

Distribution Agreement

In presenting this dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this dissertation. I retain all ownership rights to the copyright of the dissertation. I also retain the right to use in future works (such as articles or books) all or part of this dissertation.

Sarah E. Finch (Fillwock)

Date

Enriching Open-Domain Dialogue Models with Predictive Social Commonsense

By

Sarah E. Finch (Fillwock)
Doctor of Philosophy

Computer Science and Informatics

Jinho D. Choi, Ph.D.
Advisor

Joyce C. Ho, Ph.D.
Committee Member

Fei Liu, Ph.D.
Committee Member

David Traum, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D., MPH
Dean of the James T. Laney School of Graduate Studies

Date

Enriching Open-Domain Dialogue Models with Predictive Social Commonsense

By

Sarah E. Finch (Fillwock)
B.S., Michigan State University, MI, 2018

Advisor: Jinho D. Choi, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2024

Abstract

Enriching Open-Domain Dialogue Models with Predictive Social Commonsense By Sarah E. Finch (Fillwock)

The advancement of open-domain dialogue systems represents a significant goal of artificial intelligence, aiming to create more engaging and human-like interactions between machines and users. A key challenge in this domain is equipping these systems with a profound understanding of human experiences, the nuances of which are often subtly implied rather than explicitly stated in conversations. Social commonsense resources aim to support the comprehension of human experiences, including capturing commonsense knowledge about people’s motivations, the causes of events, emotions, and more. However, the existing datasets and methodologies for social commonsense integration into dialogue applications suffer from low coverage, sparse detail, and contextual redundancy, thereby impeding their capability to promote meaningful dialogue interactions. Recognizing these limitations, this dissertation explores the enhancement of open-domain dialogue systems through improved integration of social commonsense knowledge.

This dissertation is structured around three core objectives: developing a reliable evaluation framework for assessing commonsense capability in dialogue models, creating a novel dataset of contextually novel commonsense inferences tailored for dialogue, and integrating these inferences into dialogue models to enhance their conversational abilities. The first objective is addressed through the introduction of Annotation of Behaviors in Chat Evaluation (ABC-Eval), a binary behavior-based evaluation framework that offers a more objective and grounded assessment of dialogue models’ commonsense reasoning capabilities. The second objective is achieved with the development of ConvoSense, which is the largest dataset of its kind to provide novel commonsense inferences designed specifically for dialogue contexts. Finally, the third objective culminates in the presentation of Commonsense Inference Generate-Select-Respond (CSI-GSR), a novel approach that leverages the rich pool of commonsense inferences from ConvoSense to generate dialogue responses.

The findings of this dissertation highlight the current capabilities of LLM-based dialogue models and the benefits of incorporating predictive commonsense inferences for response guidance. The work on ABC-Eval reveals that commonsense errors are highly prevalent in neural dialogue systems, thus highlighting the importance of improving commonsense capabilities of dialogue models. The work on ConvoSense produces powerful resources and models for capturing multi-faceted and predictive social commonsense inferences for dialogue. The work on CSI-GSR showcases the utility of these multi-faceted and predictive social commonsense inferences for advancing response specificity to its dialogue context. Collectively, this body of work supports the pursuit of more nuanced, contextually aware, and intelligent human-computer interactions.

Enriching Open-Domain Dialogue Models with Predictive Social Commonsense

By

Sarah E. Finch (Fillwock)
B.S., Michigan State University, MI, 2018

Advisor: Jinho D. Choi, Ph.D.

Acknowledgments

First and foremost, I extend my deepest gratitude to my advisor, Dr. Jinho Choi, for his guidance throughout the many years of my Ph.D. journey. His expertise, insight, and mentorship have been invaluable to my growth as a researcher. I am also immensely grateful to my committee members, Dr. Joyce Ho, Dr. Fei Liu, and Dr. David Traum, for their valuable feedback and constructive criticism, providing new perspective that have enriched my work. My heartfelt thanks go to my labmates and peers for the collaborative spirit and countless learning moments we have experienced together. The camaraderie and support within our lab have made this journey more enjoyable and rewarding.

I cannot express enough gratitude to my family for their endless support, encouragement, and love. Special thanks to my husband, James Finch, who embarked on this Ph.D. journey alongside me. Sharing this path with you, with all its ups and downs, has been a source of strength, inspiration, and companionship. Your collaboration, encouragement, and unwavering support have meant the world to me.

This experience would not have been the same without the contribution of each of these individuals. I am profoundly thankful for the privilege of knowing such an extraordinary group of people.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Central Thesis and Research Questions	4
1.3	Research Contributions	5
1.3.1	Commonsense Evaluation for Dialogue Models (Ch. 3)	5
1.3.2	Commonsense Dataset for Dialogue Models (Ch. 4)	6
1.3.3	Commonsense-Augmented Dialogue Model (Ch. 5)	6
1.4	Organization	7
2	Background	8
2.1	Large Language Models	8
2.2	Commonsense Resources for Dialogue	9
2.3	Commonsense Augmentation for Dialogue Models	14
2.4	Commonsense Evaluation for Dialogue Models	17
3	Annotation of Behaviors in Chat	18
3.1	Introduction	18
3.2	ABC-Eval Development	20
3.2.1	Collecting Behavior Label Candidates	21
3.2.2	Pilots and Development	21
3.2.3	Final ABC-Eval Design	27

3.3	Existing Evaluation Methods	36
3.4	Evaluation Collection	37
3.5	Evaluation Analysis	39
3.5.1	Interpretability	39
3.5.2	Importance	40
3.5.3	Sensitivity	41
3.5.4	Coverage & Distinctness	42
3.6	Cost	43
3.7	Dialogue Model Insights	45
3.8	Conclusion	45
4	ConvoSense: Generating Commonsense for Dialogue	47
4.1	Introduction	47
4.2	ChatGPT Dialogue Commonsense Generation	48
4.2.1	Prompt Engineering	48
4.2.2	Evaluating ChatGPT-generated Commonsense	50
4.2.3	Results	51
4.3	Constructing Commonsense Datasets	54
4.3.1	ConvoSense: New ChatGPT-generated Dataset	54
4.3.2	HumanGen: Human-generated Datasets	58
4.3.3	Dataset Statistics	59
4.4	Generative Commonsense Models	59
4.4.1	Training and Decoding Strategies	59
4.4.2	Model Configuration	61
4.5	Generative Model Evaluation	61
4.5.1	Automatic Reference Metrics	62
4.5.2	Human Evaluations	65
4.6	Conclusion	66

5	Generate-Select-Respond: Commonsense-Augmented Dialogue Model	67
5.1	Introduction	67
5.2	Approach	68
5.3	Approach Development	70
5.3.1	Commonsense Inference Generate-Select-Respond: CSI-GSR	70
5.3.2	Pilot Study	78
5.3.3	Approach Improvement	84
5.3.4	Few-shot Learning	89
5.4	Evaluation	96
5.4.1	Models	96
5.4.2	Test Data	99
5.4.3	Metrics	99
5.4.4	Metric Reliability	105
5.4.5	Results	105
5.5	Conclusion	110
6	Conclusion	112
6.1	Research Contributions	112
6.2	Future Work	113
6.2.1	Behavioral Evaluation of LLM-based Dialogue Models	114
6.2.2	Commonsense-augmented Dialogue Models “In-the-Wild”	114
6.2.3	Improving Social Commonsense Integration for Open-Domain Dialogue	115
6.2.4	Alternative Commonsense Integrations	115
6.2.5	Improving Open-source LLMs with Commonsense	116
6.2.6	Enhancing Topical Saliency in Dialogue Models	116
6.2.7	Refining Response Content Control	116
	Bibliography	118

List of Figures

1.1	Example dialogue demonstrating the utility of commonsense for intelligent interactions.	2
3.1	Example of commonsense contradiction labels in human-machine dialogue. . .	19
3.2	Interface for uninterpretable	28
3.3	Interface for antisocial	29
3.4	Interface for preference info and life info	30
3.5	Interface for empathetic and lack of empathy	31
3.6	Interface for commonsense contradiction	32
3.7	Interface for fact usage and fact contradiction	33
3.8	Interface for self contradiction , partner contradiction , and redundant	34
3.9	Interface for ignore , irrelevant , follow-up , and topic switch	35
3.10	IAA (Krippendorff’s alpha) for all metrics. Error bars denote 95% bootstrap confidence intervals, using bias-corrected and accelerated confidence intervals with $k = 10,000$ Monte Carlo case resamples. $!Soc_b$ and $!Int_b$ ’s confidence intervals are largely due to a low rate of positive examples (see Figure 3.13).	40
3.11	Extent to which each evaluation metric can explain variance in conversation quality by fitting a univariate regression model (R^2 for predicting Qua_d using linear regression, McFadden’s Pseudo- R^2 for predicting Qua_c using logistic regression). Comparative metrics cannot predict Qua_d so only results for variance of Qua_c are shown.	40

3.12	Incremental validity of metrics within 4 evaluation methods, obtained using backwards stepwise regression. Points represent the extent to which a model can explain variance in quality (R^2 for predicting Qua_d with a linear model, McFadden’s pseudo- R^2 for predicting Qua_c with a logistic model) using all metrics on the same line and to the left as predictors. Filled marker symbols denote steps where the model’s predictors all contributed positively to adjusted R^2 or adjusted pseudo- R^2 values; otherwise, marker symbol is unfilled. Comparative metrics cannot be used to predict Qua_d so only results for explaining variance of Qua_c are shown.	43
3.13	Proportions of turns expressing undesirable behaviors, with 95% Wilson score confidence intervals.	45
4.1	Average % of <i>new \mathcal{E} detailed</i> inferences out of all positive novelty inferences for each data source.	53
4.2	Cause and Attribute inferences written by humans (top, green) and generated by ChatGPT (bottom, blue).	53
4.3	Desire and Desire _o inferences in the ConvoSense dataset.	56
5.1	The module pipeline of CSI-GSR with finalized Enumeration prompt design for the prompt-based ChatGPT modules (S, R) (Section 5.3.1). Underline denotes instruction that changes for -single and -multi variants, with -single shown here (Section 5.3.2).	69
5.2	Proportion of instances where the specific model’s output is selected as the preferred option.	79
5.3	Average scores achieved by each dialogue model on the Likert-style metrics of Informativeness and Quality, where higher is better . Error bars denote 95% Student’s t confidence intervals. * denotes statistically significant differences (paired t-test, $\alpha < 0.01$).	106

5.4 Average proportions of turns of each dialogue model expressing the behaviors of Commonsense Contradictions, violations of Conversational Expectations, and Irrelevant, where **lower is better**. Error bars denote 95% Wilson score confidence intervals. * denotes statistically significant differences (McNemar's test, $\alpha < 0.01$). 108

List of Tables

2.1	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. [77], [2] Hwang et al. [32], [3] Ghosal et al. [21], [4] Zhou et al. [111]. Counts are truncated to the nearest order of magnitude. * indicates the type was included but no human-verified instances of it are present.	13
3.1	The final set of dialogue dimensions for human evaluation from Finch and Choi [14]. Dimensions above the dashed line are dimensional characteristics of dialogue models. The Quality dimension below the dashed line is an overall performance assessment.	20
3.2	The distribution of dialogues and annotators for each evaluation pilot. Dashed line delineates a subset (below line) of the total annotations (above line) produced in that pilot. α : Krippendorff’s alpha, U : Undergraduate annotators, G : Graduate level annotators, ALL : Undergraduate and graduate annotators.	23

3.3	The 16 behavior labels within ABC-Eval. Row separators denote annotation task groupings. [1] Gopalakrishnan et al. [24], [2] Higashinaka et al. [28], [3] Mehri and Eskenazi [56], [4] Mehri and Eskenazi [57], [5] Phy et al. [64], [6] Sanguinetti et al. [75], [7] Beattie et al. [3], [8] Sun et al. [83], [9] Xu et al. [100], [10] Rashkin et al. [70], [11] Smith et al. [80], [12] Majumder et al. [52], [13] Rashkin et al. [68], [14] Zhong et al. [104], [15] Zhou et al. [108], [16] Zhou et al. [112], [17] Gupta et al. [26], [18] Honovich et al. [29], [19] Santhanam et al. [76], [20] Shuster et al. [79], [21] Li et al. [45], [22] Nie et al. [60], [23] Welleck et al. [93], [24] Xu et al. [101]	27
3.4	The seven labels for dialogue characteristics for Likert and Comparative evaluations (taken from Finch and Choi [14]), henceforth referred to using their abbreviations and colors. The label for holistic quality is also shown below the dashed line.	36
3.5	The number of statistically significant differences detected by each metric when comparing bot-pairs using z-tests of proportions (ABC-Eval), t-tests (Turn Likert and Dialogue Likert), and sign tests (Comparative) at three significance thresholds.	42
3.6	Payment per annotation task in USD. The payment for Turn Likert is per label whereas the indicated payment for Dialogue Likert and Comparative covers all labels, due to how the annotation tasks were constructed (Section 3.2.3).	44
3.7	The data collection costs for each task in USD. T ime is the median completion time in minutes for one dialogue. T hrough P ut represents the number of completed dialogues per hour. E stimated C ost is calculated using median completion time, 400 dialogues, and \$20/hr rate. O ur C ost is the total amount paid in this work to collect a dataset of 400 conversations (single-annotated).	44

4.1	A ChatGPT 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).	49
4.2	Question and answer prefixes used for generating each inference type from GPT for dialogue contexts. The ten inference types included in ConvoSense are represented by Type	50
4.3	The average % ($\sigma < 2\%$) of total samples (#) tested as reasonable (R) and novel (N), with discordance probabilities in parentheses. *: statistical significance (McNemar’s test, $\alpha = 0.05$). 90 more samples are used for ComFact due to its greater number of inference types.	52
4.4	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.	57
4.5	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.	59
4.6	Reference metric results on test splits. Columns BS denote Bertscore. <u>Underline</u> indicates best metric with statistical significance under Bonferonni multi-test correction, except where indicated by † (t-test, $\alpha = 0.05$).	64

4.7	Percentage of reasonable (R) and novel (N) inferences from each model. <u>Underline</u> 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).	66
5.1	The Inference Selection (S) and Response Generation (R) prompts used for the Enumeration prompt design. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated commonsense inferences, respectively, of a provided example. Placeholder “{Output from S}” is replaced by the output from ChatGPT from the Inference Selection module.	73
5.2	The Inference Selection (S) and Response Generation (R) prompts used for the Structured Composition prompt design. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated commonsense inferences, respectively, of a provided example. Placeholder “{Output from S}” is replaced by the output from ChatGPT from the Inference Selection module.	74
5.3	The Inference Selection (S) and Response Generation (R) prompts used for the Group Partition prompt design. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated commonsense inferences, respectively, of a provided example. Placeholder “{Output from S}” is replaced by the output from ChatGPT from the Inference Selection module.	75
5.4	The Inference Selection (S) and Response Generation (R) prompts used for the Planning prompt design. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated commonsense inferences, respectively, of a provided example. Placeholder “{Output from S}” is replaced by the output from ChatGPT from the Inference Selection module.	76

5.5	Prefixes used to transform the generated commonsense inferences into natural language sentences when used as inputs into the Inference Selection and Response Generation prompts.	77
5.6	The average quality score for each prompt design.	77
5.7	The prompt used for native response generation of ChatGPT. Placeholder “{context}” is replaced by the dialogue context of a provided example.	79
5.8	Error analysis of generated responses from each CSI-GSR model variant, on responses that fail to be preferred to responses from native ChatGPT. Analysis results show the percentage of responses belonging to various error categories, where categories are mutually exclusive.	80
5.9	Example of Suboptimal Follow-up Topics response error category. The segment of the response that expresses the error type is bolded.	82
5.10	Example of Context Mixup response error category. The segment of the response that expresses the error type is bolded.	82
5.11	Example of Topic Order response error category. The segment of the response that expresses the error type is bolded.	83
5.12	Example of Presentation of Speculative Inferences as Facts response error category. The segment of the response that expresses the error type is bolded.	83
5.13	Example of Longwindedness response error category. The segment of the response that expresses the error type is bolded.	84
5.14	Examples of human-curated few-shots.	92
5.15	The Inference Selection (S) and Response Generation (R) prompts used for few-shot in-context-learning. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated commonsense inferences, respectively. Placeholder “{examples}” is replaced by the few-shot examples in the specified format.	93

5.16	Percentage (%) of each commonsense type selected by the CSI-GSR-icl approach.	96
5.17	The prompt used for native response generation of Llama2. Placeholder “{context}” is replaced by the dialogue context of a provided example.	97
5.18	The prompt used for state-of-the-art commonsense-augmented open-domain dialogue model, ChatGPT + Doctor. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated rationale, respectively, of a provided example. The turns in the context are prefixed in alternating fashion by nominal tags “A” and “B”, where the last turn is the context is prefixed by “B”.	98
5.19	Example commonsense outputs, where applicable, and response outputs for the models under study.	100
5.20	Instructions for the Conversational Expectations evaluation label.	103
5.21	Examples given to the annotators for the Conversational Expectations evaluation label. The correct label and explanation is shown in the right-hand column for the final “A” response of the dialogue context shown in the left-hand column.	104
5.22	Measure of annotator reliability for each metric under study. The inter-annotator agreement (Krippendorff’s α), average performance on each training quiz (Training %), and average overall training performance (Training Avg.) are shown. Averages are calculated across all annotators and Training % is formatted as <i>quiz #1 quiz #2 quiz #3</i>	105
5.23	Examples highlighting discrepancies in informativeness between responses based on Chae et al. [8].	107
5.24	Examples of responses that are positive examples of violations of Conversational Expectations, along with explanations of the violation.	110

List of Algorithms

1	Metric Aggregation	62
2	Few-shot Examples Construction	90

Chapter 1

Introduction

1.1 Motivation

Open-domain human-computer dialogue is a long-standing research objective in Artificial Intelligence (AI). To meet this objective, automated dialogue systems must engage in open-ended conversation with a human user, adapting fluently and intelligently to the topics that are introduced. In open-domain dialogue, there are no constraints on the topics that can be covered, requiring models that are able to carry on conversation across all dialogue situations. However, one commonality underlying open-ended conversation is that human users exhibit a preference for engaging with dialogue systems on discussions about life experiences and personal information, which has been shown in both human-human chat [13, 58], in chat between humans and embodied AI characters [72], and through experimentation during the 3rd Amazon Alexa Prize Socialbot Grand Challenge in 2019 [16]. Therefore, it becomes evident that open-domain dialogue systems must possess the ability to comprehend and reason about users’ experiences, intentions, and emotions to foster fulfilling interactions.

The dialogue depicted in Figure 1.1 serves as a compelling illustration of the role of experiential reasoning in fostering intelligent human-computer interactions. For instance, when the user expresses their fondness for the movie “Dune” in Turn 1, the system can

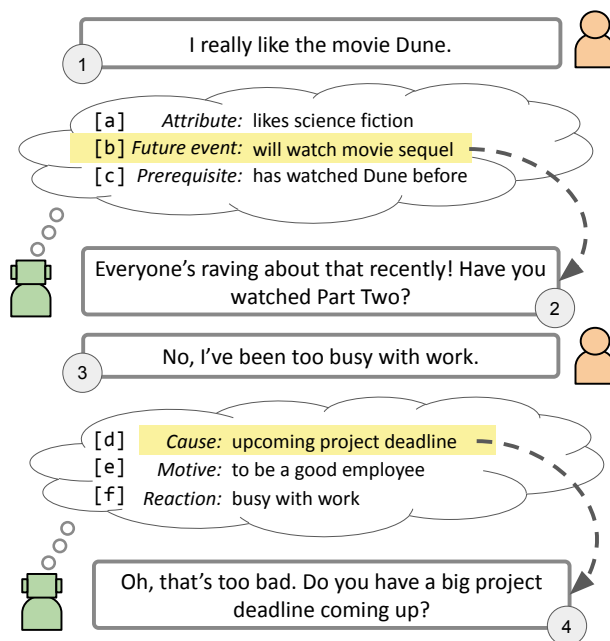


Figure 1.1: Example dialogue demonstrating the utility of commonsense for intelligent interactions.

infer various relevant details about the user’s movie-watching preferences ([a]) and make predictions about likely future plans of the user ([b]). These inferences enrich the contextual understanding of the user’s input and facilitate the generation of insightful follow-up responses, such as inquiring about the user’s thoughts on the movie’s sequel in Turn 2. Similarly, when the user mentions being occupied with work in Turn 3, the system can contemplate likely reasons behind their busy schedule such as an upcoming project deadline ([d]) and can infer that the user is motivated to be a good employee ([e]), leading to a more nuanced discussion about work-related matters.

In human-human dialogues, the ability to draw such inferences stems from shared world-views, enabling individuals to comprehend, deduce, and predict based on mutual experiences—a phenomenon commonly known as “commonsense”, particularly “social commonsense” [10]. Speculative commonsense inferences, which provide plausible information based on known facts, greatly enhance dialogue interactions by enabling intelligent follow-up engagements. Therefore, for dialogue models to achieve human-like understanding and adeptness in

navigating diverse conversation scenarios, they benefit from possessing robust capabilities in generating predictive social commonsense inferences related to shared human experiences commonly encountered in dialogue.

Over time, there has been a concerted endeavor to create datasets that facilitate commonsense reasoning. Early works focused on representing semantic characteristics of words and their relationships to one another, including WordNet [12], FrameNet [2], and ConceptNet [82]. Lately, efforts have shifted toward building datasets encompassing social- and event-based commonsense, such as ATOMIC [32], that operate on phrasal descriptions of situations. This new wave of datasets targets complex human concepts, including emotions, desires, and motivations. Datasets such as ATOMIC hold promise for dialogue applications as they provide insights directly relevant to human experience; however, a drawback lies in their lack of contextual awareness as they hinge on independent situational descriptions 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 [62, 35]. While there have been several attempts to create datasets that facilitate commonsense inferences specifically for dialogue contexts, existing datasets tend to cover only a small set of commonsense types, lack in-depth details [19], and restate information already present in the conversation rather than predicting novel information [21] (Fig. 1.1, [c], [f]).

