achieves a remarkable 93% reasonability and 98% novelty, averaging 3.4 unique inferences per example. Indeed, similar results hold for the Top-1 output per model. Moreover, when considering the positive novelty inferences in the Top-5 setting, we observe that 75% are annotated as *detailed* for ConvoSenseM but only 7% for HumanGenM, indicating a vast improvement.

However, our experiments also reveal that our trained model does not outperform HumanGenM on inference diversity. This is surprising as the conversational expert judged inferences to be unique at an average rate of 95% in ℂonvo𝕊ense in additional analyses (Appx. C), whereas the human data is much less diverse. It is clear the ConvoSenseM does not achieve the same degree of inference diversity as the underlying data.

4 Conclusion

This study introduces ℂonvo𝕊ense, a substantial dataset of commonsense inferences for dialogue, surpassing existing human-written datasets in novelty and detail. Models trained on ℂonvo𝕊ense demonstrate superior performance in reasonability, novelty, and detail compared to those trained on other datasets, whether aiming for a single-best inference or a diverse set. However, further research is needed to fully capture the diversity of inferences present in ℂonvo𝕊ense within the trained models.

References

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.

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.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

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.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization.

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.

Siqi Shen, Deepanway Ghosal, Navonil Majumder, Henry Lim, Rada Mihalcea, and Soujanya Poria. 2022. Multiview Contextual Commonsense Inference: A New Dataset and Task. ArXiv:2210.02890 [cs].

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Pei Zhou, Hyundong J. 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*.

A GPT Prompt

An example of the final GPT prompt design, specifically tailored for the `Desire` inference type, is illustrated in Table 3. The question and answer prefix pairs used for the 10 covered commonsense inference types are shown in Table 4.

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 3: 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 prefix).

B Examples

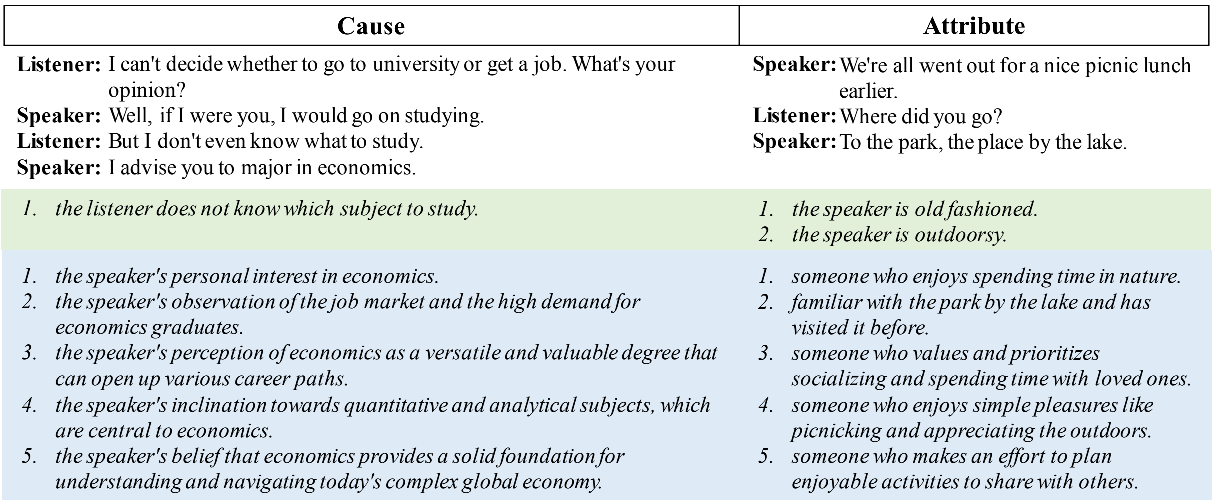
Examples of the inferences written by humans and the analogous inferences generated by GPT are shown in Figure 2. Illustrative examples from the ℂonvo𝕊ense dataset are shown in Figure 3.

C ConvoSense Evaluation

The results in Section 2 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 well. To explicitly verify this, we conduct an evaluation of the ℂonvo𝕊ense dataset where the expert annotator described in Section 3 evaluates the generated inferences for 100 ℂonvo𝕊ense examples (508 total inferences; average 5.08 inferences per example) for reasonability and novelty (Sec. 2), and inference clustering (Sec. 3). Table 5 presents the results, confirming the high reasonability, novelty, detailedness, and diversity of the ℂonvo𝕊ense dataset.

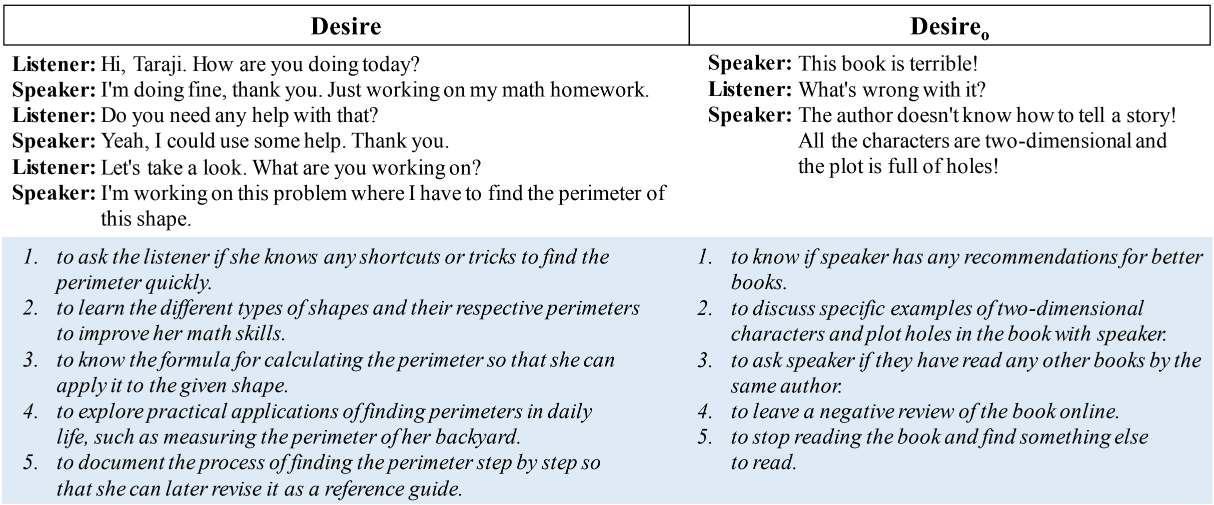
Type	Question	Answer Prefix
Subsequent	What might happen after what Speaker just said?	After this, ...
Cause	What could have caused the last thing said to happen?	This was caused by...
Prerequisite	What prerequisites are required for the last thing said to occur?	For this to happen, it must be true that...
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...

Table 4: Question and answer prefixes used for generating each inference type from GPT for dialogue contexts.



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

Figure 2: Cause and Attribute inferences written by humans (top, green) and generated by GPT (bottom, blue).



Desire	Desire$_o$
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.	**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!
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.*	*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 $\mathbb{C}$onvo$\mathbb{S}$ense dataset.

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

Table 5: Human evaluation results on $\mathbb{C}$onvo$\mathbb{S}$ense examples, showing the % of reasonable, novel, and detailed inferences, and the average number (or %) of unique inferences per example.