The limitations of these commonsense resources lead to downstream impacts on the works that attempt to leverage the commonsense knowledge towards improving dialogue models. The majority of such work has focused on utilizing static, entity-based resources like ConceptNet, rather than social commonsense resources. For those works that do recognize the utility of social commonsense, they primarily narrow their focus to empathetic dialogue, which tend to focus only on a subset of social commonsense that is related to emotion prediction. Only a small handful of works approach the task of improving open-domain dialogue models by using social commonsense [111, 8]. Their results present evidence that social commonsense has a positive impact on dialogue models; however, their strategies for

commonsense integration are limited. When generating a response, these models often only consider a single commonsense candidate, even though social commonsense covers a wide range of types that are not all equally applicable to each dialogue situation [111].

Furthermore, it is a recurring theme in existing commonsense-augmented dialogue works to fail to perform an evaluation of the commonsense ability of their approaches, instead focusing on other aspects of the response such as its quality or relevance to the context. Those works that do attempt to evaluate the responses for their commonsense adherence use methods that fail to provide a clear and complete definition of commonsense [110], often polluting the definition with other aspects such as general context appropriateness and relevance [113]. Consequently, it is impossible to know where the field stands in terms of commonsense capability of dialogue systems.

Recognizing the limitations in existing commonsense evaluation, resources, and models for dialogue, this dissertation aims to address each in turn, culminating in a novel dialogue model that leverages a powerful predictive social commonsense generator and whose capabilities can be reliably evaluated. By proposing new approaches to improve commonsense evaluation of dialogue models, commonsense inference prediction for dialogue, and commonsense integration into dialogue models, this dissertation contributes to the advancement of open-domain dialogue modeling.

1.2 Central Thesis and Research Questions

The central thesis of this dissertation is that social commonsense knowledge serves as a rich reservoir of semantic material for open-domain dialogue. To be effective as semantic material for dialogue responses, social commonsense inferences must be multi-faceted (i.e. are diverse and cover a wide range of types) and must be contextually novel (i.e. predictive of new information with respect to the dialogue context). When this is true, then social commonsense can provide a spectrum of predictive ideas that dialogue models can use for generating highly

contextualized follow-up responses. By considering such social commonsense inferences as a pool of content candidates for response generation, dialogue models can produce responses that are not only coherent and contextually appropriate but also exhibit a deeper understanding of content of the conversation, like that observed in Figure 1.1.

This dissertation addresses the central thesis by addressing the following questions:

Research Questions	
1. Does a binary behavior-based evaluation framework result in more reliable and informative evaluation of dialogue model characteristics?	Ch. 3
2. Is it possible to reliably obtain multi-faceted and contextually novel social commonsense inferences for dialogue?	Ch. 4
3. Does access to a multi-faceted and contextually novel social commonsense pool improve dialogue responses?	Ch. 5

1.3 Research Contributions

In summary, the research contributions of the studies in this dissertation are as follows.

1.3.1 Commonsense Evaluation for Dialogue Models (Ch. 3)

I propose Annotating Behaviors in Chat Evaluation (ABC-Eval), a binary behavior-based evaluation of dialogue model characteristics. It is a human evaluation method that specifically measures the rate of behaviors including commonsense contradictions, irrelevant responses, and self-contradictions. Compared to previous evaluation methods, my research shows that ABC-Eval offers an improved method for quantifying dialogue model characteristics, including commonsense capability of dialogue models. The contributions are:

- Design of 16 behavior labels for dialogue model evaluation, along with training exercises

with automatic feedback for annotator onboarding.

- Enhanced interpretability, predictive power for dialogue quality, and sensitivity to between-bot differences relative to existing Likert and Comparative evaluations.
- Capability to clearly reveal that commonsense contradictions are one of the largest sources of errors produced by neural dialogue models.

1.3.2 Commonsense Dataset for Dialogue Models (Ch. 4)

I introduce ConvoSense, a dataset rich in commonsense inferences tailored for dialogue that is automatically curated using Large Language Models. Using this ConvoSense dataset, this study also trains smaller yet still highly capable generative commonsense inference models for dialogue. The contributions are:

- Dataset of commonsense inferences for dialogue that surpasses existing datasets in terms of contextual novelty and is competitive in terms of inference reasonability.
- Improved coverage of diverse commonsense inferences for dialogue, which is crucial for representing the multi-faceted nature of commonsense reasoning.
- A generative commonsense inference model that surpasses existing models in both novelty and reasonability of the outputted inferences.

1.3.3 Commonsense-Augmented Dialogue Model (Ch. 5)

I present Commonsense Inference Generate-Select-Respond (CSI-GSR), a novel three-stage approach based on Large Language Models (LLMs) for generating dialogue responses from ConvoSense inferences, and compare it against the native dialogue capabilities of prominent LLMs, including ChatGPT3.5 and Llama2. The contributions are:

- Enhancement of response specificity and informativeness while maintaining overall response quality.

- Evaluation of dialogue capabilities of native LLMs, including commonsense contradictions, response relevance, and response informativeness.

1.4 Organization

The remainder of this dissertation is organized as follows. Chapter 2 summarizes existing work in commonsense for dialogue, covering resources, models, and evaluation. Chapter 3 presents ABC-Eval, the behavior-based dialogue evaluation method that introduces a reliable commonsense evaluation for dialogue. Chapter 4 presents ConvoSense, the dataset of commonsense inferences for dialogue that achieves the best contextual novelty and coverage. Chapter 5 presents CSI-GSR, a novel select-and-respond approach to commonsense-augmented dialogue models that leverages ConvoSense inferences, and compares it against the native dialogue capabilities of prominent LLMs. Chapter 6 concludes the dissertation.

Chapter 2

Background

2.1 Large Language Models

Large Language Models (LLMs) have emerged as a transformative force in the field of natural language processing, exhibiting an unprecedented ability to perform a wide array of tasks through instructional prompting [22, 39, 114, 51]. LLMs utilize an auto-regressive Transformer architecture [89] which is pretrained on a massive corpus of self-supervised data, where the primary training objective is to predict the next word given previous context. The advancements of modern LLMs arise from the size of their pretraining data and the size of the models themselves (billions to hundreds of billions of parameters) [85]. The greatest success to date has come from closed-source LLMs [92], such as OpenAI’s GPT [5] and Google’s Gemini [84], although much effort is being put into developing competitive open-sourced models, such as Meta’s Llama2 [85]. LLMs are often finetuned on instruction datasets, such as ChatGPT and Llama2-Chat, which teach the LLM to operate in a collaborative interface rather than in a text-completion interface. Special finetuning techniques are employed to optimize instruction-finetuned LLMs, such as Reinforcement Learning with Human Feedback which improves generation capabilities by aligning outputs to human preference data [61]. These finetuning techniques result in LLMs with high performance on completing requested

tasks given just a natural language prompt that contains textual descriptions, instructions, and optional examples of the desired outcomes.

LLMs have demonstrated remarkable capabilities in various dialogue applications. For one, studies have shown that LLMs can create datasets that are on par with, or even superior to, established datasets such as DailyDialog, BlendedSkillTalk, and TopicalChat [38, 9]. This ability not only highlights the inherent dialogue generation proficiency of LLMs but also underscores their potential as efficient tools in dialogue system development. Furthermore, the application of LLMs as dialogue agents has also seen significant advancements, often surpassing the performance of state-of-the-art models that were specifically trained on dialogue datasets. For instance, providing GPT-3 Davinci with in-context examples of empathetic dialogues has yielded better outcomes than models trained directly on empathetic dialogue datasets [41]. Furthermore, a modular approach to harnessing LLMs for dialogue generation has been explored, wherein the dialogue generation process is segmented into multiple phases, such as dialogue history summarization and information retrieval. Research indicates that such modular frameworks achieve superior performance compared to their non-modular counterparts [40, 106, 50, 91].

2.2 Commonsense Resources for Dialogue

Early work on commonsense focused predominantly on lexical, taxonomic, semantic, and physical properties related to entities and verbs, such as the widely recognized WordNet [12], FrameNet [2], and ConceptNet [81].

WordNet Fellbaum [12] present a large database of the English language, WordNet, covering nouns, verb, adjectives, and adverbs. WordNet groups English words into sets of synonyms called synsets, which are characterized by concise definitions and contain semantic relations to other synsets. These semantic relations cover categories including definitional (synonymy, antonymy), taxonomic (hyponymy/hypernymy), physical (meronymy/holonymy),

manner (troponymy), and causal (entailment).

FrameNet Baker et al. [2] capture the linguistic frames underlying word usage in natural language in their FrameNet resource. A “frame” refers to a conceptual structure that represents a particular scenario, situation, or concept. This structure includes a collection of slots that represent the roles that different elements play in the concept being represented. For example, in the frame of “Eating”, the slots include the eater, the food, the manner of eating, and so on. Each frame is associated with a number of words that evoke it, which are called “lexical units”. FrameNet also records relationships between the frames, covering temporal relationships like precedence, taxonomic relationships like inheritance, and causal relationships like causative and inchoative.

ConceptNet Speer et al. [82] present a large-scale general knowledge graph that connects words and phrases using labeled edges, or relations. It was collected from many sources such as expert-created resources, crowd-sourcing, and games with a purpose, and contains 36 relations. The information contained in ConceptNet is predominantly on the concept and entity-level, covering lexical, taxonomic, and physical properties of different words [32].

Although informative and often used for earlier dialogue approaches, especially ConceptNet, these resources encounter challenges towards their application to dialogue specifically. For one, they present abstracted representations of discussed concepts that require additional sophisticated processing to link to the specific contextualized scenarios, events, and situations described in dialogue. Additionally, these resources do not have substantial coverage of social commonsense which is highly relevant to dialogue. In light of these challenges, follow up efforts have sought to better capture commonsense specifically for dialogue. In particular, efforts have shifted toward building datasets encompassing social- and event-based commonsense, such as ATOMIC [32], and for curating commonsense inferences specifically tailored for dialogue contexts, such as Cider [20], ComFact [19], Cicero[21], and Reflect [111].

ATOMIC Sap et al. [77] construct the original ATOMIC dataset specifically to cover the social aspects of commonsense that are not well-covered by previous resources. They focus on commonsense types of prerequisites, personal characteristics, event effects, emotional reactions, personal desires, intentions, and motivations in the context of short phrase descriptions of events that can happen in people’s lives (e.g. “PersonX went for a run”). Various versions of ATOMIC have since been developed, expanding its coverage.

All ATOMIC datasets structure their commonsense as a premise (which is a short phrase describing some event), an inference type, and a reasonable inference that fits the definition of said type (which is also a short phrase like the premise). The highest quality ATOMIC dataset is **ATOMIC10X_{high}** which was constructed from filtered results from GPT3 [94]. Human-generated examples from the precursor dataset **ATOMIC2020** [32] were used as in-context learning examples in the GPT3 prompt to produce new commonsense inferences for each type. A classifier was developed to predict whether a commonsense inference was reasonable for its premise by training on 10,000 GPT3 generations annotated by humans for their reasonability. This classifier was then used to filter the large-scale GPT3 generations to produce the final **ATOMIC10X_{high}** dataset.

ATOMIC has the disadvantage of a lack of contextual awareness as its commonsense tuples 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 [62, 35].

COMET Bosselut et al. [4] develop a generative model of commonsense knowledge, termed COMET. COMET is a Transformer language model which is trained to produce the phrase tail of a commonsense knowledge tuple given the phrase head and the relation. Hwang et al. [32] use the ATOMIC2020 dataset to train a COMET model, with best performance observed when BART is used for the base model.

Cider Ghosal et al. [20] present CIDER, a dataset in which pairs of utterances in a dialogue

context are labeled by humans with derivable commonsense inferences. For example, in the pair of utterances (“SpeakerA: Why were you late for work?”, “SpeakerB: I missed the bus”), the event “missed the bus” would be linked to the event “late for work” with a causal inference type. By design, Cider arises from a span-based extraction annotation setting and is restricted to contain only those inferences that are explicitly revealed in the dialogue utterances.

ComFact Gao et al. [19] map dialogue utterances to reasonable inferences from the existing ATOMIC2020 dataset [32] by using exact string matching and embedding similarity. Subsequently, human annotators verify the relevance of the retrieved inferences.

Cicero Human participants are tasked with composing responses to five social 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 of Cicero produces a single inference for each example [21], whereas the second version produces multiple examples of both good and bad inferences [78].

Reflect Zhou et al. [111] supplies both human-generated social commonsense inferences covering five types and next utterance responses that could be derived from a specified commonsense inference. The inferences are collected by instructing human participants to answer a commonsense question, while the next-utterance responses are composed by new human participants who are provided with the dialogue context and one of the human-generated inferences.

For dialogue-centric commonsense datasets (Cider, ComFact, Cicero, Reflect), 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

Type	Label(s)	Definition(s)	COM	CIC	REF
Subsequent	isBefore Subsequent- Events	What could happen after this? [2] What subsequent event happens or could happen following the Target? [3] What might happen after? [4]	*	22K	600
Antecedent	isAfter	What could have happened before this? [2] What might have happened before? [4]	*		600
Cause	xReason Cause	What could be the cause of this event? [2] What is the event that directly causes or could cause Target? [3]	80	21K	
Prerequisite	xNeed Prerequisites	What does X need to do before the event can happen? [1] What is or could be the prerequisite of Target? [3]	1K	10K	
Motivation	xIntent Motivation	Why does X cause the event? [1] What is an emotion or basic human drive that motivates or could motivate Target? [3]	800	12K	
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] What is Speaker feeling now? [4]	300		600
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] What is Responder feeling now? [4]	70	6K	600
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 2.1: 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. [77], [2] Hwang et al. [32], [3] Ghosal et al. [21], [4] Zhou et al. [111]. Counts are truncated to the nearest order of magnitude. * indicates the type was included but no human-verified instances of it are present.

of relations, they are limited to a small number of dialogues [19], whereas others capture a large number of dialogues but on a limited set of relations [21]. Table 2.1 summarizes the distribution of different inference types in three of the dialogue-focused commonsense datasets and presents a mapping of synonymous types among them based on the provided definitions.

In addition, a few challenges exist in these datasets. For example, the inferences in these datasets are often too succinct and derive only straightforward conclusions with minimal elaboration [19], which do not convey implicit commonsense. Furthermore, 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 [21]. Moreover, although multiple plausible inferences can be drawn from a single dialogue context, only a few datasets support this multifaceted nature [78], impeding the development of models capable of generating diverse inferences, and thus, limiting their utility in real applications.

2.3 Commonsense Augmentation for Dialogue Models

There have been many efforts to integrate commonsense into dialogue models, encompassing a variety of sources of commonsense and strategies of integration.

ConceptNet-grounded Dialogue Models A popular direction in existing work is to deduce topic shifts based on the concepts that are connected through commonsense relations to terms previously mentioned in the dialogue. These works tend to use existing commonsense knowledge graphs, primarily ConceptNet, and retrieval-based lookups to obtain commonsense tuples that are within one or two hops to the words in the dialogue context. A defining element of these works is that the decoder can choose to either generate a word from the base vocabulary or choose to copy a specific word from the retrieved commonsense. Some works use graph attention to integrate the retrieved commonsense (CCM [107], ConceptFlow [102], CARE [105], KEMP [43], CAKF [99], CNTF [88], ECCF [44]) whereas others linearize the retrieved tuples to treat them as additional text input (ConKADI [97], SAKDP [98]). Additionally, some works further modify the commonsense resource using statistical properties of dialogue datasets to add additional tuples representing common history-response concept-pairs [44] and emotion-related concepts [105], or extract a subset of ConceptNet concepts that are related to common life stressors [31]. Approaches such as these that are based on retrieval from static commonsense knowledge resource are limited by the coverage of the resource. Zhou et al. [113] attempt to mitigate this constraint of static ConceptNet by training a model to generate intermediate commonsense assertions before generating the dialogue response, using ConceptNet commonsense assertions mapped to dialogue contexts in existing datasets

using exact string matching and embedding similarity.

ATOMIC-grounded Empathetic Dialogue Models Recently, a few works have used the generative model COMET trained on ATOMIC as the source of commonsense information in order to better represent the social and experience-oriented nature of conversations with human users. However, all of these models target either empathetic or emotional support response generation specifically, rather than open-domain responses. Empathetic response generation tasks the model with recognizing and acknowledging the emotional state of the user based on what they just said and producing supportive reactions [69]. Similarly, emotional support conversations task the model with identifying the problem that the user is facing and utilizing an appropriate psychology-grounded interaction strategy to comfort, encourage, and aid the user in coping with their problems [48]. Both of these tasks narrow the scope of possible responses that are targeted by such dialogue models and represent only a subset of the interactions that can occur in natural open-domain conversations. Works including Sabour et al. [74], Tu et al. [86], and Liu and Kilicoglu [49] use encoded outputs from various commonsense types from COMET-ATOMIC as additional input with the dialogue context, either through direct text concatenation or concatenation of their embeddings, used to train the dialogue model. On the other hand, Fu et al. [18] develop a ChatGPT-based dialogue agent that ingests COMET-ATOMIC inferences. They use COMET-ATOMIC to generate two user inferences (emotion, desire) and to generate in-context learning examples for two system inferences (emotion, intention) by running on the ground truth responses. Then, the user inferences and in-context learning examples are inputted in the prompt to ChatGPT, which is instructed to output the appropriate system inferences and a final response for the given dialogue context.

Social-commonsense-grounded Open-domain Dialogue Model Only a couple of recent works have attempted to integrate social commonsense to open-domain dialogues. Zhou et al. [111] collect Reflect, a dataset of human-written commonsense inferences for

five commonsense types for dialogue contexts (Section 2.2) and corresponding responses which are human-written under instruction to express a provided inference. They finetune BlenderBot on their collected dataset of commonsense-guided responses where the input is a dialogue context followed by one commonsense inference question and the output is the answer followed by the response. However, they find that finetuning BlenderBot to be conditioned on the inference question does not improve human perception of quality of responses against BlenderBot finetuned on their data to output the response directly. Furthermore, they find that prompting GPT3-davinci to output responses by thinking about one commonsense inference question with few-shot in-context learning examples performs best when compared to their finetuned models. Through evaluation performed by Amazon Mechanical Turkers in which the sensibility, specificity, and interestingness of the generated responses are binarily annotated, they observed notable improvements over native GPT3 response generation. Specifically, sensible responses increased by 6%, specificity by 14%, and interestingness by 9% across 50 examples. Chae et al. [8] take a different approach, employing ChatGPT3.5 to generate rationales linking dialogue context to responses, where each rationale is composed of three question-answer pairs where the questions were derived from 11 types of commonsense reasoning. After filtering the rationales produced by ChatGPT using two automatic measures of the suitability of the rationale for response generation of a given dialogue context, they trained an OPT1.3B model to generate rationales for dialogue contexts, and used the generations of this model as additional input to ChatGPT3.5 in a prompt designed for response generation. Their evaluation, conducted via MTurkers, demonstrated statistically significant enhancements in naturalness (67 vs 33) and specificity (58 vs 42), but not for consistency (53 vs 47) or engagingness (52 vs 48) against responses from native ChatGPT3.5. Both of these approaches restrict the consideration of commonsense in a greedy fashion when generating a response, failing to capitalize on the inherent multi-faceted nature of commonsense reasoning for dialogue situations.

2.4 Commonsense Evaluation for Dialogue Models

Most works on commonsense-grounded dialogue models do not perform an explicit evaluation of the commonsense properties of the generated responses. Instead, they evaluate the generated responses in terms of various characteristics, such as relevance to the context and specificity. Although such characteristics may be influenced by the integration of commonsense, they encompass broader concepts that represent more than just the commonsense properties.

For commonsense-specific evaluations, two previous works utilize two different evaluation procedures, each with their own concerns. Zhou et al. [113] use a pairwise comparison evaluation to determine which responses follow commonsense more although they do not provide any definition of commonsense to their annotators and the eliciting question that they use conflates commonsense with appropriateness and relevance to the context. Zhou et al. [110] instruct human annotators to rate the commonsense plausibility of responses on a scale from 1 to 10, but their evaluations produce concerningly large variance in the scores, calling into question the validity of their evaluation procedure.

Chapter 3

Annotation of Behaviors in Chat

3.1 Introduction

In any dialogue system development, assessing the system’s performance on relevant aspects is crucial. Take, for instance, evaluating the commonsense understanding of the bot illustrated in Figure 3.1. The bot’s responses include statements like building a life-size airplane model (Turn 3) and gifting it to a brother who dislikes planes (Turn 5), both of which seem implausible. These instances indicate a lack of commonsense understanding, as they conflict with innate knowledge of likely situations in the world. This capability of pointing to specific responses that express a lack of commonsense understanding gives rise to the idea that it is possible to evaluate the commonsense capability of this bot by labeling all of the turns in which it produces a commonsense mistake. Under this, a concrete measure of the expected rate of commonsense mistakes made by a bot can be obtained, which is intuitively interpretable as an indication of its commonsense capability. This principle is the motivation behind the first research question: does a binary behavior-based evaluation framework result in more reliable and informative evaluation of dialogue model characteristics? In particular, the aim is to accurately measure distinct attributes of a dialogue model pertinent to its effectiveness as a conversational agent, thereby achieving a nuanced and multi-dimensional

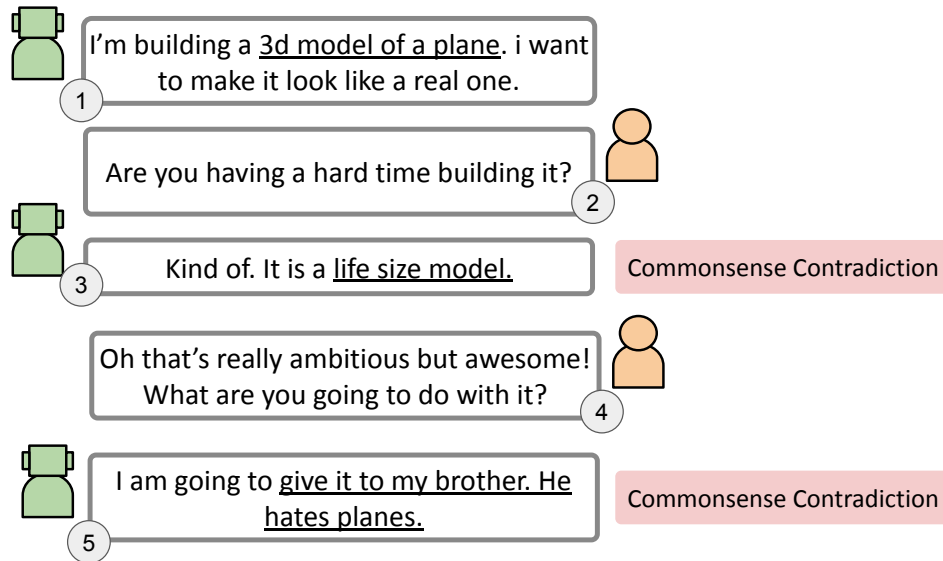


Figure 3.1: Example of commonsense contradiction labels in human-machine dialogue.

evaluation of its performance. Such an evaluation is intended to be complementary to holistic evaluations of overall model performance (e.g. overall quality on a Likert scale) as it enables a higher granularity into the strengths and weaknesses of the model that help to understand and explain the human perception of the quality achieved by a specific dialogue model [87].

In this chapter, I present the Annotation of Behaviors in Chat Evaluation (ABC-Eval), a comprehensive evaluation framework for dialogue models that targets fine-grained response assessment which is developed in collaboration with another graduate student [17]. ABC-Eval is an evaluation protocol that measures rates of important dialogue model behaviors, including irrelevant responses, self contradictions, and factual hallucinations. Importantly, ABC-Eval enables tangible evaluation of the commonsense capability of dialogue models through measuring the rate of commonsense contradictions expressed in the outputted responses. Through extensive statistical analyses, the behavior labels from ABC-Eval are shown to achieve high inter-annotator reliability, discriminate differences between different dialogue models, and capture large degrees of variance in human perception of overall response quality. Following these analyses, ABC-Eval is verified to enable reliable and informative

fine-grained assessment of dialogue model characteristics that contribute useful information related to overall dialogue quality.

3.2 ABC-Eval Development

Typically, when dialogue models are evaluated for fine-grained characteristics, such as their capability for emotional understanding, conversational consistency, or response relevance, these characteristics are defined and evaluated in a high-level manner. The most commonly evaluated dimensions and their definitions as provided to evaluators are shown in Table 3.1, as synthesized in my work on surveying the evaluation procedures of open-domain dialogue works accepted to top-tier Natural Language Processing venues [14]. The evaluations typically follow one of two methodologies, either Likert ratings or pairwise comparisons. Likert rating evaluation instructs human annotators to rate the dialogue system responses on a set of dialogue dimensions using numeric scales, most often 1 to 5. Pairwise comparisons instruct human annotators to select the most fitting response on a set of dialogue dimensions when shown multiple options generated from different dialogue models.

Dimension	Definition
Grammaticality	Responses are free of grammatical and semantic errors
Relevance	Responses are on-topic with the immediate dialogue history
Informativeness	Responses produce unique and non-generic information that is specific to the dialogue context
Emotional Understanding	Responses indicate an understanding of the user’s current emotional state and provide an appropriate emotional reaction based on the current dialogue context
Engagingness	Responses are engaging to user and fulfill the particular conversational goals implied by the user
Consistency	Responses do not produce information that contradicts other information known about the system
Proactivity	Responses actively and appropriately move the conversation along different topics

Quality	The overall quality of and satisfaction with the dialogue

Table 3.1: The final set of dialogue dimensions for human evaluation from Finch and Choi [14]. Dimensions above the dashed line are dimensional characteristics of dialogue models. The **Quality** dimension below the dashed line is an overall performance assessment.

One major limitation of these common evaluation procedures is that they fail to provide detailed information on the dialogue models under study. Obtaining a decimal score or a win percentage on one of the typical dimensions reveals some information on the response

characteristics achieved but does not uncover the specific behaviors that are going right or wrong for a particular dialogue model. For the dialogue shown in Figure 3.1, it could be indicated that this bot receives a score of 2 out of 5 for commonsense, which follows the typical method of evaluating commonsense. However, it is difficult to say exactly what such a numerical score means, other than some abstract assessment of low commonsense. Instead, by following the behavior labeling approach, specific turns of the bot are identified that express commonsense problems, thereby obtaining a concrete measure of the expected rate of commonsense mistakes made by a bot which is intuitively more informative as an indication of its commonsense capability. As a result, this motivates the utility of a behavior evaluation procedure for dialogue models and leads to the development of ABC-Eval.

ABC-Eval is developed in two stages: (1) collecting a set of behavior label candidates and (2) developing and piloting the annotation instructions and procedure.

3.2.1 Collecting Behavior Label Candidates

Based on a review of recent work in open-domain dialogue modeling and evaluation, the characteristics of chatbot responses related to conversation quality are identified. These characteristics include those presented as error cases, evaluation metrics, or desirable response features. Due to its high coverage of error categories, Higashinaka et al. [28] is the primary source of inspiration for many of the behavior labels. However, their presented taxonomy is further improved by considering additional labels based on characteristics of chat presented by other work, and by further refining their error categories for clarity. The development process of these behavior label candidates is discussed next.

3.2.2 Pilots and Development

The labels, annotation instructions, procedure, and software application for ABC-Eval are created using an iterative process of development and piloting. The goals of this development process are to curate the behavior labels to be included in ABC-Eval and to identify the best

annotation interface for these included behavior labels for the most reliable annotations.

14 students are invited to serve as evaluators for piloting the evaluation. The pilots are performed on conversations collected between a high-performing neural dialogue model (Blender [73]) and one of the ABC-Eval developers. To avoid overfitting the evaluation design, a new set of conversations is used for each pilot round. Three pilot rounds are conducted, making revisions to the evaluation setup after manually reviewing each round’s annotation.

Pilot 1 consists of a single annotation task, which covers 17 behavior labels. Annotators are given instructions and definitions for all 17 behavior labels and instructed to label the turns of a provided dialogue with all of the labels that are applicable. A single dialogue is annotated by two to three different annotators, and four dialogues in total are used in this pilot, each annotated by a different group of annotators. Students of both the graduate level and undergraduate level participate in this pilot. The labels are: `uninterpretable`, `antisocial`, `redundant`, `irrelevant`, `correct fact`, `incorrect fact`, `commonsense contradiction`, `correct profile`, `contradict profile`, `correct self context`, `contradict self context`, `correct partner context`, `contradict partner context`, `request`, `present`, `ignore request`, and `ignore present`.

Pilot 2 consists of six annotation tasks, each covering a different set of behavior labels. The instructions and definitions have been updated from Pilot 1. Annotators are given these new instructions and definitions for the set of behavior labels for a particular annotation task, and instructed to label the turns of a provided dialogue with those labels that are applicable. Each annotation task is applied to a different dialogue, where at least two annotators perform the annotation task. The same annotator is allowed to perform more than one annotation task, although not all did. Students of both the graduate level and undergraduate level participate in this pilot. The six annotation tasks are: interpretability (`uninterpretable`), conversational consistency (`correct profile`, `contradict profile`, `contradict self context`, `contradict partner context`, and `redundant`), sociality (`antisocial`), ency-

clopedic knowledge (`correct fact`, `incorrect fact`), commonsense (`commonsense contradiction`), and turn-taking (`request`, `present`, `follow-up`, `topic-switch`, `irrelevant`, `ignore`).

Pilot 3 follows a similar setup to Pilot 2. There are three major differences. First, all annotators perform all of the annotation tasks. Second, the instructions and definitions have been updated, based on the results from Pilot 2. Finally, training exercises have been implemented for each annotation task in order to teach the annotators how to identify each behavior label before they then perform the annotations on the desired dialogues. It should also be noted that only undergraduate students participate in this pilot.

	Dialogues	Annotators		
	#	#	Type	α
Pilot 1	4	11	ALL	0.18
	2	4	G	0.39
Pilot 2	6	5	ALL	0.49
	6	4	G	0.50
Pilot 3	6	4	U	0.43
	6	4 (screened)	U	0.45

Table 3.2: The distribution of dialogues and annotators for each evaluation pilot. Dashed line delineates a subset (below line) of the total annotations (above line) produced in that pilot. α : Krippendorff’s alpha, U: Undergraduate annotators, G: Graduate level annotators, ALL: Undergraduate and graduate annotators.

Table 3.2 presents a summary of the pilot rounds. It is important to note that these piloting rounds are not necessarily directly comparable to one another when taken as a whole, since the annotators and dialogues to be annotated varied between each round. Instead, I will discuss below the major takeaways afforded by different splits of the pilots that informed the final design of ABC-Eval.

Task Formulation The decision to format ABC-Eval into several small annotation tasks, each with a tailored subset of the behavior labels, is made from the results of Pilot 1. In Pilot 1, there was only a single annotation task, in which the annotators had to consider all labels at the same time. The overall interannotator agreement is quite low ($\alpha = 0.18$), which is concerning. Based on ad-hoc feedback from the pilot annotators, the consensus is that this setup demanded an unreasonable cognitive load on annotators due to the large number of labels to keep track of.

For Pilot 2, the behavior labels are split into groups based on their similarity and each group is structured as a new annotation task. Table 3.2 shows the boost to interannotator agreement between Pilots 1 and 2. However, this agreement increase could have resulted from an increase in the quality of the annotators (as Pilot 2 is composed primarily of annotators with a graduate-level education whereas Pilot 1 is more evenly split between annotators with an undergraduate-level education and graduate-level education). To remove this confound, the agreement in Pilots 1 and 2 is also calculated for only graduate-level annotators. Although it is less dramatic, there remains an increase in agreement from 0.39 to 0.50, which encourages the decision to maintain the smaller annotation tasks. Dividing the annotation into tailored tasks seems to reduce the cognitive load on annotators, thus allowing them to perform more accurate annotations per task with minimal overhead.

Training and Screening Manual analysis of the pilot annotations from Pilots 1 and 2 reveal some recurring annotation mistakes, arising from misunderstandings of the detailed guidelines for the tasks. In an attempt to correct such misunderstandings, a three-round training procedure is introduced for each task, where an annotator must complete the training before being allowed to participate in the evaluations.

Each round of training for a particular annotation task consists of one curated conversation with ground-truth labels and explanations that are shown as feedback to the annotator after they complete the training round. The results of Pilots 1 and 2 are used in order to develop

these training conversations as follows:

1. **Label Specification Construction:** A label specification is constructed that identifies frequent dialogue situations or patterns that reveal either a positive or negative expression of the behavior label. This specification aims to provide a definition of the label that is example-based, by exemplifying different dialogue cases and how they should be annotated for a particular behavior label. These cases focus on covering the turns which received a high rate of incorrect annotations from the annotators in the pilot rounds. These specifications are not necessarily exhaustive or theoretically complete; instead, this serves as an instructional guideline for the commonly observed misconceptions.
2. **Training Conversation Selection:** Three new conversations between Blenderbot and a human (from a collection within my lab) are selected for each annotation task to be used as training conversations for it. This selection is manually done by ranking the conversations on their coverage of the cases identified in the label specifications relevant to the annotation task.
3. **Training Conversation Modification:** The selected conversations are heavily revised by hand to ensure that all of the cases identified in the specification are adequately represented, most often by inserting new utterances that correspond to underrepresented cases.

To evaluate the utility of this training process, a third pilot is conducted using four undergraduates. It is observed that there is a general upwards trend in annotation performance between the training rounds for the annotators, suggesting that the training is aiding in the annotation accuracy for the annotators. The final agreements are calculated in two manners, in order to quantitatively assess the impact of the training. First, the annotations from all annotators are considered, regardless of annotator performance on the training, resulting in an agreement of 0.43. Second, only those annotations from annotators who scored highly in

the training are considered. For this, annotators were filtered out of each task based on the number of mistakes that they made in the training. The filtering criteria was a maximum of 1 mistake on the training for interpretability and sociality annotation tasks and a maximum of 2 mistakes for the other tasks. After filtering, the agreement is slightly improved at 0.45.

Due to the small nature of this pilot, it cannot be concluded whether this difference is meaningful. However, ad-hoc feedback from the annotators suggested that the training rounds are useful towards their understanding of the tasks, although the amount of training did increase the overall workload of participation.

Pilot Conclusion At the end of these three pilot rounds, it is noticed that emotional understanding and bot character depth are not well represented in the piloted behavior labels. To remedy this, two additional annotation tasks are developed and tested outside of the previous pilots. These two annotation tasks consist of the labels `empathetic` and `lack of empathy`, and `preference info` and `life info`, respectively. In addition, the decision is made to exclude a number of low-performing and lower priority behavior labels in order to minimize the overall scope and cost of this behavior-based evaluation framework, including `use profile`, `contradict profile`, `request`, and `present`. The next section will describe the final ABC-Eval design in full, based on the incorporation of all of the decisions discussed for the pilots.

3.2.3 Final ABC-Eval Design

Label	Abbr.	Description	Inspired by
Uninterpretable	!Int_b	It is difficult to understand the intended meaning of part or all of the response.	1, 2, 3, 4, 5, 6
Antisocial	!Soc_b	The response is insulting, hateful, or excessively vulgar.	2, 7, 8, 9
Preference Info	Pre_b	The response expresses the bot’s preferences, wishes, or values.	10, 11
Life Info	Lif_b	The response shares information about the bot’s life or experiences.	
Empathetic	Emp_b	The response shows an understanding and reacts appropriately to someone’s emotions.	11, 12, 13
Lack of Empathy	!Emp_b	The bot misunderstands or reacts inappropriately to someone’s emotions.	
Commonsense Contradiction	!Com_b	The response misunderstands or contradicts common knowledge.	2, 14, 15, 16
Fact Usage	Fac_b	The response accurately incorporates encyclopedic or expert knowledge.	1, 2, 11, 17, 18, 19, 20
Fact Contradiction	!Fac_b	The response hallucinates or inaccurately presents encyclopedic or expert knowledge.	
Self Contradiction	!Sel_b	The bot contradicts something it said earlier in the dialogue.	2, 3, 6, 20, 21, 22, 23
Partner Contradiction	!Par_b	The bot contradicts or misremembers something the user said earlier in the dialogue.	
Redundant	Red_b	The response inappropriately repeats information presented earlier in the dialogue.	
Ignore	Ign_b	The response ignores what the user just said.	1, 2, 3, 6, 24
Irrelevant	!Rel_b	The response interrupts the current topic of discussion by presenting unrelated information.	
Follow-up	Fol_b	The response explores, elaborates on, or asks about the ideas shared in the previous turn.	
Topic Switch	Top_b	The response introduces a new topic of conversation.	

Table 3.3: The 16 behavior labels within ABC-Eval. Row separators denote annotation task groupings. [1] Gopalakrishnan et al. [24], [2] Higashinaka et al. [28], [3] Mehri and Eskenazi [56], [4] Mehri and Eskenazi [57], [5] Phy et al. [64], [6] Sanguinetti et al. [75], [7] Beattie et al. [3], [8] Sun et al. [83], [9] Xu et al. [100], [10] Rashkin et al. [70], [11] Smith et al. [80], [12] Majumder et al. [52], [13] Rashkin et al. [68], [14] Zhong et al. [104], [15] Zhou et al. [108], [16] Zhou et al. [112], [17] Gupta et al. [26], [18] Honovich et al. [29], [19] Santhanam et al. [76], [20] Shuster et al. [79], [21] Li et al. [45], [22] Nie et al. [60], [23] Welleck et al. [93], [24] Xu et al. [101] .

Based on the piloting and development work described previously, the final ABC-Eval procedure includes 16 binary behavior labels divided between 8 independent annotation tasks, where each annotation task has corresponding training materials. Table 3.3 presents the names and definitions of the 16 behavior labels, and highlights the previous works that inspired the development of each behavior label. In each annotation task, human evaluators are provided with definitions and examples of the behavior labels associated with that task and asked to annotate every chatbot turn in a given human-chatbot conversation with each behavior label, if it is applicable to that turn. Evaluators complete these tasks using a custom web application based on the ParlAI evaluation interface. The 8 annotation tasks are described next.

1. **Interpretability** covers one label: **uninterpretable**, which identifies nonsensical language in the turns. The interface used for this annotation task is shown in Figure 3.2.

Interpretability of Dialogue Responses

Identify all uninterpretable responses.

A response is uninterpretable if it:

- is highly ambiguous and does not have one specific and clear meaning
- takes a lot of effort to understand what the intended meaning is
- contains any phrase that is difficult to derive meaning from, usually due to the use of nonsense words or grammar problems

Pay careful attention and DO NOT MARK if the problem is that the response...

- is illogical or contradicts common knowledge
- does not make sense as an answer to the previous question
- is off-topic or contradictory to the context
- contains a phrase that is technically ungrammatical, but conveys a clear meaning in the context
- uses a made-up word that is easily and completely interpretable in the context
- uses slang words or phrases. If you are unsure whether some phrase is slang, you should look up the phrase online (e.g. on urban dictionary) to check whether the response is interpretable.

Help

Sam: I saw a great movie yesterday.
Alex: I like.

Uninterpretable: Alex's response is incomplete.

Sam: I think pizza is the best food.
Alex: No, ramen.

Not uninterpretable: Even though Alex's response is not a complete sentence, it represents a complete thought ("No, ramen is the best food") because of the context.

Alex: I just got back from swimming at the beach but I haven't left my house yet today so I haven't gone swimming yet.

(Sam) hi

(Alex) hi, how are you today? I'm having a hard time keeping up with my life.
 This response is uninterpretable.

(Sam) Oh no, you sound stressed out a bit. What is giving you a hard time?

(Alex) I get into a lot of fights, especially when.
 This response is uninterpretable.

(Sam) Fights with people you know, like family or friends?

(Alex) mostly friends and family. They are tired of dealing with my anger issues.
 This response is uninterpretable.

(Sam) I hope you can get along better with the people you care about soon.

(Alex) Thanks, I am working on it. I want to better see control bad. So, what do you like to do?
 This response is uninterpretable.

(Sam) I play baseball.

(Alex) do you play professionally or do you just play baseball?
 This response is uninterpretable.

(Sam) It is just a community league for fun. I am definitely not good enough to play professionally. Do you play?

You are on task 1 / 1 for dialogue 1 / 3

Next

Figure 3.2: Interface for uninterpretable

2. **Sociality** covers one label: **antisocial**, which identifies offensive language in the turns.

The interface used for this annotation task is shown in Figure 3.3.

Antisocial Behaviors in Dialogue

Identify all responses that are insulting, hateful, or excessively vulgar.

A response demonstrates antisocial behavior if any of the following apply. The response...

- clearly insults the other speaker
- is excessively degrading to things, people, or ideas the other speaker values
- uses excessively vulgar, profane, or explicit language
- expresses a clear and disturbing intolerance towards a group of people

The following characteristics do NOT demonstrate antisocial behavior. The response...

- expresses an opinion different from that of the other speaker
- argues with the other speaker in good faith

Note that what constitutes antisocial behavior is subjective, so some responses might be difficult to judge. If you are unsure whether a response demonstrates antisocial behavior, "judge whether most people would find the response to be shockingly inappropriate in its context".

Help

Sam: How are you?
Alex: Bad, because you're so stupid.

Antisocial: Alex is insulting Sam.

Sam: My friend took me to the aquarium yesterday.
Alex: Your friend sounds annoying.

Antisocial: Alex's response is excessively degrading to someone Sam cares about.

Sam: I think pizza is the best food.
Alex: You're crazy.

Not antisocial: Although Alex's response insults Sam if taken literally, the most likely function of Alex's response is to express an opinion, rather than to degrade Sam.

(Sam) hi

(Alex) hey, how are you? i've been surfing all day. what do you do for a living?
 This response exhibits antisocial behavior.

(Sam) I have always wanted to surf but I don't live near an ocean. I am an artist.

(Alex) You must be poor then.
 This response exhibits antisocial behavior.

(Sam) I make enough money to get by. What do you do?

(Alex) I own a pet store. I have to work all the time, which is kind of shit, but I have really great employees that make it a lot better.
 This response exhibits antisocial behavior.

(Sam) that's cool. what kind of pet store do you own?

(Alex) We sell all of the things you need for most common pets, like food, beds, accessories, and toys. do you have any pets?
 This response exhibits antisocial behavior.

(Sam) I have a cat named Dash.

(Alex) dash is a stupid name for a cat. i love cats.
 This response exhibits antisocial behavior.

You are on task 1 / 1 for dialogue 1 / 3

Next

Figure 3.3: Interface for antisocial

3. Character Depth covers two labels: `preference info` and `life info`, which identify turns in which character-like information (preferences, opinions, life experiences, etc.) is shared. The interface used for this annotation task is shown in Figure 3.4.

Personal Information Sharing in Dialogue

Identify responses where the speaker is explicitly sharing either of two kinds of personal information:

- Information about the **preferences or values** held by the speaker
- Information about the speaker's **life**

Responses should not be marked if the identified preference or life information must be inferred based on what the speaker has said.

Preference and value information includes:

- The speaker's likes, dislikes, and favorites
- Attitudes the speaker has that reveal their preferences
- Wishes, wants, and desires held by the speaker
- Principles and values held by the speaker

Information about the speaker's life includes:

- Experiences the speaker has had or might have in the future
- Characteristics about the speaker (e.g. age, location, tendencies, habits, plans, skills)
- Characteristics of people, places, or things directly related to the speaker

Note that for this task, the speaker's thoughts, beliefs, and mood do NOT count as personal information and should not be labeled.

Help

Sam: *What's your favorite movie?*
Alex: *Probably Inception.*

Preference/value info: Alex is sharing her movie preferences by saying Inception is her favorite movie.

Sam: *I got back from Ireland a few days ago.*
Alex: *Ireland is a great place to take a trip.*

Preference/value info: Alex shares an attitude towards Ireland.

Sam: *What have you been up to?*
Alex: *I went to the movies today with my friend.*

(Sam) Hi.

(Alex) Hello, how are you?
 This response shares information about the speaker's preferences/values.
 This response shares information about the speaker's life.

(Sam) I'm in a great mood today. How are you?

(Alex) I'm doing well, I'm glad you're in a good mood today!
 This response shares information about the speaker's preferences/values.
 This response shares information about the speaker's life.

(Sam) Thanks! I just got through my last couple midterm exams today in my college classes.

(Alex) That's good to hear. What are you studying in college? I'm still in high school.
 This response shares information about the speaker's preferences/values.
 This response shares information about the speaker's life.

(Sam) I'm studying psychology.

(Alex) Psychology always sounded interesting to me. What do you want to do with your degree once you graduate?
 This response shares information about the speaker's preferences/values.
 This response shares information about the speaker's life.

(Sam) I really want to get into a PhD program in clinical psychology, but it's really competitive. I'm not sure my grades are good enough.

You are on task 1 / 1 for dialogue 1 / 3

Next

Figure 3.4: Interface for preference info and life info

4. Emotional Understanding covers two labels: **empathetic** and **lack of empathy**, which identify positive and negative cases of emotional appropriateness in the turns, respectively. The interface used for this annotation task is shown in Figure 3.5.

Empathy in Dialogue

Identify responses where the speaker shows an understanding—or lack of understanding—of someone’s emotions, especially the emotions of the person they are talking to.

A response shows empathy when the speaker:

- clearly demonstrates an understanding of their partner’s emotions
- reacts with the appropriate sentiment or emotion to their partner’s shared experience
- understands or appropriately reacts to someone else’s experience or emotions
- appropriately reassures, encourages, or supports their partner

A response displays a lack of empathy when:

- the tone, emotion, or sentiment of the response is clearly inappropriate OR
- the response has an inappropriate lack of emotion

It is possible for a response to be neutral, and neither show an understanding nor be missing an understanding of emotions. In this case, neither category should be selected for the response.

Help

Sam: I graduated college last week, so I've been celebrating.
Alex: I'm sorry to hear that.

Lack of Empathy: Alex's response attempts to sympathize with Sam, which is inappropriate because Sam is sharing good news.

Sam: I graduated college last week, so I've been celebrating.
Alex: Congratulations!

Empathy: Alex's response appropriately congratulates Sam, demonstrating an understanding that Sam feels accomplished.

Sam: My dog passed away last week.
Alex: Was your dog long-haired or short-haired?

Lack of empathy: Alex's response has an inappropriate lack of emotion, and does not sufficiently acknowledge or sympathize

(Sam) Hi.

(Alex) Hello, how are you?

Is this response empathetic?

Yes, the speaker demonstrates an understanding of their partner's emotions.

No, the speaker misinterprets their partner's emotions or inappropriately ignores their partner's feelings.

Not applicable, the speaker neither shows an understanding or misunderstanding of their partner's emotions.

(Sam) I'm good, just a little stressed about work.

(Alex) Oh, I'm sorry to hear that. What do you do for a living, if you don't mind me asking?

Is this response empathetic?

Yes, the speaker demonstrates an understanding of their partner's emotions.

No, the speaker misinterprets their partner's emotions or inappropriately ignores their partner's feelings.

Not applicable, the speaker neither shows an understanding or misunderstanding of their partner's emotions.

(Sam) I'm a software engineer. We have a project deadline coming up is all.

(Alex) That's good to hear. Do you think you'll be able to finish it by the deadline?

Is this response empathetic?

Yes, the speaker demonstrates an understanding of their partner's emotions.

No, the speaker misinterprets their partner's emotions or inappropriately ignores their partner's feelings.

Not applicable, the speaker neither shows an understanding or misunderstanding of their partner's emotions.

(Sam) I don't know. I guess we're about to find out.

(Alex) I wish you the best of luck. I'm sure you'll knock it out of the park.

Is this response empathetic?

Yes, the speaker demonstrates an understanding of their partner's emotions.

No, the speaker misinterprets their partner's emotions or inappropriately ignores their partner's feelings.

Not applicable, the speaker neither shows an understanding or misunderstanding of their partner's emotions.

You are on task 1 / 1 for dialogue 1 / 3

Next

Figure 3.5: Interface for empathetic and lack of empathy

5. **Commonsense** covers one label: `commonsense contradiction`, which identifies turns in which violations of commonsense knowledge are expressed. The interface used for this annotation task is shown in Figure 3.6.

Contradicting Common Knowledge in Dialogue

Identify all responses that contradict common knowledge.

Common knowledge is knowledge that:

- Is learned through direct experience, rather than from reading or being taught
- Almost everyone knows and agrees on

To identify contradictions of common knowledge, judge whether a **vast majority** of people would agree that the response doesn't make sense because the response:

- contradicts common knowledge
- makes unfounded assumptions
- is highly illogical or self-contradictory
- asks a question where the answer is already obvious

It is **NOT** considered a contradiction of common knowledge if the chatbot (Alex) pretends to be involved in human activities (such as eating, having a job, etc.).

Note that you also **SHOULD NOT MARK** responses that don't make sense for reasons other than common knowledge contradiction, such as:

- off-topic responses
- responses that don't have any clear meaning (e.g. overly vague or ill-formed responses)

In some cases, a response might make a claim that is technically possible, but is implausible because of something that is common knowledge. In these cases, **mark highly implausible responses as contradictions of common knowledge ONLY IF the implausible claims are not appropriately explained during the conversation.**

Help

Alex: I went surfing at the lake this weekend.

Contradiction: surfing generally requires ocean waves, making the claim implausible.

Alex: I think Kanye West has done a great job as president.

No contradiction: even though the response is factually inaccurate, it does not violate common knowledge. This is

(Sam) hi

(Alex) hi there, how are you? what do you think about ghosts? i think they are real.
 This response contradicts common knowledge.

(Sam) I'm doing okay. I don't think ghosts are real.

(Alex) i don't think they are real either, but i do believe in them. do you have any hobbies?
 This response contradicts common knowledge.

(Sam) I like to sit on the couch, eat potato chips, and watch tv. I am kind of a geek.

(Alex) You probably like playing sports too. All of my friends do.
 This response contradicts common knowledge.

(Sam) No, not really. I am not that coordinated.

(Alex) Oh, I see. I am not ever going there, but not now and definitely not tomorrow. What else do you like to do?
 This response contradicts common knowledge.

(Sam) I build models, like model airplanes and trains. My basement is kind of being overrun by them at this point.

(Alex) that sounds like a lot of fun. i would love to be able to build something like that.
 This response contradicts common knowledge.

(Sam) Yeah, it is a lot of work, but I just get lost in the process and end up having a blast.

You are on task 1 / 1 for dialogue 1 / 3

Next

Figure 3.6: Interface for commonsense contradiction

6. Encyclopedic Knowledge covers two labels: `fact usage` and `fact contradiction`, which identify positive and negative cases of encyclopedic knowledge being expressed, respectively. The interface used for this annotation task is shown in Figure 3.7.

Incorporating Knowledge in Dialogue

Overview

The goal of this task is to identify when factual knowledge is being used within a dialogue, and to discriminate whether it is an accurate or inaccurate usage.

Instructions

Factual knowledge is information presented about the world that is often learned by reading or being taught.

In order to be considered factual information for this task, the presented information must be:

- **public:** includes entities or facts that are relevant to the broader public, not just personal information.
- **specific:** mentions specific entities or facts, not just broad categories or generalizations.
- **objective:** presents objective information, not just opinions or personal beliefs.

Some examples of factual information include:

- historical or news events
- public figures or organizations
- publications or works of art, including books and movies
- scientific information
- specific named products or services
- etc.

Do not mark a turn as using factual knowledge if the turn could be interpreted as expressing:

- opinions or value judgements
- estimates or predictions
- personal information about the speaker or their partner
- information about things in either speaker's life that are not publicly relevant

It is possible for there to be more than one claim about factual knowledge being made in a turn, in which case you need to consider all relevant claims when completing the task.

Mark usages of factual knowledge as accurate when the claim is true and can be verified by trustworthy sources, and as inaccurate when the claim is false, highly implausible, or clearly misleading.

Help

If you are unsure whether to mark the checkbox or how to answer one of the questions, a help section is provided below for each type of question. Clicking on the help icon next to each question will open and jump to the relevant help section.

(Sam) hi

(Alex) hi, how are you? I love to travel. When did you get back from your vacation?

Alex's response incorporates or assumes at least one fact.

Do you know whether ALL of the facts that Alex uses or assumes are accurate?

Yes, I know for sure ALL facts are accurate.

No, I know for sure that one of the facts is inaccurate, false, or highly implausible.

It is misleading for Alex to claim or assume one of the facts, because there is no way that Alex knows whether that fact is accurate.

I don't know for sure whether ALL of the facts are accurate.

Take 60 seconds to search ALL unknown facts on the internet. Does your search verify or falsify ALL the facts?

ALL facts are accurate; a credible source verified the facts in my search.

One of the facts is inaccurate; a credible source falsified the fact or revealed that it is highly implausible.

My search revealed multiple credible sources that disagreed about whether one of the facts was true.

I couldn't find enough credible evidence in my search to either verify or falsify one of the facts.

(Sam) I love to travel too but I haven't taken a vacation recently. What countries have you visited?

(Alex) I went to India a while ago and it was amazing.

Alex's response incorporates or assumes at least one fact.

(Sam) Where did you go in India when you visited last?

(Alex) I went to Bangalore, which is the capital of India.

Alex's response incorporates or assumes at least one fact.

(Sam) What did you think of Bangalore?

(Alex) I think it's a beautiful city with a lot of history and culture.

Alex's response incorporates or assumes at least one fact.

You are on task 1 / 1 for dialogue 1 / 3

Next

Figure 3.7: Interface for fact usage and fact contradiction

7. **Conversation Consistency** covers three labels: self contradiction, partner contradiction, and redundant, which identify turns that express deteriorations in conversational consistency. The interface used for this annotation task is shown in Figure 3.8.

Speaker Consistency in Dialogue

Overview

This task explores how consistent Alex's responses are within a dialogue. For each of Alex's responses, you must decide all of the consistency labels that are applicable, if any.

It is often the case that MORE THAN ONE label is applicable to a single turn. It is important to select the appropriate label for EVERY occurrence, not just the first time it happens.

The 3 consistency labels you are considering are:

- 1. Contradicts Self Context**

The speaker says something that doesn't make sense, either because it is a contradiction of something they have said previously or it is extremely implausible based on the information they have already shared.

- 2. Contradicts Partner Context**

The turn is inconsistent with what Sam has said in the conversation history.

For this label, judge whether the turn demonstrates that the speaker has forgotten or misremembered what their partner has said earlier in the conversation.

- 3. Redundant**

The speaker repeats something either of the speakers has said earlier in the conversation in a way where most people would find it clearly inappropriate, annoying, or unnatural.

Note that many cases of repetition are appropriate, such as when a speaker...

- reiterates an idea as an acknowledgement
- elaborates on an earlier point
- repeats something for additional emphasis
- summarizes a discussion or thought

Help

Details and examples for each label are provided in their respective help sections below.

Contradicts Self Context +

Contradicts Partner Context +

(Sam) hi there

(Alex) Hello, how are you? I love visiting India, it is my favorite place to visit.

Select ALL labels that apply:

Select...

Contradicts self context

Contradicts partner context

Redundant

(Alex) I went to Bangalore, which is the capital of India. It was very crowded.

Select ALL labels that apply:

Select...

(Sam) What makes it stand out as your favorite place to visit then?

(Alex) I think it was because it was so new to me. I had never been to a big city before. I have always lived in New York City.

Select ALL labels that apply:

Select...

(Sam) I see. Well, maybe you will return one day.

(Alex) I hope so! I would love to go back and explore more of the city and the surrounding area. Have you ever been to India?

Select ALL labels that apply:

Select...

You are on task 1 / 1 for dialogue 1 / 3

Next

Figure 3.8: Interface for self contradiction, partner contradiction, and redundant

8. Turn-taking covers four labels: **ignore**, **irrelevant**, **follow-up**, and **topic switch**, which identify positive and negative cases of appropriate turn-taking behaviors. The interface used for this annotation task is shown in Figure 3.9.

Transitions during Dialogue

Instructions

Categorize responses based on their appropriateness to the context they appear in. There are 3 facets to consider for this task

1. Acknowledgements

A response is an appropriate acknowledgement when it is clear that the speaker is responding directly to what their partner just said. This could be from an explicit acknowledgement (e.g. saying "okay" at the beginning of the response), or from a response where the speaker implicitly shows an understanding of what their partner said.

Responses that are not appropriate acknowledgements make it seem like the speaker ignored their partner.

Note that there are some contexts that do not require an acknowledgement in the response, usually because the previous turn did not present any new ideas or questions.

2. Topic Switches

A topic switch occurs when the response causes the conversation to focus on a new topic or idea that is not a subpoint of the current discussion topic.

If a new talking point is introduced but it fits under the current overall topic of the conversation, then this is not a topic switch.

If the response only presents ideas, questions, or information that are directly following up on specific questions or ideas presented in the previous turn, then this is also not a topic switch.

3. Relevance

A response is appropriately relevant to the dialogue when it continues the current discussion or naturally transitions to a new topic of conversation.

Responses that are not appropriately relevant feel abrupt and interrupt the discussion, usually because they present questions or ideas that are unrelated to the previous turn.

Help

If you are unsure of the correct answer to a question, a help section is provided below for each type of question. Clicking on the help icon next to each question will open and jump to the relevant help section.

Does Alex appropriately acknowledge Sam with this response?

(Sam) hi

(Alex) hi, how are you? i just got back from a dance class. what do you do?

(Sam) I'm okay. I play video games but its cool that you are a dancer.

(Alex) yeah, i love video games. what kind of video games do you play? do you have a favorite?

You are on task 1 / 1 for dialogue 1 / 3

[Next](#)

Figure 3.9: Interface for **ignore**, **irrelevant**, **follow-up**, and **topic switch**

Training and Screening To improve annotation consistency and detect poorly performing evaluators, automated training sessions are included for each annotation task. Each session consists of three conversations that evaluators annotate using an identical procedure and web interface to the corresponding task. The three conversations used for each session are hand-crafted to represent a variety of positive and negative examples of the behavior labels for the corresponding task. The gold annotations for each training conversation are hidden from evaluators during the annotation; however, after completing each training conversation,

any disagreements between the evaluator’s annotations and gold labels are displayed along with an explanation to help the evaluator improve. The evaluator’s performance on the third conversation of each training session is used to screen evaluators, where performance is measured by the number of turns where their annotations disagree with gold labels. Evaluators are eligible to complete the work on a task if they make mistakes on fewer than two turns for the Sociality and Interpretability tasks, or on fewer than three turns for the other six tasks.

3.3 Existing Evaluation Methods

To evaluate ABC-Eval as a fine-grained evaluation method, three popular approaches to evaluations of dialogue characteristics are included for comparison in this study: Dialogue Likert, Turn Likert, and Comparative. Through this comparison, the benefits afforded by ABC-Eval for dimensional evaluation of dialogue characteristics will be assessed.

Label	Dialogue Likert	Turn Likert	Comparative
Consistency	Con _d	Con _t	Con _c
Emotion	Emo _d	Emo _t	Emo _c
Understanding	Eng _d	Eng _t	Eng _c
Engagingness	Eng _d	Eng _t	Eng _c
Grammaticality	Gra _d	Gra _t	Gra _c
Informativeness	Inf _d	Inf _t	Inf _c
Proactivity	Pro _d	Pro _t	Pro _c
Relevance	Rel _d	Rel _t	Rel _c
Quality	Qua _d	Qua _t	Qua _c

Table 3.4: The seven labels for dialogue characteristics for Likert and Comparative evaluations (taken from Finch and Choi [14]), henceforth referred to using their abbreviations and colors. The label for holistic quality is also shown below the dashed line.

Dialogue Likert Annotators provide dialogue-level ratings from 1 (least) to 5 (most) for the seven characteristic labels shown in Table 3.4. The dimension set proposed in Finch

and Choi [14] is used, which results from a detailed survey of characteristics used in chat evaluation. Bot-level metrics are calculated as the mean rating across all bot dialogues.

Turn Likert Annotators provide turn-level ratings on the same scale and labels as those used for Dialogue Likert. The dialogue-level metric is measured as the mean rating of a single dialogue’s turns. The bot-level metric is calculated as the mean rating of all turns in all bot dialogues.

Comparative Annotators select the dialogue in which chatbot responses better fit a label definition from a side-by-side pair of dialogues, also using the seven characteristic labels in Table 3.4. A “neither” option is allowed, only for cases where the evaluator cannot distinguish which dialogue is a better fit. Bot-level metrics are calculated as bot pair win/tie/loss proportions between pairing of their dialogues.

For each evaluation method described previously, overall quality annotations are also included.

3.4 Evaluation Collection

To compare ABC-Eval against the existing evaluation methods, it is necessary to collect evaluations from each of these methods on the same underlying conversation dataset. To this end, human-bot conversations are collected using high-performing models for the time. Based on a literature review of state-of-the-art open-domain multi-turn dialogue models and subsequent pilot studies within my lab, the following four dialogue models are selected:

- Blender2 [95], a high-performing large-scale general-purpose neural dialogue model trained on several dialogue datasets.
- Emora [16], a rule-based chatbot from the Amazon Alexa Prize Socialbot Grand Challenge [67] with the all-time highest final score.

- Blender-Decode [60], a Blender-based dialogue model that ranks generated response candidates by their consistency to the context.
- BART-FiD-RAG [79], a knowledge-grounded open-domain dialogue model using a fusion-in-decoder and retrieval-augmented-generation techniques.

Conversations are collected between 46 university undergraduate students and the dialogue models in a remote text-based conversation setting using a Javascript-based interactive web interface that I modified from the ParlAI framework. Links to the web interface are sent to each interactor with instructions to be completed within 2 weeks. For each link, the interactor completes two conversations with a random pair of chatbots, for a minimum of 30 turns per conversation, with a similar open-ended, topic-free chatting environment to Adiwardana et al. [1]. Interactors are asked to rate seven dimensions as well as overall quality (Table 3.4) of each conversation after its completion on a 1-5 Likert scale, and to select the better-fitting conversation for the same seven dimensions and overall quality after each conversation-pair (ties allowed). The final conversation dataset includes 400 human-bot dialogues (100 dialogues per chatbot), averaging 30.3 turns per dialogue (11.3 tokens per user turn).

After the conversations are collected, external evaluations are then conducted. ABC-Eval, Dialogue Likert, Turn Likert, and Comparative evaluations are performed on these 400 conversations using an external annotation company SurgeHQ. Consequently, this study consists of the collection of 40 labels per conversation. This collection is split into 18 independent evaluation tasks as follows:

- 8 ABC-Eval tasks, each composed of 1 to 4 labels as denoted by groupings in Table 3.3
- 1 Dialogue Likert task, composed of all 8 labels from Table 3.4 completed in random order
- 8 Turn Likert tasks, each composed of 1 label from Table 3.4
- 1 Comparative task, composed of all 8 labels from Table 3.4 completed in random order

The 18 evaluation tasks are posted on SurgeHQ’s annotation platform (<https://www.surgehq.ai>) to be completed by dedicated remote workers (Surgers) with experience in NLP annotation. Each time an evaluator connects to one of the tasks, they are assigned a randomly selected conversation to annotate. This study is allocated a group of 125 Surgers, chosen by a SurgeHQ employee based on high annotation performance on past projects. Evaluators are compensated per annotated conversation per task, at an estimated rate of \$20/hr. Evaluators are allowed to annotate up to 60 conversations per task.

The final evaluation dataset consists of 400 conversations, each with results for all 40 labels. Additionally, a randomly-selected subset of 100 conversations (and 50 of the conversation pairs) is evaluated a second time by a different Surger in order to measure inter-annotator agreement (IAA).

3.5 Evaluation Analysis

In this section, I present the results of four analyses comparing ABC-Eval against the existing evaluation methods for fine-grained dialogue characteristic evaluations. These analyses demonstrate that ABC-Eval achieves better interpretability (Section 3.5.1), better coverage of distinct characteristics of chat that impact quality (Section 3.5.2 and Section 3.5.4), and overall higher measurement sensitivity (Section 3.5.3) than alternative evaluation methods.

3.5.1 Interpretability

The reliability of interpreting each metric’s annotation instructions is measured by calculating IAA using the set of 100 double-annotated conversations. High agreement between annotators demonstrates that different people can reliably come to the same conclusions about how a metric’s definition applies to each chatbot response.

The results in Figure 3.10 show the IAA for all metrics from each evaluation method (left to right: ABC-Eval (orange), Turn Likert (blue), Dialogue Likert (red), and Dialogue

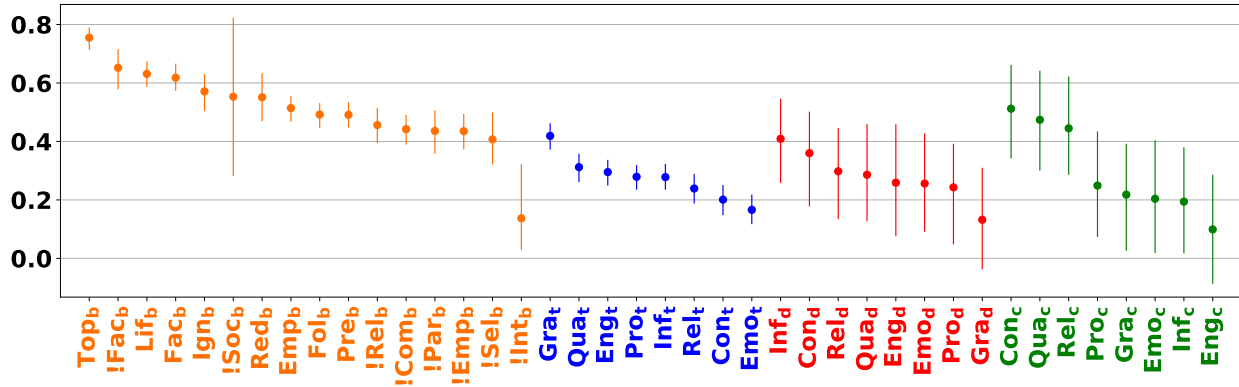


Figure 3.10: IAA (Krippendorff’s alpha) for all metrics. Error bars denote 95% bootstrap confidence intervals, using bias-corrected and accelerated confidence intervals with $k = 10,000$ Monte Carlo case resamples. $!Soc_b$ and $!Int_b$ ’s confidence intervals are largely due to a low rate of positive examples (see Figure 3.13).

Comparative (green)). Overall, it can be seen that the definitions of most ABC-Eval metrics can be interpreted more reliably than the definitions of most Turn Likert, Dialogue Likert, and Dialogue Comparative metrics. Likert-style and comparative-style annotations appear to have similar interpretability, although Qua_c is a notable exception that produced higher agreement than Qua_d .

3.5.2 Importance

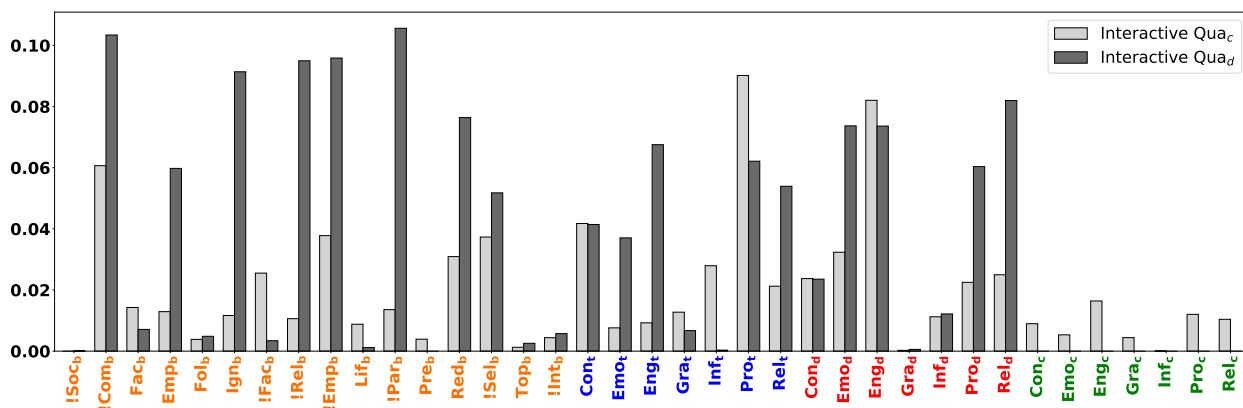


Figure 3.11: Extent to which each evaluation metric can explain variance in conversation quality by fitting a univariate regression model (R^2 for predicting Qua_d using linear regression, McFadden’s Pseudo- R^2 for predicting Qua_c using logistic regression). Comparative metrics cannot predict Qua_d so only results for variance of Qua_c are shown.

The importance of each metric is estimated by a predictive validity analysis that measures the extent, to which the metric can predict conversation quality (Figure 3.11). Conversation quality is represented by Qua_d and Qua_c from interactors that participated in the conversations (Section 3.4) to avoid cases where the same evaluator produced the quality label and explanatory metric. The predictive validity of each metric is measured by fitting univariate linear or logistic regression models to predict Qua_d or Qua_c , respectively.

Qua_c is represented as a binary encoding, where 0 and 1 represent choosing the first and second conversation, respectively. Any conversation pairs in which the interactor could not distinguish a difference in quality between conversations are excluded, and models are fitted on the remaining set of 184 conversations. To use non-comparative predictors for predicting Qua_c , the difference in metric value between each pair of conversations is used.

These results show that dialogue quality is substantially related to emotional understanding metrics (Emo , Emp_b , $!Emp_b$), relevance-related metrics (Rel , $!Rel_b$, Ign_b), and consistency metrics (Con , $!Sel_b$, Red_b , $!Par_b$). Within these metric groupings, ABC-Eval metrics are overall more predictive of quality than their Likert or Comparative analogs, while Comparative metrics are least predictive of quality.

3.5.3 Sensitivity

The sensitivity of each metric is investigated following the procedure used by Li et al. [42] where hypothesis tests are conducted to count the number of statistically significant differences that each metric is able to detect between the six pairings of the four chatbots (Table 3.5). To make results comparable, the conversations used for hypothesis testing are downsampled to 32 conversations per bot for the Dialogue Likert, Turn Likert, and ABC-Eval metrics to match the 32 conversation-pairs per bot-pair produced by the Comparative evaluation.¹

The results show that the Likert evaluations are more sensitive than the Comparative evaluation for most labels. ABC-Eval metrics have a wide range of sensitivity, with many

¹Only 192 of the 200 dialogue pairs are evaluated with Comparative labels due to a collection mistake

α	!So c _b	!Com b	Fac b	Emp b	Fol b	Ign b	!Fac b	!Rel b	!Emp b	Lif b	!Par b	Pre b	Red b	!Sel b	Top b	!Inf b	Con t	Emo t	Eng t	Gra t	Inf t	Pro t	Qua t	Rel t	Con d	Emo d	Eng d	Gra d	Inf d	Pro d	Qua d	Rel d	Con c	Emo c	Eng c	Gra c	Inf c	Pro c	Qua c	Rel c
0.01	0	1	4	3	5	1	4	2	4	2	0	4	2	5	5	2	1	1	2	4	4	4	4	2	1	2	1	0	0	3	1	0	1	0	1	0	1	1	0	0
0.05	0	2	5	3	5	2	6	2	5	2	1	4	3	5	5	3	2	1	3	5	4	4	4	3	2	3	2	3	2	3	1	2	1	0	1	0	1	2	1	0
0.1	1	3	5	3	5	2	6	2	5	3	1	4	3	5	5	3	2	3	3	5	4	4	4	4	2	4	2	5	3	3	3	2	2	0	1	0	2	2	1	2

Table 3.5: The number of statistically significant differences detected by each metric when comparing bot-pairs using z-tests of proportions (ABC-Eval), t-tests (Turn Likert and Dialogue Likert), and sign tests (Comparative) at three significance thresholds.

ABC-Eval metrics appearing to be more sensitive analogs of similar Likert metrics. For example, the results show that *!Sel_b* and *Red_b* are more sensitive than *Con*, that *Fac_b* and *!Fac_b* are more sensitive than *Inf*, and that *Emp_b* and *!Emp_b* are more sensitive than *Emo*. On the other hand, the Likert-style *Rel* metric shows similar or slightly superior sensitivity compared to the analogous *Ign* and *!Rel* behavior metrics.

3.5.4 Coverage & Distinctness

The coverage and distinctness of the ABC-Eval metrics is investigated via incremental validity analysis. For this analysis, backwards stepwise regression is performed that determines (1) the ability of an evaluation method as a whole to explain conversation quality, and (2) whether each metric contributes distinct information about quality above and beyond other metrics (Figure 3.12). Specifically, a multivariate regression model is fit to each of the four evaluation methods under study. These models are fit similarly to those presented in Section 3.5.2, but include all non-quality metrics within an evaluation method as predictors. Then, predictors are removed from each model one at a time based on a beam search ($k=100$) of which removed predictor results in the smallest decrease in model fitness (adjusted R^2 or adjusted pseudo- R^2). This stepwise regression analysis is performed twice to predict both Qua_d and Qua_c given by interactors, similar to the analysis in Section 3.5.2.

The results show that ABC-Eval has overall better coverage than other evaluation methods for explaining conversation quality. Furthermore, most ABC-Eval metrics that have a strong relationship with conversation quality appear to be appropriately distinct in the information they provide, especially *!Emp_b*, *!Sel_b*, *Red_b*, *!Rel_b*, *Emp_b*, *!Com_b*, and *Ign_b*.

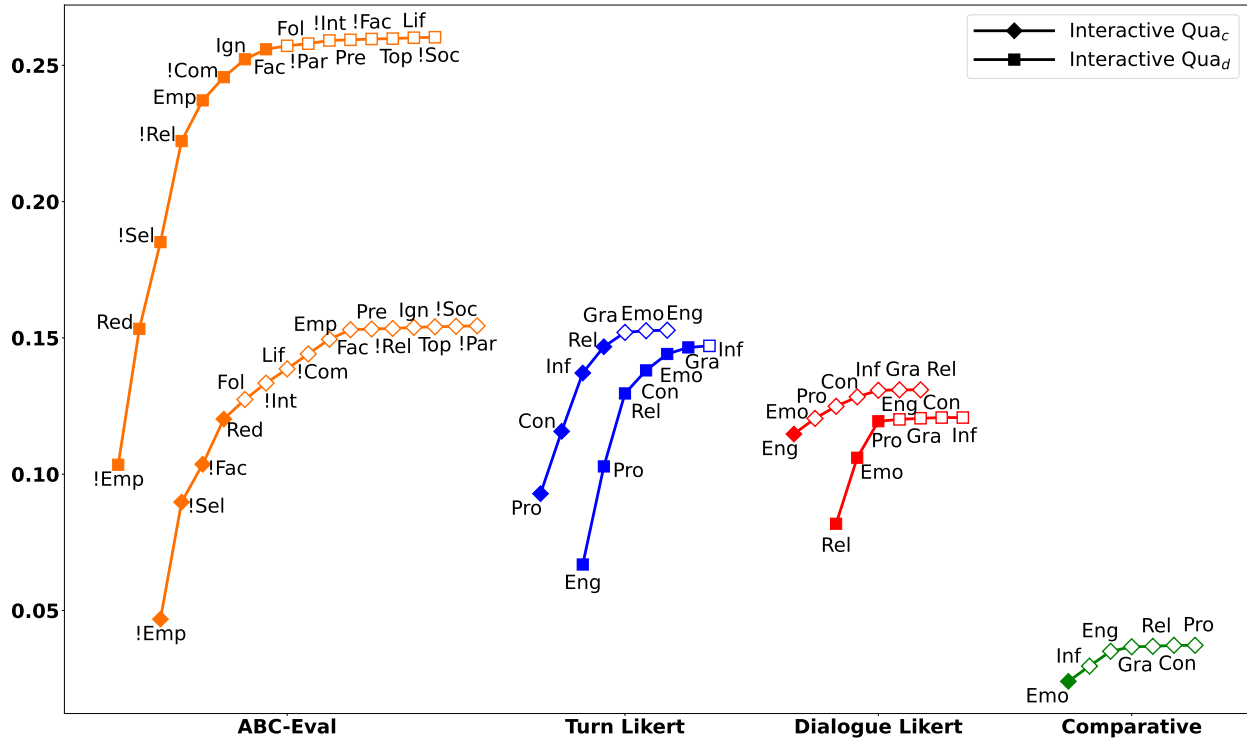


Figure 3.12: Incremental validity of metrics within 4 evaluation methods, obtained using backwards stepwise regression. Points represent the extent to which a model can explain variance in quality (R^2 for predicting Qua_d with a linear model, McFadden’s pseudo- R^2 for predicting Qua_c with a logistic model) using all metrics on the same line and to the left as predictors. Filled marker symbols denote steps where the model’s predictors all contributed positively to adjusted R^2 or adjusted pseudo- R^2 values; otherwise, marker symbol is unfilled. Comparative metrics cannot be used to predict Qua_d so only results for explaining variance of Qua_c are shown.

Similar distinctness can also be seen in Turn Likert metrics, whereas dialogue-level metrics show relatively low distinctness.

3.6 Cost

Tables 3.6 and 3.7 provide the payment rates per task used in this study and the subsequent overall costs of each evaluation method, respectively. It is clearly shown that turn-level annotation tasks are significantly more costly to gather compared to dialogue-level tasks. Nevertheless, the previous analyses suggest that such turn-level annotation tasks, particularly ABC-Eval, offer richer insights for finely assessing dialogue models. However, it’s crucial to

recognize that not all 16 ABC-Eval behavior metrics are equally relevant across all dialogue applications, rather subsets of behaviors can be identified that capture the desired dialogue characteristics of a particular application. Hence, the most pragmatic evaluation approach, ensuring both utility and efficiency in model performance assessment, likely involves a blend of ABC-Eval and other metrics. For example, a dialogue model focusing on commonsense knowledge targets commonsense understanding, response relevance, response informativeness, and overall quality. This can be achieved by combining ABC-Eval metrics such as `commonsense contradictions` and `irrelevant` with Likert metrics like `informativeness` and `quality`. Such an evaluation scheme provides detailed insights into desired characteristics while remaining practical in terms of cost and annotator effort. In fact, this evaluation approach is found useful in practice later in this dissertation (Chapter 5).

Task	Payment	Task	Payment
Uninterpretable	\$0.63	Antisocial	\$0.44
Preference Info Life Info	\$0.70	Empathetic Lack of Empathy	\$1.15
Commonsense Contradiction	\$0.92	Fact Usage Fact Contradiction	\$1.96
Self Contradiction Partner Contradiction Redundant	\$0.87	Ignore Irrelevant Follow-up Topic Switch	\$1.87
Dialogue Likert	\$0.60	Turn Likert	\$0.70
Comparative	\$1.43		

Table 3.6: Payment per annotation task in USD. The payment for Turn Likert is per label whereas the indicated payment for Dialogue Likert and Comparative covers all labels, due to how the annotation tasks were constructed (Section 3.2.3).

Metric	TI	TP	EC	OC
Dialogue Likert	2.81	21.37	374.36	240.00
Comparative	4.35	13.81	289.68	286.67
Turn Likert	19.94	3.01	2658.40	2240.00
ABC-Eval	25.60	2.34	3413.58	3422.67

Table 3.7: The data collection costs for each task in USD. **TI**me is the median completion time in minutes for one dialogue. **Throug**h**P**ut represents the number of completed dialogues per hour. **E**stimated **C**ost is calculated using median completion time, 400 dialogues, and \$20/hr rate. **O**ur **C**ost is the total amount paid in this work to collect a dataset of 400 conversations (single-annotated).

3.7 Dialogue Model Insights

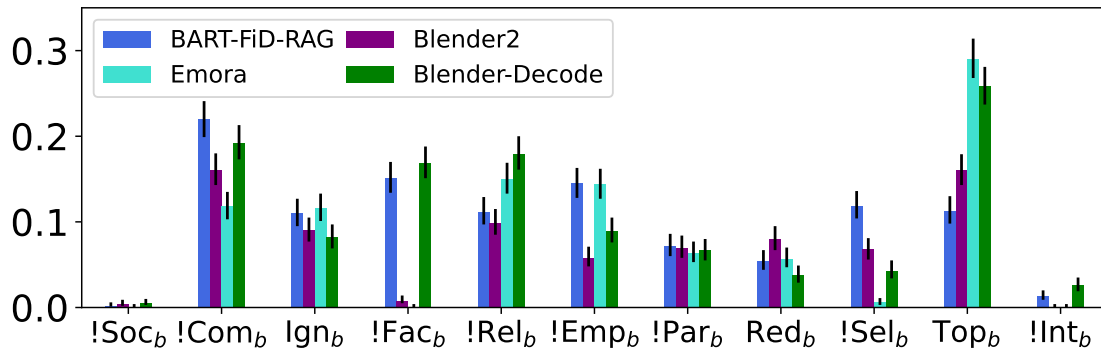


Figure 3.13: Proportions of turns expressing undesirable behaviors, with 95% Wilson score confidence intervals.

Given that ABC-Eval is a powerful fine-grained evaluation framework for dialogue models, I next consider the insights it provides towards the weaknesses of current neural dialogue models, specifically focusing on commonsense understanding. Figure 3.13 demonstrates the rates of undesirable model behaviors observed in the 400 collected conversations. The results show that a clear weakness in neural dialogue models is in commonsense reasoning, since commonsense violations are present in about 15-20% of the bots’ responses (!Com_b). In addition, consistency issues are prevalent across all bots: self-contradictions and partner contradictions appear in about 10% of the bots’ responses overall (!Sel_b, !Par_b). The behaviors of self- and partner-contradictions often reveal commonsense errors as well, since the occurrence of two mutually exclusive events as what is observed during such contradictions is a violation of commonsense. The importance of commonsense in dialogue is further corroborated by the predictive validity results shown in Figure 3.11, in which it is clear to see that commonsense violations and partner contradictions are two of the strongest predictors of overall dialogue quality.

3.8 Conclusion

This study demonstrates the power of a binary behavior-based evaluation framework for assessing dialogue models in high granularity through the development and application of

ABC-Eval. The metrics covered in ABC-Eval have better interpretability (Section 3.5.1), a wider coverage of distinct characteristics of chat that impact quality (Section 3.5.2 and Section 3.5.4), and overall higher measurement sensitivity (Section 3.5.3) than alternative evaluation methods of dialogue model characteristics. In light of this, ABC-Eval is a compelling method for obtaining a detailed view of specific strengths and weaknesses of dialogue models through the measurement of different behavior rates. By presenting verified evaluation setups for 16 different behavior labels, ABC-Eval enables the selection of the most pertinent subset for a particular dialogue application. This can serve as a complement to the coarse-grained yet more efficient evaluations of overall dialogue model quality, often collected through Likert ratings, in order to better understand and explain the specific characteristics of different dialogue models in detail. Following this, application of ABC-Eval to neural dialogue models circa 2020 and 2021 reveal that commonsense reasoning is a crucial vulnerability, thereby motivating the further efforts undertaken in this dissertation.

Chapter 4

ConvoSense: Generating Commonsense for Dialogue

4.1 Introduction

Given the weakness to commonsense observed in earlier neural dialogue models (Chapter 3), it becomes pertinent to develop better dialogue commonsense resources that can be leveraged in downstream dialogue applications. Towards this, this chapter targets the second research question: Is it possible to reliably obtain multi-faceted and contextually novel social commonsense inferences for dialogue? In this chapter, I present the collection of ConvoSense, a dataset of social commonsense inferences for dialogue [15]. As discussed in Section 2.2, existing commonsense resources for dialogue tend to be misaligned with dialogue contexts, recycle known context information as commonsense inferences, derive only straightforward conclusions with minimal elaboration, or contain inferences of low diversity. As such, it is necessary to develop a more suitable social commonsense resource that can be used to produce multiple novel inferences for a given dialogue context. To this end, ConvoSense is the first dialogue-centric social 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. Compared to existing datasets, ConvoSense shows greater contextual novelty, a higher number of inferences per example, and enhanced inference diversity while maintaining exceptional reasonableness (Section 4.5.2). I will also discuss in this chapter the results of several experiments towards developing generative models for producing inferences for dialogue contexts (Section 4.4), which show that models trained on ConvoSense excel in generating plausible inferences with higher detail and novelty, compared to ones trained on existing datasets (Section 4.5). Importantly, the trained commonsense generator (ConvoSenseM*) from this work will be leveraged as a powerful social commonsense resource for the commonsense-augmented dialogue approach covered next in Chapter 5.

4.2 ChatGPT Dialogue Commonsense Generation

In order to support the development of a large-scale and high coverage commonsense dataset for dialogue that improves upon existing works, I hypothesize that I can leverage large language models (LLMs) to accomplish this task in an efficient and low-cost manner. From initial pilot tests of both closed-source (ChatGPT) and open-sourced LLMs (Vicuna and Llama), I find that ChatGPT provides greater reliability in following specific instructions and produces commonsense inferences of overall better quality than the open-sourced LLMs. Consequently, I choose to rely on ChatGPT in this work.

4.2.1 Prompt Engineering

Prior to crafting the full ConvoSense dataset, I empirically assess ChatGPT’s efficacy in generating reasonable and novel commonsense inferences for dialogue. To mitigate any unintended bias from in-context examples in the ChatGPT prompt, I adopt a zero-shot generation framework using ChatGPT.¹ During the development process of the ChatGPT prompts for this task, I observe that the inferences generated from ChatGPT frequently

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

contain detailed and rich information, thus addressing one of the major limitations of existing works. In addition, to encourage novel inferences from ChatGPT, I include the instruction “Your answers should provide novel information that is not explicitly shared in the conversation.” as seen in Table 4.1. I 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 accompanied by 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**) slots in the prompt. For every dialogue context, the 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. Finally, the ChatGPT output, commencing with the header *Answers* and adopting a list-like format with newline separation, is parsed to extract the generated inferences. An example of the final prompt design is illustrated in Table 4.1 and the questions and answer prefixes employed for the fifteen identified social commonsense inference types derived from the existing dialogue-centric datasets (Table 2.1) are detailed in Table 4.2.

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 4.1: A ChatGPT 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).

Type	Question	Answer Template
<u>Subsequent</u>	What might happen after what Speaker just said?	After this, ...
<u>Antecedent</u>	What events happened before the situation that Speaker just shared?	Before this, ...
<u>Cause</u>	What could have caused the last thing said to happen?	This was caused by...
<u>Prerequisite</u>	What prerequisites are required for the last thing said to occur?	For this to happen, it must be true that...
<u>Motivation</u>	What is an emotion or human drive that motivates Speaker based on what they just said?	Speaker is motivated...
<u>Attribute</u>	What is a likely characteristic of Speaker based on what they just said?	Speaker is...
<u>Reaction</u>	How is Speaker feeling after what they just said?	Speaker feels...
<u>Reaction_o</u>	How does Listener feel because of what Speaker just said?	Listener feels...
<u>Desire</u>	What does Speaker want to do next?	As a result, Speaker wants...
<u>Desire_o</u>	What will Listener want to do next based on what Speaker just said?	As a result, Listener wants...
<u>Constituents</u>	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 4.2: Question and answer prefixes used for generating each inference type from GPT for dialogue contexts. The ten inference types included in ConvoSense are represented by Type.

4.2.2 Evaluating ChatGPT-generated Commonsense

To evaluate the quality of ChatGPT-generated commonsense inferences for dialogues, I compare their **reasonableness** and **novelty** against inferences from human datasets. First, I sample dialogue contexts and their corresponding inferences from the existing human datasets of ComFact, Cicero, and Reflect (Section 2.2). For each dataset, an equal number of examples are sampled for each inference type provided in the dataset (Table 2.1). For every sampled example, I then prompt ChatGPT to produce inferences of the same type as that provided by the human and randomly select one from the generated list. Finally, two human annotators from the SurgeHQ crowdsourcing platform are presented with the dialogue context, inference question, and both the ChatGPT- and human-generated inferences and asked to categorize them.

Reasonableness Most prior works assess their inferences based on human-judged reasonableness [32, 21, 78, 111]. An inference is deemed reasonable if it makes sense in, is relevant to, and is consistent with the provided dialogue context. I follow Hwang et al. [32], 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 new contextual information. In my study, annotators explicitly evaluate the extent to which an inference contributes fresh information to the conversation, categorized as: *new & detailed*, *new & simple*, and *purely repetitive*.

Since I aim to elicit the natural commonsense understanding learned by each annotator through their life experience in the annotation tasks, I 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, I provide a description of the task with definitions of the different categories. The 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. I verified through pilots that this approach resulted in reliable and sensible annotations.

4.2.3 Results

Following the evaluation approach presented in Hwang et al. [32], the two metrics in this work are converted into binary representations. Thus, labels [*always/likely*, *sometimes/possible*] are categorized as *positive* and [*never/farfetched*, *invalid/nonsense*] are considered *negative* reasonableness. Similarly, [*new & detailed*, *new & simple*] are designated as *positive*, and [*purely repetitive*] is classified as *negative* novelty. In cases of annotator disagreement, one of

the annotators’ decisions is randomly selected. To mitigate the potential noise introduced by this random selection, I repeat the process 100 times and report the average result, only confirming statistical significance when every selection yields a significant result.

Many agreement metrics, including Cohen’s kappa, have been shown to be inaccurate measurements of annotator agreement in cases where there is a large class imbalance due to their calculation of chance agreement [34, 96, 65]. Gwet’s AC1 inter-annotator agreement metric has been demonstrated to overcome the vulnerability to class imbalance due to its different calculation for chance agreement, thus producing more accurate inter-annotator agreement measurements for such situations [27]. Considering the reported quality of the existing datasets and my preliminary assessments of ChatGPT-generated inferences, I expect much higher rates of positive classes than negative ones, resulting in a class imbalance. As a result, I utilize Gwet’s AC1 metric for inter-annotator agreement in this work. The annotators obtain AC1 values of 0.8 and 0.6 for reasonableness and novelty, respectively, implying substantial agreement.

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

Table 4.3: The average % ($\sigma < 2\%$) of total samples (**#**) tested as reasonable (**R**) and novel (**N**), with discordance probabilities in parentheses. *: statistical significance (McNemar’s test, $\alpha = 0.05$). 90 more samples are used for ComFact due to its greater number of inference types.

Table 4.3 demonstrates that ChatGPT attains comparable reasonableness in its generated inferences as those derived from humans, even exceeding the reasonableness of the inferences in ComFact with statistical significance. Notably, the results also indicate that ChatGPT *surpasses* the novelty of the human-generated inferences for the majority of the existing datasets. Furthermore, ChatGPT outputs achieve higher detail than that observed from

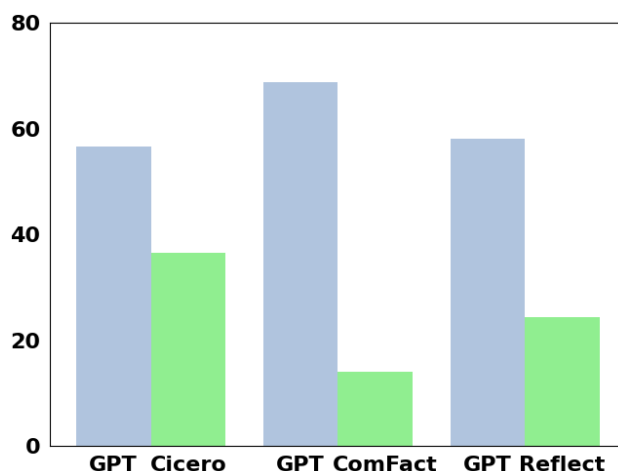


Figure 4.1: Average % of *new* & *detailed* inferences out of all positive novelty inferences for each data source.

human-generated inferences. Figure 4.1 shows the percentage of *new* & *detailed* inferences out of all positive novelty inferences for each data source, clearly demonstrating the superiority of ChatGPT inferences in terms of their expressed detail. Example inferences from ChatGPT and humans are shown in Figure 4.2.

Cause	Attribute
<p>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.</p>	<p>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.</p>
<p>1. <i>the listener does not know which subject to study.</i></p>	<p>1. <i>the speaker is old fashioned.</i> 2. <i>the speaker is outdoorsy.</i></p>
<p>1. <i>the speaker's personal interest in economics.</i> 2. <i>the speaker's observation of the job market and the high demand for economics graduates.</i> 3. <i>the speaker's perception of economics as a versatile and valuable degree that can open up various career paths.</i> 4. <i>the speaker's inclination towards quantitative and analytical subjects, which are central to economics.</i> 5. <i>the speaker's belief that economics provides a solid foundation for understanding and navigating today's complex global economy.</i></p>	<p>1. <i>someone who enjoys spending time in nature.</i> 2. <i>familiar with the park by the lake and has visited it before.</i> 3. <i>someone who values and prioritizes socializing and spending time with loved ones.</i> 4. <i>someone who enjoys simple pleasures like picnicking and appreciating the outdoors.</i> 5. <i>someone who makes an effort to plan enjoyable activities to share with others.</i></p>

Figure 4.2: Cause and Attribute inferences written by humans (top, green) and generated by ChatGPT (bottom, blue).

4.3 Constructing Commonsense Datasets

Given my assessment of high-quality, novel, ChatGPT-generated commonsense inferences, I next construct a substantial conversational commonsense dataset, termed ConvoSense.

4.3.1 ConvoSense: New ChatGPT-generated Dataset

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

Dialogue Selection I 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 ChatGPT covering situations based on ATOMIC commonsense tuples [37]. For cost practicality, ConvoSense is constructed to contain 10,000 training dialogues, 1,000 validation dialogues, and 1,000 test dialogues each.

To encourage diversity in ConvoSense, I employ BERTopic [25], which clusters the dialogues selected from SODA into groups using dimension reduction technique UMAP [55] and HDBSCAN clustering algorithm [54] on embeddings of the dialogues.² I configure the hyperparameters³ to effectively group dialogues while maintaining a well-balanced distribution of group lengths based on manual verifications. As a result, I 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, I randomly select one dialogue from the n groupings, where each dialogue contains at

²The `all-mpnet-base-v2` model is used for embedding.

³neighbors: 5, components: 5, min_cluster_size: 2.

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, I set n values as $[10K, 1K, 1K]$ for assembling the training, validation, and test splits, respectively.

Utterance Selection For each selected dialogue, I determine which utterance to perform inference generation on. The goal is to select utterances that are semantically rich, since commonsense reasoning is most applicable in situations in which there is much shared information to ingest about the conversational partner on which to make inferences [109].⁴ Manual review of the topic keywords identified for each group during the BERTopic grouping suggest that the topic keywords can serve as useful approximations of the rich semantic information expressed in the dialogue. Similar to Zhou et al. [109] who identify the most similar dialogue utterance to the grounding scenario description of the dialogue for this selection step, I use the topic keywords to pinpoint the most topically salient utterance in each dialogue which is then used as the focal utterance for commonsense inference generation. In addition, utilizing the topic keywords to select the utterance also helps to ensure that the topical diversity afforded by the grouping is maintained. Consequently, in this work, the focal utterance is determined 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, I 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, I introduce nominal tags for the two participants. The terminal utterance is labeled as *Speaker*, and its preceding utterance is labeled as *Listener*. These

⁴Utterances which focus less on information-sharing (e.g. acknowledgment, question-asking, etc.) may have different requirements or integrations with commonsense reasoning, which is left to future work.

nominal tags are then assigned in alternating order to the remaining utterances.

Inference Types For each preprocessed dialogue, ChatGPT generates inferences for all included commonsense types following the procedure in Section 4.2.1. Specifically, ten commonsense types are included: **Subsequent**, **Cause**, **Prerequisite**, **Motivation**, **Attribute**, **Reaction**, **Reaction_o**, **Desire**, **Desire_o**, and **Constituents** (highlighted in Table 4.2). These ten types are selected from the overall set of fifteen social commonsense types identified from previous works based on their usage frequency in existing datasets and their lack of semantic overlap. The rationale behind the decisions of what to exclude are as follows:

- The set **Effect**, **Effect_s**, and **Effect_o** is excluded because such information is highly related to **Subsequent** as well as **Desire_s** and **Desire_o** in many cases.
- **Antecedent** is excluded due to high overlap with **Cause** and **Prerequisite**.
- **Obstacle** is excluded due to its low usage in existing datasets and a lack of compelling examples of its applicability to dialogue situations.

Illustrative examples of ConvoSense are shown in Figure 4.3.

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 4.3: Desire and Desire_o inferences in the ConvoSense dataset.

Data Quality The results in Section 4.2.3 demonstrate that ChatGPT is generally capable of producing high-quality commonsense inferences regardless of the underlying dialogue source. Consequently, applying ChatGPT to generate commonsense inferences for the SODA dialogues is expected to perform with similar high quality. To explicitly verify this, I 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 reasonableness and novelty of an inference (Sec. 4.2.2) and performing inference clustering to measure per-example output diversity. Table 4.4 presents the results, confirming the high reasonableness, novelty, detailedness, and diversity of the inferences in the ConvoSense dataset.

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

Table 4.4: 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.

Error Analysis I next perform an error analysis on the unreasonable inferences identified by the human judge. I 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 4.3), or by their attribution to the wrong conversational participant (26%; Desire_o examples #4-5 in Figure 4.3). 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 predominately

accurate representations of commonsense understanding, although they can suffer from lack of precision regarding situational nuances and speaker roles.

4.3.2 HumanGen: Human-generated Datasets

To compare ConvoSense against the existing human-generated datasets, I combine the three human-generated datasets of ComFact, Cicero, and Reflect (Section 2.2) into a solitary dataset, termed HumanGen.⁵ 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 cohesive format, I perform the following preprocessing steps.

First, I align the commonsense types in the human datasets with the specifications of Table 4.2 to identify the appropriate commonsense inference questions for each instance. Then, I combine consecutive utterances from the same speaker to ensure every dialogue turn represents a distinct speaker. Lastly, I apply *Speaker* and *Listener* tags in a similar manner to ConvoSense (Figure 4.3). Since human-generated inferences often contain nominal references to specific target entities, I 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, I 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, I randomly assign A/B tags to each conversation. On the other hand, Reflect includes original speaker names; thus, I replace them with A/B tags accordingly. Since the speaker name frequently appears in Reflect’s inferences, I uniformly replace it with “*the speaker*”, aligning with the prevalent format in Cicero.

⁵Many 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.

4.3.3 Dataset Statistics

Table 4.5 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.

	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 4.5: 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.

4.4 Generative Commonsense Models

4.4.1 Training and Decoding Strategies

With the rich and diverse multi-inference examples provided in ConvoSense, training commonsense generation models that produce versatile outputs is likely well-supported. Yet, a key query remains: how can this versatility be best achieved by the model?

A common method of enhancing diversity in generative outputs is to modify the decoding strategy [23, 90, 33]. Through preliminary testing, I observe that diverse beam search decoding with Hamming distance reward following Vijayakumar et al. [90] improves the output diversity with less impact on accuracy compared to other methods.

On the other hand, Cao and Wan [7] propose modifying the model architecture by introducing latent variables to guide output variety. However, this approach only approximates

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, I explore the performance of three strategies for diverse generation of commonsense inferences.

Monomorphic Beam Search (M) This model receives as input a dialogue context \mathbf{C} consisting of the previous six utterances delimited by their corresponding speaker tags, the current response \mathbf{r} for which to generate inferences, and a commonsense question \mathbf{q} pertaining to one of the ten inference types (Table 4.2) in the following format:

$$\mathbf{C} \backslash n \mathbf{r} \backslash n \backslash n [\text{Question}] \mathbf{q} \backslash n [\text{Answer}]$$

It is trained to output a single inference i . During training, instances with multiple correct inferences 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. [90].

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 follow:

$$(1) \mathbf{i}_1; (2) \mathbf{i}_2; (3) \mathbf{i}_3; \dots$$

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 [33].

4.4.2 Model Configuration

I 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 [66] 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 hyperparameters are selected through grid-search based on their results on the validation sets.⁶ 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, I develop the HumanGenP model exclusively on multi-inference instances to facilitate learning of polymorphic outputs.

4.5 Generative Model Evaluation

I evaluate the six generative models (Section 4.4.2) on the ten commonsense inference types (Table 4.2) that exist in both the HumanGen (Section 4.3.2) and ConvoSense (Section 4.3.1) datasets. The model performance is evaluated using automatic reference metrics (Section 4.5.1)

⁶The 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”.

and human evaluations of reasonableness and novelty (Section 4.5.2).

4.5.1 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 my multi-inference dialogue generation objective. To address this, I structure my 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 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( $a$ )

```

Algorithm 1 demonstrates the **PolyAgg** aggregation function. The **PolyAgg** aggregation function 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 [6], which seeks to 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, I can accurately measure reference set coverage and prevent models that generate mere surface-level variations from scoring highly on datasets with diverse references. I use SciPy’s linear sum assignment solver, then calculate the mean

of the assigned scores for the final metric value. Dou et al. [11] 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, I 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:

$$C = \frac{|outs_e|}{|refs_e|} \quad (4.1)$$

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.

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

$$\frac{\sum_{e \in E} PolyAgg(outs_e, refs_e) * C * |refs_e|}{\sum_{e \in E} |refs_e|} \quad (4.2)$$

I use this evaluation scheme with three automatic metrics to measure the performance of the models. I include the traditional ngram-matching BLEU metric with $n \in [1, 4]$ [63], the token-level embedding-based metric BertScore⁷ [103], and sentence cosine similarity using

⁷BertScore: [microsoft/deberta-xlarge-mnli](https://github.com/microsoft/deberta-xlarge-mnli)

SentenceBert⁸ [71].

	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 4.6: 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, $\alpha = 0.05$).

Results I 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 4.2, 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 4.6 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 approach 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.

⁸SentenceBert: `all-mpnet-base-v2`

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, I 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 reasonableness of the generated inferences, but rather a difference in inference content between the datasets, which is supported by evidence in Section 4.2.3 showing that human-written generations are more often repetitive with the dialogue context than ChatGPT generations. To obtain a direct measure of the quality of the generated model inferences, a human evaluation is discussed next in Section 4.5.2.

4.5.2 Human Evaluations

I also evaluate the models through human assessment, in both the Top-1 and Top-5 setting. Based on automated evaluation outcomes, I 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 reasonableness and novelty of an inference (Sec. 4.2.2) and performing inference clustering to identify unique generated inferences per example.

Table 4.7 demonstrates ConvoSenseM*’s superior performance compared to the HumanGen models. ConvoSenseM* achieves a remarkable 93% reasonableness and 98% novelty, averaging

	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 4.7: 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**).

3.4 unique inferences per example. Indeed, similar results hold even when considering the Top-1 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, I 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.

4.6 Conclusion

This study has confirmed the feasibility of generating multi-faceted, contextually novel commonsense inferences for dialogue. The presented dataset, ConvoSense, surpasses existing commonsense resources by offering a more diverse range of inferences that afford higher contextual novelty for their dialogue contexts. The comprehensive evaluation demonstrates that models trained on ConvoSense outperform those trained on other datasets in producing plausible, detailed, and novel commonsense inferences across various dialogue scenarios. This achievement provides a state-of-the-art generative commonsense model that can be leveraged to improve downstream dialogue applications.

Chapter 5

Generate-Select-Respond: Commonsense-Augmented Dialogue Model

5.1 Introduction

This chapter targets the third research question: Does access to a multi-faceted and contextually novel social commonsense pool improve dialogue responses? In this chapter, I present Commonsense Inference Generate-Select-Respond (CSI-GSR), a three-stage dialogue model that considers a pool of commonsense inferences relevant to the current dialogue situation when generating the follow-up response. CSI-GSR specifically leverages the ConvoSense-trained commonsense generator from Chapter 4, which is the only existing model of its kind that affords access to a multi-faceted and contextually novel commonsense pool for dialogue contexts. CSI-GSR is compared against the native response generation capabilities of prominent LLMs, including ChatGPT3.5 and Llama2, as well as the state-of-the-art in commonsense-augmented open-domain dialogue modeling, ChatGPT + Doctor [8]. It is found that CSI-GSR substantially advances the informativeness of the responses, improving

the level of detail and specificity of the responses to the dialogue context, over all considered models while still maintaining overall response fluidity in the greater dialogue context.

5.2 Approach

Following the success of modular LLM-based dialogue agents (Section 2.1), I present a commonsense-augmented dialogue approach that contains a pipeline of LLM modules directing the response generation strategy, hereafter referred to as **CSI-GSR**. It consists of the following modules: (1) a powerful commonsense inference (**CSI**) generator that outputs diverse and contextually novel inferences (**G**), (2) a strategic inference selector (**S**), and (3) an inference-grounded responder (**R**). Figure 5.1 illustrates the proposed pipeline with examples.

Inference Generation (G) First, the pool of commonsense inferences are generated for a given dialogue context D using ConvoSenseM* (Chapter 4.4). This results in the commonsense inference set I , a set of commonsense inference candidates that are diverse and relevant to the dialogue context.

Inference Selection (S) Second, the commonsense inference selection ChatGPT module plans which inferences are useful towards generating an interesting and appropriate response to the dialogue context. It is tasked with selecting k inferences at a time from the full set of inferences I by being prompted to carefully consider each inference and strategically determine which inferences are the most useful, relevant, and interesting for writing the next response in the dialogue context. The selected inferences are outputted as the list I' and are a subset of I where $1 \leq |I'| < |I|$. This produces a set of commonsense inferences that are curated specifically for their utility towards generating a good follow-up response.

Response Generation (R) Third, the ChatGPT response generator module takes I' and D and outputs the next response r . It is instructed to synthesize the semantic content

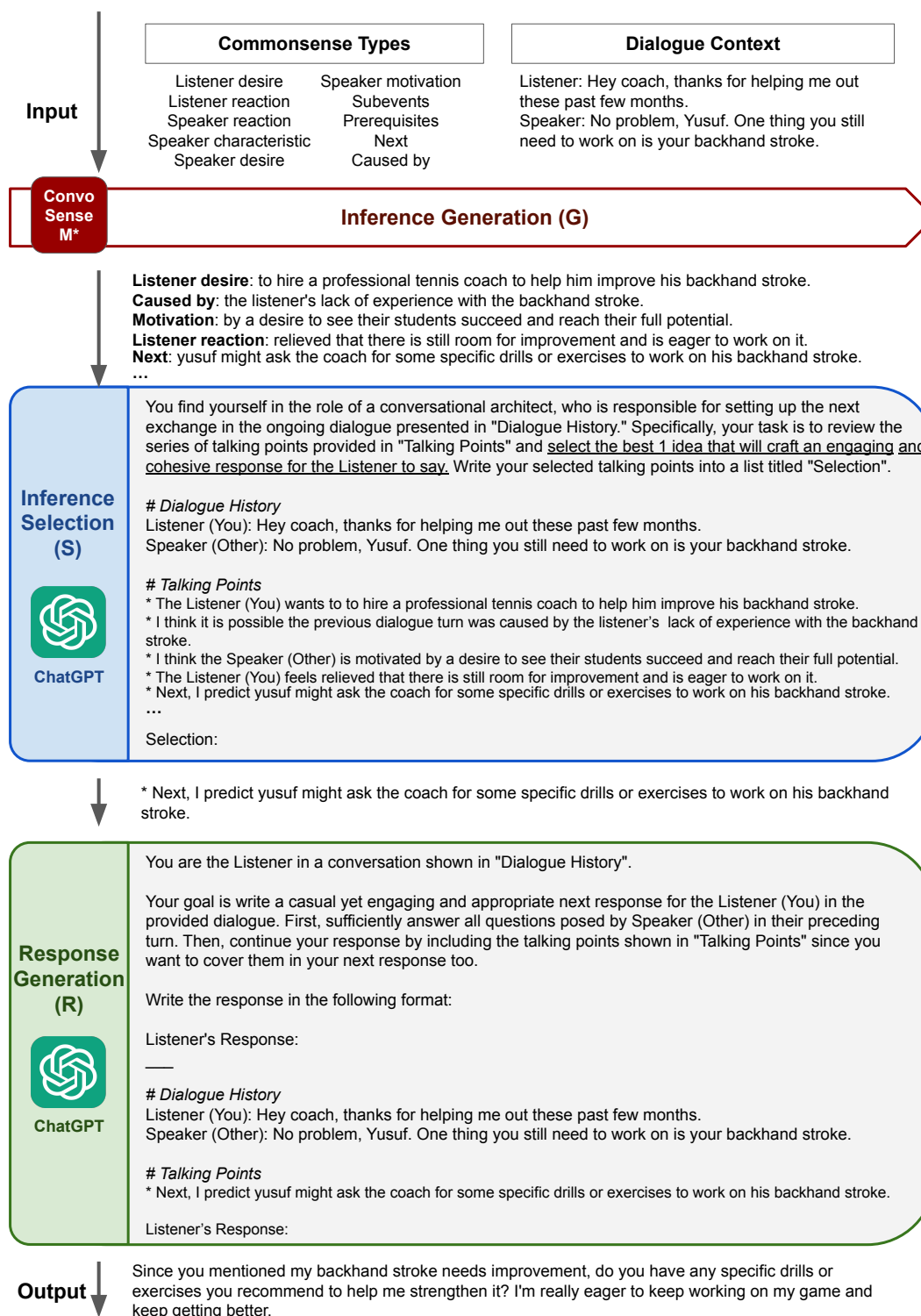


Figure 5.1: The module pipeline of CSI-GSR with finalized **Enumeration** prompt design for the prompt-based ChatGPT modules (S, R) (Section 5.3.1). Underline denotes instruction that changes for -single and -multi variants, with -single shown here (Section 5.3.2).

provided in the set of multiple inferences into an engaging and appropriate response. This produces a dialogue response that is natural and engaging in the dialogue by being grounded on k selected commonsense inferences.

5.3 Approach Development

The performance of an approach that relies on prompting LLMs is determined by the prompt that is provided to the LLM. The phrasing of the task and its instructions thus play a pivotal role in the outcomes achievable through such methods. However, this characteristic introduces a unique challenge when developing approaches using prompt-based LLM techniques. Namely, the reliance on natural language introduces an intractable exploration space of prompts, since there are countless ways of composing a natural language prompt to accomplish a singular task. It is impossible to test all possible natural language prompts, and there is no highly reliable way of estimating the impact of specific modifications, as of now.

5.3.1 Commonsense Inference Generate-Select-Respond: CSI-GSR

In light of these challenges of prompt development, I first identify a suitable Inference Generation module before focusing on the LLM-based Inference Selection and Response Generation modules, which utilize the output of the Inference Generation module.

Inference Generation (G)

Recent work on LLMs, including ChatGPT3.5, has revealed that they tend to overlook information situated in the “middle” of the prompt, especially as the length of the prompt grows [47]. With this in mind, I decide to limit the number of inferences that are generated from the Inference Generation module, in order to better situate the subsequent modules to reason over the inferences that they are provided. Specifically, the Inference Generation module will output a set of ten commonsense inferences in total, one corresponding to each

of the ten commonsense types covered in the ConvoSenseM* model. I first experiment with outputting the top-1 inference for each type, but observe that there can be high semantic overlap between the inferences outputted between different types. Since the hypothesis to be tested in this work specifically targets the utility of diverse and multi-faceted social commonsense inferences as response content candidates, the level of diversity in the inferences should be maximized to the greatest extent possible. Thus, overlapping inferences are undesirable. To mitigate this, I next experiment with ways of improving the diversity of the final set of commonsense inferences. Notably, the results in Chapter 4 demonstrate that the top-5 outputs of the ConvoSenseM* model attain a remarkable 68% inference uniqueness while maintaining high reasonableness and novelty. In practice, this means that within a specific dialogue context and for a given type of commonsense, an average of 3 to 4 inferences represent unique semantic information from one another out of the 5 that are outputted. In light of this, I implement an approach for the Inference Generation module that outputs the top-5 outputs for each inference type t in ConvoSense and then selects a single inference for each type such that between-type cosine similarity of inference embeddings is minimized. Manual analysis of the results indicates that this approach improves the diversity of the inference set outputted by the Inference Generation module, without too substantial of an impact on the reasonableness of the inferences.

Inference Selection (S) and Response Generation (R)

Next, I focus on on the Inference Selection and Response Generation modules, which will ingest the outputs from the previously described Inference Generation module as part of their prompts. I first identify a small set of prompts that are likely to accomplish the desired tasks of Inference Selection and Response Generation, based on trial-and-error experimentation, and then perform a more thorough performance comparison between this finite set. I aim for high diversity in the considered prompts, where the prompts are designed to approach the task from different perspectives, rather than being surface-form modifications to the vocabulary

and phrasings of a common approach. In this way, even though I am only considering a small set of prompts in full, I approximate wide coverage of the prompt search space to optimize the likelihood of finding a successful strategy. In total, I compare the performance of 4 distinct strategies for the Inference Selection and Response Generation modules of CSI-GSR:

Enumeration the Inference Selection module generates a list of interesting and engaging talking points intended to inspire a compelling follow-up response. The Response Generation module then incorporates these talking points into the generated response. The prompts are shown in Table 5.1.

Structured Composition the Inference Selection module is prompted to identify the best composition of commonsense inferences for constructing a response by choosing a main commonsense inference to drive the response and a set of supplementary inferences to provide supporting details for the main point. The Response Generation module receives these grounding inferences categorized into two groups: the main point and supporting points. The prompts are shown in Table 5.2.

Group Partition the Inference Selection module is instructed to partition commonsense inferences into groups, each of which could potentially generate a suitable response, and to then identify the optimal group. The Response Generation module then receives the chosen group for generating the response. The prompts are shown in Table 5.3.

Planning the Inference Selection module is prompted to articulate a natural language description outlining the plan for the subsequent response. The explanation must include an indication of a selected set of commonsense inferences. The Response Generation module receives both the selected inferences and the explanation to construct the response accordingly. The prompts are shown in Table 5.4.

For all prompts, the dialogue context is provided as a sequence of turns, separated by newlines,

Enumeration	
S	<p>You find yourself in the role of a conversational architect, who is responsible for setting up the next exchange in the ongoing dialogue presented in “Dialogue History.” Specifically, your task is to review the series of talking points provided in “Talking Points” and curate a selection of 2 to 4 non-repetitive ideas that will craft an engaging and cohesive response for the Listener to say when combined. You want to strike a balance between selecting enough talking points to construct an interesting response without selecting too many that would result in a longwinded, overwhelming, or unfocused response. Write your selected talking points into a list titled “Selection”.</p> <pre># Dialogue History {context} # Talking Points {inferences} Selection:</pre>
R	<p>You are the Listener in a conversation shown in “Dialogue History”.</p> <p>Your goal is write a casual yet engaging and appropriate next response for the Listener (You) in the provided dialogue. First, sufficiently answer all questions posed by Speaker (Other) in their preceding turn. Then, continue your response by including the talking points shown in “Talking Points” since you want to cover them in your next response too.</p> <p>Write the response in the following format:</p> <pre>Listener’s Response: --- # Dialogue History {context} # Talking Points {Output from S} Listener’s Response:</pre>

Table 5.1: The Inference Selection (S) and Response Generation (R) prompts used for the Enumeration prompt design. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated commonsense inferences, respectively, of a provided example. Placeholder “{Output from S}” is replaced by the output from ChatGPT from the Inference Selection module.

 Structured Composition

S	<p>You are a conversational architect, who is responsible for creating a response outline for the next response in the ongoing dialogue presented in “Dialogue History”.</p> <p>A response outline consists of 1 MAIN POINT and 1 - 3 SUPPORTING POINTS that the Listener will say in their immediate next response. You are given a set of possible talking points shown in “Talking Points” from which you need to put together the outline. The foremost purpose of MAIN POINT is to answer all of the questions asked in the previous turn by the Speaker. If there are any questions to answer, then “Main Point” should consist of a question-answering action (e.g. the Listener (You) will answer the Speaker’s questions).</p> <p>If there are no questions to answer, then MAIN POINT either proactively contributes new information to the conversation from the perspective of the Listener, or follows up to elaborate on what was shared by the Speaker. The MAIN POINT must be taken from the list “Talking Points”.</p> <p>SUPPORTING POINTS either function as transitions into the main point by introducing its relevance to the conversation, or as trailing addendums in order to offer additional details expounding on the content of the MAIN POINT. All supporting points must be taken from the list “Talking Points”.</p> <p># Dialogue History {context}</p> <p># Talking Points {inferences}</p> <p>Main Point:</p>
R	<p>You are the Listener in the ongoing conversation shown in “Dialogue History”, who is provided an outline of how to structure your next response in the form of “Main Point” and “Supporting Points”.</p> <p>Begin by determining how to express the main point outlined in the “Main Point” section, focusing on either proactively contributing new information from the perspective of the Listener or following up on the Speaker’s input for further elaboration.</p> <p>Then, integrate the “Supporting Points” into your response, using them as transitions to establish the relevance of the main point or as trailing addendums to provide additional details that enhance the content of the main point.</p> <p>Pay careful attention to the flow of your response, ensuring a coherent and logical progression between main and supporting points, and also consider the overall flow of the conversation given your response to make sure it is natural and cohesive.</p> <p>It is critical that ALL points in both “Main Point” and “Supporting Points” need to be mentioned in your response.</p> <p># Dialogue History {context}</p> <p># Talking Points {Output from S}</p> <p>Listener’s Response:</p>

Table 5.2: The Inference Selection (S) and Response Generation (R) prompts used for the Structured Composition prompt design. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated commonsense inferences, respectively, of a provided example. Placeholder “{Output from **S**” is replaced by the output from ChatGPT from the Inference Selection module.

Group Partition	
S	<p>First, think about how to split the talking points in “Talking Points” into groups of 2 or 3 relevant talking points that can be said together to form an appropriate and interesting response in the ongoing conversation shown in “Dialogue History”. The only thing you should output is what you think the best group is for making the next response. Output it as a list labeled “Best Group”.</p> <p># Talking Points {inferences}</p> <p># Dialogue History {context}</p>
R	<p>Write the Listener’s next response for ongoing conversation shown in the “Dialogue History” based on the talking points in “Best Group” as “Listener’s response”. Make sure the response is natural as the next response in the conversation, synthesizes the talking points in the group together, and avoids being longwinded or rambling.</p> <p># Dialogue History {context}</p> <p># Best Group {Output from S}</p>

Table 5.3: The Inference Selection (S) and Response Generation (R) prompts used for the Group Partition prompt design. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated commonsense inferences, respectively, of a provided example. Placeholder “{Output from **S**}” is replaced by the output from ChatGPT from the Inference Selection module.

Planning	
S	<p>You are teaching someone to be a better conversationalist through a demonstration of constructing a response from the Listener in the example dialogue shown in “Dialogue History”. First, explain how you would use the talking points shown in “Talking Points” to construct the next response as the Listener in the ongoing conversation shown in “Dialogue History” in “Explanation” field. Then, output the list of talking points needed to execute your explanation in “Selection” field.</p> <p># Dialogue History {context}</p> <p># Talking Points {inferences}</p> <p>Explanation:</p>
R	<p>You are teaching someone to be a better conversationalist through a demonstration of constructing a response from the Listener in the example dialogue shown in “Dialogue History” from a set of provided talking points in “Talking Points”. First, sufficiently answer all questions posed by Speaker (Other) in their preceding turn. Then, continue your response by including the talking points shown in “Talking Points”.</p> <p>Write the response in the following format:</p> <p>Listener’s Response: ---</p> <p># Dialogue History {context}</p> <p># Talking Points {Output from S}</p> <p>Listener’s Response:</p>

Table 5.4: The Inference Selection (S) and Response Generation (R) prompts used for the Planning prompt design. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated commonsense inferences, respectively, of a provided example. Placeholder “{Output from **S**}” is replaced by the output from ChatGPT from the Inference Selection module.

where each turn is prefixed by a nominal tag. The terminal turn is prefixed as “Speaker (Other)”, and its preceding utterance is labeled as “Listener (You)”. These nominal tags are then assigned in alternating order to the remainder of the turns. These nominal tags encourage ChatGPT to take on the role of the Listener in the conversation when constructing the follow-up response. Additionally, the commonsense inferences generated from the Inference Generation module (ConvoSenseM*) are prefixed with natural language text based on their type as indicated by Table 5.5 to transform the generated inferences into complete natural language sentences before being used as input in a prompt.

Type	Prefix
Cause	I think it is possible the previous dialogue turn was caused by
React_o	The Listener (You) feels
React	I think the Speaker (Other) feels
Subsequent	Next, I predict
Attribute	I think the Speaker (Other) is
Desire_o	The Listener (You) wants
Constituent	I think it is possible the previous dialogue turn depends on
Motivation	I think the Speaker (Other) is motivated
Prerequisite	I think it is possible the previous dialogue turn requires
Desire	I think the Speaker (Other) wants

Table 5.5: Prefixes used to transform the generated commonsense inferences into natural language sentences when used as inputs into the Inference Selection and Response Generation prompts.

Enumeration	Structured Composition	Group Partition	Planning
3.45	3.05	2.88	2.48

Table 5.6: The average quality score for each prompt design.

To identify the best prompt, I rate the quality of the outputted responses from each prompt design on 20 dialogues from the development set of the ConvoSense dataset. Outputs from both `gpt3.5-turbo-0613` and `gpt3.5-turbo-1103` are included. Because the OpenAI model lifecycle deprecates snapshots relatively quickly, it is important to provide a more generalized assessment of each prompt, rather than be restricted to a specific snapshot. Table 5.6 shows the average quality score for each prompt, with the highest performance being observed for

the **Enumeration** prompt. As a result of its superiority to the other prompts under study, the **Enumeration** design is chosen to be used for the CSI-GSR approach.

5.3.2 Pilot Study

Having optimized the prompt design for CSI-GSR in the previous section, a pilot study is conducted to estimate the performance impact of this response generation approach. Two versions of CSI-GSR are investigated: **CSI-GSR-single** and **CSI-GSR-multi**. For both variants, the Response Generation module is kept the same, but it receives a different set of grounding commonsense inferences from two different Inference Selection modules as follows:

CSI-GSR-single restricts the Inference Selection module to select the best, single commonsense inference to guide the follow-up response, with the instruction “select the best 1 idea that will craft an engaging and cohesive response for the Listener to say” placed at the underlined portion of the prompt shown in Figure 5.1.

CSI-GSR-multi prompts the Inference Selection module to select multiple commonsense inferences that provide interesting and useful information to guide the follow-up response, with the instruction “curate a selection of 2 to 4 non-repetitive ideas that will craft an engaging and cohesive response for the Listener to say when combined. You want to strike a balance between selecting enough talking points to construct an interesting response without selecting too many that would result in a longwinded, overwhelming, or unfocused response” placed at the underlined portion of the prompt shown in Figure 5.1.

The two variants of CSI-GSR (**CSI-GSR-single** and **CSI-GSR-multi**) are compared against one another, and against the baseline response generation capability of ChatGPT. For baseline performance, ChatGPT is instructed to produce the next follow-up response given a dialogue context using the prompt shown in Table 5.7.

To perform this comparison, an external conversational AI expert, unaffiliated with this

ChatGPT

```
# Dialogue History
{context}

You are the Listener in a conversation shown in “Dialogue History”.

Your goal is write a casual yet engaging and appropriate next response for the Listener
(You) in the provided dialogue.

Write the response in the following format:

Listener’s Response:
---
```

Listener’s Response:

Table 5.7: The prompt used for native response generation of ChatGPT. Placeholder “{context}” is replaced by the dialogue context of a provided example.

study, is shown a response from CSI-GSR-single, CSI-GSR-multi, and ChatGPT for 60 dialogues from the development set of ConvoSense. They are instructed to select which of the three responses is the best option for the follow-up response in the dialogue context of that example, with no ties allowed. Figure 5.2 shows the win-rate for each model.

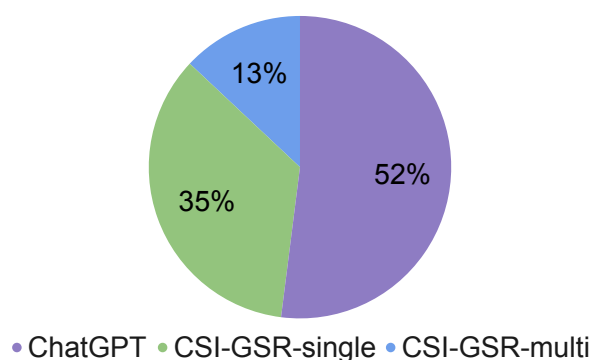


Figure 5.2: Proportion of instances where the specific model’s output is selected as the preferred option.

Surprisingly, responses from ChatGPT are substantially preferred by the evaluator, with a win-rate of 52%, outperforming both versions of CSI-GSR. Furthermore, it can be seen that CSI-GSR-single is preferred over CSI-GSR-multi. These results suggest that the incorporation

of social commonsense under the current CSI-GSR pipeline does not improve the quality of the dialogue responses over what ChatGPT is able to do natively, and that response quality is further degraded with multi-inference grounding.

	CSI-GSR-multi	CSI-GSR-single
Longwinded	40%	17%
Suboptimal Follow-up Topic	38%	67%
Context Mixup	11%	3%
Topic Order	7%	0%
Presentation of Speculative Inferences as Facts	4%	14%

Table 5.8: Error analysis of generated responses from each CSI-GSR model variant, on responses that fail to be preferred to responses from native ChatGPT. Analysis results show the percentage of responses belonging to various error categories, where categories are mutually exclusive.

In an effort to understand these unexpected results better, the external evaluator is asked to provide a short explanation of their response preferences for the examples. Then, an error analysis is performed, guided by the provided explanations. Table 5.8 shows the rates of different errors affecting the responses generated by both variants of CSI-GSR, where the categories are:

Longwindedness Responses are unfocused and unnaturally verbose, resulting in an overwhelming dialogue experience.

Suboptimal Follow-up Topics In some cases, the semantic content introduced in the response is a relevant but less important aspect of the topic being discussed, which feels unnatural as a follow-up response in the dialogue.

Context Mixup In this type of error, the response is misplaced in the conversation, referencing a previous turn rather than the most recent one.

Topic Order Sometimes, although the response content is appropriate, the order in which it is expressed feels disjointed or unnatural.

Presentation of Speculative Inferences as Facts Errors in this category occur when the selected inference is predictive of future events, yet is expressed in the response as if the future event already happened.

Based on the results of this pilot study and error analysis, it is clear that ChatGPT is unable to leverage the external resource of commonsense inferences in a manner that improves response generation, suffering from a multitude of errors on both how to select salient follow-up topics from the commonsense inferences as well as how to incorporate them into a coherent and appropriate response to the current dialogue.

In general, CSI-GSR is vulnerable to producing a response that expresses an unexpected follow-up topic, such as by choosing to focus on an aspect of what is being discussed that has low salience or low likelihood (e.g. proposing a snack break at the onset of a homework assistance session, Table 5.9). Both multi and single variants of CSI-GSR are dominated by this error type. Furthermore, CSI-GSR also encounters issues with coherence of the outputted responses, such as producing a response that is a follow-up to some historical utterance rather than the current state of the dialogue (Table 5.10), outputting thoughts in a response in an unnatural order (Table 5.11), or presenting selected speculative commonsense inferences as observed events in the world (Table 5.12).

Specific to CSI-GSR-multi, longwinded responses also are a critical limiting factor, in which the responses are unfocused and present too much information at one time rendering them overwhelming and hard to follow (Table 5.13). This is most likely due to the selection of inferences that each present independent talking points, which are then concatenated into a single response. Based on the decision to optimize the Inference Generation module for diversity, this effect of the inferences in the pool being independent talking points is encouraged. It could be the case that allowing for greater semantic similarity between the inferences helps to mitigate this longwindedness effect. However, as the main hypothesis being tested in this study focuses on the utility of diverse commonsense inferences for dialogue models, resolving this issue specific to multi-inference grounding is left for future work.

Dialogue Context	
<p>A: Hey, you dropped your pencil. B: Thanks, I appreciate it. A: No problem, I know how it feels to be in need of a pencil and not have one. B: Yeah, it's frustrating. But at least you were there to help me out. Thanks again. A: You're welcome. So, what were you working on? B: Just some homework for my math class. It's due tomorrow and I'm still struggling with it. A: Let me take a look. I'm pretty good at math and maybe I can help you understand it better. B: That would be great! Thank you so much for your help, Jan.</p>	
CSI-GSR	ChatGPT
<p>You're welcome, Max. I'm glad I could help you out. Math can be tough sometimes, but don't worry, we'll figure it out together. Before we dive into the homework, though, how about we take a break and grab a snack? It's always good to recharge and refuel before tackling a challenging task. What do you say?</p>	<p>Sure, I'd be happy to help! What specific topic or problem are you struggling with in your math homework?</p>

Table 5.9: Example of **Suboptimal Follow-up Topics** response error category. The segment of the response that expresses the error type is bolded.

Dialogue Context	
<p>B: Hey, Sarah! I'm so glad you came to the beach with me today. A: Yeah, it's really nice out here. The weather is perfect and the water looks so inviting. B: I know! I can't wait to go for a swim. But first, let's build a sandcastle! A: That sounds like fun. What do you need me to do? B: Well, first we need to find a good spot to build our castle. Somewhere where the sand is nice and firm. Do you see a spot that looks good to you?</p>	
CSI-GSR	ChatGPT
<p>I'm glad I came too! The beach is always so relaxing. Building a sandcastle sounds like a great idea. Let's start by looking for a spot with firm sand. I see a spot over there that looks promising. What do you think?</p>	<p>Sure, how about over there? The sand looks pretty packed and it's close enough to the water to enjoy the view while we build.</p>

Table 5.10: Example of **Context Mixup** response error category. The segment of the response that expresses the error type is bolded.

Dialogue Context	
A: Hi, I'm Ryken. B: Hi, I'm James. I need a lawyer for a divorce.	
CSI-GSR	ChatGPT
Hi James, nice to meet you. I'm sorry to hear that you're going through a divorce. I can definitely help you with that. Do you have any specific concerns or questions about the divorce process? It's important to address any issues or uncertainties you may have so that I can assist you in the best way possible. And I'm glad that you've reached out for legal help. It takes courage to take this step, but it's important to have the support and guidance you need during this difficult time.	Hi James, nice to meet you. I'm sorry to hear that you're going through a divorce. I can definitely help you find a lawyer who specializes in divorce cases. Can you tell me a bit more about your situation so I can assist you better?

Table 5.11: Example of **Topic Order** response error category. The segment of the response that expresses the error type is bolded.

Dialogue Context	
B: Hey, Mom. A: Hey, Delanie. What's up? B: I crashed my bike.	
CSI-GSR	ChatGPT
Oh no, Delanie! Are you okay? I'm really sorry to hear about your bike. Do you need any help fixing it? I'm glad to hear that you didn't get hurt though. That must have been a scary experience. Let me know if there's anything I can do to help.	Oh no, are you okay? Did you get hurt?

Table 5.12: Example of **Presentation of Speculative Inferences as Facts** response error category. The segment of the response that expresses the error type is bolded.

Dialogue Context	
A: So, what can you tell me about the history of our town?	
B: Well, it was founded back in the 18th century by a group of settlers from France. They were looking for a new place to call home, and they found this land. They built a few houses and a church, and that's how the town started. Over the years, more and more people came to live here. We have records of all sorts of people living in the town, from farmers to shopkeepers to blacksmiths. There's even a story about a witch who lived here back in the 1700s!	
CSI-GSR	ChatGPT
Wow, that's really fascinating! I'm especially intrigued by the story about the witch. Could you tell me more about her? Like, what was her name and what kind of things did she do? It's so cool to hear about the history of our town and all the different people who lived here. Speaking of which, what's the current population of our town? And what is it like to live here today?	That's really fascinating! I love hearing about the origins of our town and how it has evolved over the years. It's amazing to think about all the different people and stories that have shaped our community. I'm especially intrigued by the story of the witch who lived here in the 1700s. Do you know any more details about her?

Table 5.13: Example of **Longwindedness** response error category. The segment of the response that expresses the error type is bolded.

Instead, I focus only on investigating the impact of a diverse and predictive pool of common-sense inferences as response content candidates for dialogue modeling through CSI-GSR-single.

5.3.3 Approach Improvement

One candidate explanation for the limitations observed for the CSI-GSR-single approach is that the Inference Selection module is failing to perform adequately, especially seeing as the most common error involves an unimportant or low salience follow-up topic in the generated response. Under the CSI-GSR approach, the response content should come directly from the selected inferences, thus an issue with the response could be derived from an issue with the selected inference. On the other hand, it could be the case that Inference Selection module is performing sufficiently but the Response Generation module fails to incorporate the selected inference when constructing the response. In light of these possible causes, I analyze the 60 development examples on the following two characteristics:

1. Selection Bias: one possible point of failure for CSI-GSR is in the Inference Selection

module providing a simplistic inference selection mechanism resulting in poorly selected inferences, such as by being biased towards selecting the first or last inference from the provided inference candidates; thus, I aim to detect any patterns in selection that may be indicative of poor selection capabilities.

2. **Response Faithfulness:** a second possible point of failure for CSI-GSR is in the Response Generation module failing to incorporate the selected commonsense inferences; thus, I aim to identify whether this is a current issue in order to take appropriate action to mitigate it if it is a problem

Response Faithfulness It is observed that 90% of the responses incorporate details from the selected commonsense inferences. This still leaves the question of whether the way that the inference is being expressed in the response is suboptimal, perhaps by being phrased as a question when it is more natural being posed as a thought or prediction. However, manual examination suggests that it is not the presentation of the inference content itself that is found to be problematic. Instead, the selected commonsense inference is often contributing fundamentally unnatural focal points, which would produce odd responses no matter how they are incorporated or expressed in the response. In light of this, the analysis on the Inference Selection module becomes even more pressing.

Selection Bias It is observed that the Inference Selection module has a high likelihood of selecting commonsense inferences whose semantic content directly indicates a follow-up action in the conversation, since 78.3% of the examples select an inference of type **Desire**, **Desire_o**, or **Subsequent**. For instance, the **Subsequent** inference of “Next, I predict the Listener might ask the Speaker for some specific drills or exercises to work on his backhand stroke” is providing an explicit instruction on what to say in the next response and how to say it. Selecting inferences like this that explicitly convey follow-up actions is not necessarily an indicator of poor selection behavior, since such semantic content is often salient for

conversational behaviors.

However, it is possible that ChatGPT selects these inferences because they explicitly provide guidance on what to say next in the conversation, whereas other non-action-prescribing inferences are less obviously applicable to follow-up response prediction since they involve a potentially complex extrapolation from the commonsense idea to a coherent response. For instance, a **Motivation** inference of “I think the Speaker is motivated by a desire to see their students succeed and reach their full potential” does not directly translate to an action to take in the conversation, thus it is much less straightforward how to incorporate this content into a follow-up response. One option is produce a response that acknowledges the Speaker’s commitment to their student’s success. A second option is to produce a response that asks the Speaker about the significance they place on their students’ high levels of achievement. In fact, there are many other possible ways of expressing such an inference than the two mentioned. Consequently, in order to evaluate the true utility of such inferences as response content candidates, ChatGPT would need to be capable of identifying and evaluating all of these different versions of expression. However, the suggestion from the observed selection bias is that ChatGPT is likely unable to do this well in the current implementation of CSI-GSR as evidenced by its failure to select such inferences with high expression-variability most of the time.

With this in mind, I next experiment with a modification to CSI-GSR to encourage the Inference Selection module to more fully consider *all* of the commonsense inference candidates for response generation.

Commonsense Inference Generate-Respond-Select: CSI-GRS

The goal of this modification is to improve the capability of ChatGPT to assess each commonsense inference towards its utility of contributing to the next response in the dialogue. Based on the observed behavior for CSI-GSR, I hypothesize that the Inference Selection

module is performing a simplistic selection procedure due to the disconnect between inference and follow-up response, which incurs a multi-step reasoning process to project the usage of inference in response generation, thus encouraging ChatGPT to prefer the selection of straightforward, action-prescribing inferences with their lower projection demands. One approach to improve the capability of assessing the utility of each commonsense inference for response generation is to re-order the Inference Selection and Response Generation modules, such that the first step in the response generation pipeline is to construct a response for each inference and then instruct ChatGPT to select the best response from that candidate set. In this way, the Inference Selection module does not need to perform any implicit reasoning about how a commonsense inference could contribute to the response generation, rather it is provided the actual resultant response from incorporating each inference and only needs to focus on assessing the quality of each fully formed response as the follow-up response in the dialogue.

Following this approach, I redesign CSI-GSR to CSI-GRS as follows:

Inference Generation (G) This is unchanged from CSI-GSR, using ConvoSenseM* to generate a set of t commonsense inference candidates I that are relevant to the dialogue context.

Response Generation (R) Second, the ChatGPT response generator module takes I and D and outputs $|I|$ next responses R , where each $r \in R$ corresponds to a follow-up response that incorporates one $i \in I$. It is instructed to synthesize the semantic content of each inference into an engaging and appropriate response.

Inference Selection (S) Third, the commonsense inference selection ChatGPT module assesses which inference-grounded response $r \in R$ is the best response to the dialogue context.

Upon initial experimentation, this new Inference Selection module is instructed to select the best follow-up response from the provided set of candidates from the new Response

Generation module. However, this results in an index-biased selection, preferring to select one of the first three responses from the Response Generation module for all development examples. To mitigate this effect, I explore two iterative strategies of selection in which the Inference Selection module is called $|R| - 1$ times to pinpoint the best response. The **Pruning** variant instructs ChatGPT to identify the worst response from a set of candidates. After each iteration, the selected worst response is removed from the set of candidates and ChatGPT is called again with this reduced set of candidates. This continues until only a single response remains in the set, which is considered to be the selection of the overall best response. On the other hand, the **Maximization** variant instructs ChatGPT to identify which of two response candidates is better. Upon selection, ChatGPT is called again to compare the selected response against an unseen response candidate from R . After the $|R| - 1$ -th iteration, the selected response is considered to be the selection of the overall best response.

By comparing the response outputs of the **Pruning** and **Maximization** variants on 30 dialogues from the development set of ConvoSense, it is observed that the **Pruning** variant is better. Next, I compare the quality of the responses from this winning **Pruning** variant of CSI-GRS and the original CSI-GSR approach using the same evaluation setup. This comparison analysis reveals that CSI-GRS does not produce higher quality responses, with a win-rate of only 27% against the original CSI-GSR. Upon examination of the outputs of CSI-GRS, it is observed that CSI-GRS is less biased in its inference selection; however, the responses produced by CSI-GRS are often less coherent to the previous dialogue context, failing to address the specific previous turn. This suggests that the Response Generation module struggles to adequately incorporate several of the commonsense types into coherent and appropriate responses, and that the Inference Selection module struggles to identify high-quality responses. Consequently, simply switching the order of the Inference Selection and Response Generation modules is insufficient for correcting the observed limitations of the CSI-GSR approach.

Based on the results of this experimentation on restructuring the approach, further evidence is gathered that the zero-shot approach to Inference Selection and Response Generation is unsuccessful. I next look to strategies on teaching large language models to perform a new task as the final approach to improve the performance of CSI-GSR. The common practice is to perform fine-tuning using a dataset of task examples. However, there is no such existing dataset for CSI-GSR-single and no promising approach to obtain such a dataset automatically. Because ChatGPT-based CSI-GSR-single results in low preference responses and high rates of detrimental characteristics (Pilot Study, Section 5.3.2), it is not suitable to use ChatGPT to construct a dataset, since it would construct a dataset with these undesirable characteristics which would then leak into any approach that leverages such a dataset. Consequently, the most likely option is to utilize human annotators; however, this incurs a prohibitively high cost. Recently, previous work has revealed that LLMs can obtain a significant performance gain simply by providing several examples of successful task completion, rather than finetuning on a large dataset of task examples. Motivated by such works, I explore two techniques to teach large language models to perform CSI-GSR-single successfully, leveraging recent techniques on few-shot in-context-learning and few-shot finetuning.

5.3.4 Few-shot Learning

As discussed previously, there is no suitable approach to automatically generate high-quality and successful examples of the CSI-GSR approach to dialogue response generation. As a result, I undertake the task of creating few-shot examples for CSI-GSR-single by manually performing the selection and subsequent response generation steps for a given dialogue. For each of the 10 commonsense types covered in ConvoSense, I construct 10 examples for both the Inference Selection and Response Generation modules of CSI-GSR, resulting in a total of 100 examples for each module. Algorithm 2 portrays the process that was followed for few-shot example construction. To construct these examples, dialogues are sampled from the training split of the ConvoSense dataset. Each sampled dialogue is considered in turn as to

its utility as a few-shot example for CSI-GSR response generation. If the sampled dialogue is too similar to a different few-shot or is of low quality due to repetitiveness or nonsensicalness, it is excluded from being a few-shot. After verifying the suitability of the dialogue, the quality of its corresponding inferences is considered next. These inferences are those generated by ChatGPT following the ConvoSense dataset construction process, in which one generated inference is taken for each commonsense type. From the ten inferences, the one leading to the most suitable and engaging follow-up response is selected. If the chosen inference already has ten examples in the existing few-shots, the next best one is considered until either an underrepresented inference is found or none of the remaining inferences lead to a good follow-up response. If a good inference cannot be found for the sampled dialogue, then it is excluded from being a few-shot. If a sampled dialogue passes both the dialogue quality and inference suitability verifications, then a follow-up response is crafted based on the selected inference. The dialogue, along with the selected inference and written inference-grounded response, is then included as a few-shot example. Other than the initial dialogue sampling, all of the work performed during this process is done by a human in order to construct high-quality and human-verified examples.

Algorithm 2 Few-shot Examples Construction

```

1:  $fewshots \leftarrow []$ 
2: while LENGTH( $fewshots$ ) < 100 do
3:    $dialogue, inferences \leftarrow \text{SAMPLE}(\text{ConvoSense}_{training})$ 
4:   if  $\exists shot_{dialogue} \in fewshots$  s.t. SIMILAR( $dialogue, shot_{dialogue}$ ) then
5:     continue
6:   if POOR_QUALITY( $dialogue$ ) then
7:     continue
8:    $selected_i \leftarrow \text{BEST}(inferences)$ 
9:   while  $selected_i \neq null$  and COUNT( $fewshots, \text{type}(selected_i)$ ) == 10 do
10:    REMOVE( $inferences, selected_i$ )
11:     $selected_i \leftarrow \text{BEST}(inferences)$ 
12:   if  $selected_i == null$  then
13:     continue
14:    $response \leftarrow \text{GROUNDED\_RESPONSE}(dialogue, selected_i)$ 
15:   APPEND( $fewshots, (dialogue, selected_i, response)$ )

```

Following the initial creation of few-shot examples for CSI-GSR, a second round of curation is conducted to refine the examples. This refinement aims to optimize the naturalness of responses and ensure strong connection between the response and the selected commonsense. First, I revisit each example to assess the appropriateness of the selected commonsense out of the 10 options. There is one instance in which I determine that all commonsense inferences are too unnatural as response candidates, in which case this few-shot example is replaced by a better fitting one. In addition, there are several examples that necessitated a reselection of the optimal commonsense inference to improve the naturalness of the resulting response. As a result of this refinement, there is a slight distribution shift in the number of examples per commonsense type. Specifically, **Desire** and **Motivation** have 9 examples each, **React_o** has 12 examples, and the remaining types retained 10 examples. Second, each written response is modified by ChatGPT using the instruction “Make this a better conversational response” and the resulting response is used if it improved the naturalness or engagingness of the response based on manual analysis. Table 5.14 shows examples of the human-curated few-shots resulting from this curation process.

In-Context-Learning One strong application of few-shot examples is for in-context-learning for prompt-based approaches using LLMs, where relevant few-shots are concatenated to the prompt to explicitly show the LLM the intended task behaviors [5]. Following this, I develop CSI-GSR-icl, utilizing the created few-shots as in-context-learning examples for the CSI-GSR pipeline. For the Inference Selection module, I handpick 10 examples from the 100 selection few-shots, such that there is one example of each commonsense type being selected. For the Response Generation module, the 10 response generation few-shot examples are retrieved at runtime that match the type of the selected commonsense inference being passed to the Response Generation module. In this way, the Response Generation module is given 10 strong examples of utilizing the selected commonsense type in the follow-up response. For each module, these few-shots are concatenated into the prompts shown in Figure 5.15, which

Dialogue	Inference Selection	Response Generation
<p>Speaker (Other): Hey there! Thanks for taking the time to talk with me. It's really nice out here, isn't it?</p> <p>Listener (You): Yeah, it is. I love coming to this lake to relax and clear my head. There's something about being surrounded by nature that just makes me feel at peace.</p> <p>Speaker (Other): I know what you mean. I came out here today specifically to try and de-stress a bit. work has been hectic lately and I just needed a break.</p>	<p>The previous dialogue turn was caused by a recent project deadline at work that put a lot of pressure on the speaker.</p>	<p>What's been going on at work? Some big project or deadline working you into the ground?</p>
<p>Speaker (Other): Hey, I'm glad I ran into you. I've been meaning to talk to you about something.</p> <p>Listener (You): Oh, yeah? What's up?</p> <p>Speaker (Other): Well, remember how we were talking about that new movie that came out last week? The one with the big twist ending?</p>	<p>The Speaker (Other) wants to share their own theory about the twist ending and hear the listener's thoughts on it.</p>	<p>Yeah, I remember. I've still been thinking about that ending over and over, trying to figure out what it means for the next movie in the series. You have any theories about it?</p>
<p>Listener (You): Hi there. Are you okay? You look a bit hungry.</p> <p>Speaker (Other): Yeah, I am. I haven't had anything to eat in a while.</p>	<p>Next, speaker might start feeling weak and tired due to low energy levels.</p>	<p>Do you want to get something to eat with me? I don't want you to feel too weak or tired from being hungry.</p>

Table 5.14: Examples of human-curated few-shots.

result from modifications to the **Enumeration** prompt design to allow for the inclusion of few-shot examples.

Fine-tuning There have been recent works showing the success of few-shot and low-resource fine-tuning on LLMs, especially when using parameter-efficient fine-tuning methods [36, 46, 59]. In particular, I opt to employ Low-Rank Adaptation (LoRA) [30] for few-shot fine-tuning an LLM to perform both commonsense inference selection and inference-grounded response generation for the CSI-GSR approach.

LLMs, by design, are pre-trained on massive datasets, enabling them to capture broad linguistic understanding and generalizability. However, directly fine-tuning an LLM on a

 Few-shot Enumeration

S	<p>You find yourself in the role of a conversational architect, who is responsible for setting up the next exchange in the ongoing dialogue presented in “Dialogue History.” Specifically, your task is to review the series of talking points provided in “Talking Points” and select the best 1 idea that will craft an engaging and cohesive response for the Listener to say. Write your selected talking point into a list titled “Selection”.</p> <p>Review the following examples of good selections for different pairs of “Dialogue History” and “Talking Points”.</p> <p>{examples}</p> <p>Now, select the best talking point for the following pair:</p> <p># Dialogue History {context}</p> <p># Talking Points {inferences}</p> <p>Selection:</p>
R	<p>You are the Listener in a conversation shown in “Dialogue History”.</p> <p>Your goal is write a casual yet engaging and appropriate next response for the Listener (You) in the provided dialogue. First, sufficiently answer all questions posed by Speaker (Other) in their preceding turn. Then, continue your response by including the talking points shown in “Talking Points” since you want to cover them in your next response too.</p> <p>Write the response in the following format:</p> <p>Listener’s Response: ---</p> <p>Review the following examples to understand how to write a response given a “Dialogue History” and set of “Talking Points”.</p> <p>{examples}</p> <p>Now, complete the tasks for the following situation:</p> <p># Dialogue History {context}</p> <p># Talking Points {inferences}</p> <p>Listener’s Response:</p>

Table 5.15: The Inference Selection (S) and Response Generation (R) prompts used for few-shot in-context-learning. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated commonsense inferences, respectively. Placeholder “{examples}” is replaced by the few-shot examples in the specified format.

small, task-specific dataset risks overfitting and can lead to “catastrophic forgetting”, in which the general knowledge that was encoded in the LLM during pretraining is lost during the subsequent fine-tuning stage. This can lead to the model performing well on the training data but failing to generalize to unseen examples.

Instead of directly finetuning the original weights of an LLM, Hu et al. [30] introduce LoRA, a low rank adaptation strategy for finetuning. Under LoRA, the update to a weight matrix $W \in \mathbb{R}^{d \times k}$ of a layer in the LLM is approximated by a pair of smaller matrices A and B , where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$ such that $\Delta W = AB$. Following this, the modified forward pass of the affected computations is:

$$h = W_0x + \Delta Wx = W_0x + ABx \quad (5.1)$$

Only the low-rank decomposition matrices A and B are updated through backpropagation during fine-tuning. Such parameter-efficient fine-tuning affords benefit to few-shot finetuning. For one, because the learned matrices are of low-rank (i.e. much smaller size than the original weight matrices), this could lower the risk of overfitting that arises from over-parametrized and highly complex neural structures that can memorize all of the nuances in a small dataset at the expense of generalizability. In addition, because the original weights of the LLM are frozen, LoRA can balance the LLM’s pre-trained knowledge with the ability to learn task-specific behaviors. Specifically, this design can encourage the adaptation matrices to learn to perform the task-specific behaviors given the existing general knowledge contained in the LLM itself with less forgetting. In essence, LoRA can facilitate efficient adaptation of the LLM to a new task while lowering the risks of overfitting due to the limited dataset size.

Following this, I finetune an LLM to perform CSI-GSR using the created few-shot examples using LoRA, hereafter CSI-GSR-ft. There are 100 examples of the Inference Selection module and 100 examples of the Response Generation module, approximately uniformly distributed across all 10 commonsense types. The LLM is finetuned jointly to learn to perform both the Inference Selection and Response Generation modules independently, following the

Enumeration prompt design as the input format. At inference time, a two-call generation approach is performed, where CSI-GSR-ft is first provided the Inference Selection prompt with the dialogue context and commonsense options of the datapoint, and it outputs a single selected commonsense inference. Then, CSI-GSR-ft is provided the Response Generation prompt with the dialogue context and the previously outputted inference selection, and it generates a follow-up response.

Few-shot Impact on Responses Next, I perform the bias analysis from Section 5.3.3 on the Inference Selection module of the in-context-learning few-shot approach in order to determine whether the improvement strategy of adding few-shots has an impact on the previously observed selection bias. Table 5.16 shows that a more diverse spread of selected commonsense types is achieved, overcoming the bias towards selecting only three of the types. Indeed, previously **Subsequent**, **Desire_o**, and **Desire** accounted for 78.3% of the selections, whereas under the few-shot in-context-learning approach they account for only 23.3%. The strategy of incorporating few-shots, following a uniform distribution among the available types, has encouraged the Inference Selection module to incorporate a wider variety of types of commonsense for response generation. Although the Inference Selection module still displays a tendency to select certain types, manual examination reveals that the responses are less likely to bring up unnatural talking points for the dialogue than what was observed before; thus the few-shot strategy has made improvements to the CSI-GSR approach overall. Furthermore, I assess the rate of responses that are faithful to their selected commonsense, in order to verify that the Response Generation module is not detrimentally impacted by this modification to the Inference selection module. I find that 88% of the responses are faithful to their selected commonsense, incorporating its specified semantic content into the generated response. This is a comparable faithfulness rate to what is previously observed (90%), thus confirming that faithfulness has been upheld. Given the promising results of these analyses, a full evaluation study is undertaken next.

Type	%
Motivation	35.0
Reaction	13.3
Reaction _o	13.3
Desire _o	13.3
Attribute	11.7
Desire	6.7
Subsequent	3.3
Cause	3.3
Prerequisite	0.0
Constituent	0.0

Table 5.16: Percentage (%) of each commonsense type selected by the CSI-GSR-icl approach.

5.4 Evaluation

5.4.1 Models

This evaluation aims to measure the impact of CSI-GSR on dialogue response generation. As baselines, both large language models of ChatGPT and Llama2 are included. Based on the previous section on approach development, two variants of CSI-GSR-single are included for evaluation. Finally, the state-of-the-art in commonsense-grounded response generation is also included. Thus, the following models are evaluated:

ChatGPT is instructed to write an engaging and appropriate follow-up response for the provided dialogue context. The prompt used for response generation is shown in Table 5.7.

Llama2 is instructed to write the most plausible follow-up response for the provided dialogue context. This prompt is developed through manual experimentation and analysis, similar to the development process for ChatGPT-based prompts. Llama2 is included as an additional comparison of baseline response capability of current large language models as it is one of the best-performing open-sourced large language models. Specifically, the 7 billion chat version of Llama2¹ developed by Meta is utilized in this study, which has been

¹<https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

instruction-finetuned similar to ChatGPT. The prompt used for response generation is shown in Table 5.17.

Llama2
Dialogue History {context}
Write the most plausible next response from the Listener for the dialogue shown above. Only output the next response, no preamble:

Table 5.17: The prompt used for native response generation of Llama2. Placeholder “{context}” is replaced by the dialogue context of a provided example.

ChatGPT + CSI-GSR-icl is a model that follows the novel CSI-GSR-single approach to commonsense-augmented response generation using human-curated in-context-learning few shots for both the Inference Selection and Response Generation modules. It is described in detail in Section 5.3.4.

Llama2 + CSI-GSR-ft is a model that follows the novel CSI-GSR-single approach to commonsense-augmented response generation using few-shot finetuning on human-curated training data. Since ChatGPT is closed-source with limited accessibility for finetuning, Llama2 is used as the base model to support LoRA few-shot finetuning of this approach. It is described in detail in Section 5.3.4.

ChatGPT + Doctor is the state-of-the-art commonsense-augmented open-domain dialogue model [8]. It was previously shown to outperform native ChatGPT3.5 with statistical significance in terms of response naturalness and specificity, and to outperform models that rely on other dialogue commonsense resources, including ComFact, Reflect, and Cicero. Following this approach, ChatGPT3.5 is conditioned to produce the most plausible next response to a given dialogue context by leveraging a provided commonsense rationale, where appropriate. The commonsense rationale is composed of a sequence of 3 question-answer (QA) pairs that, when taken together, aim to interpret the dialogue context and reason about

the most likely follow-up. At inference time, a trained rationale model is used to generate the rationale for a given dialogue example, which is then inserted into the ChatGPT prompt before sending the prompt to ChatGPT. The official trained rationale model² is used, which is an OPT-1.3B model finetuned on the official dataset of commonsense rationales filtered from ChatGPT. The prompt used for response generation is taken from the original paper and is shown in Table 5.18.

ChatGPT + Doctor
<p>Generate the most plausible next response considering the dialogue history. You can refer to the rationale, but you should ignore the rationale if it misleads the next response. Do not try to put too much information in the next response. You should follow the style of the history.</p> <p>Rationale: {inferences} History: {context} Next Response: A</p>

Table 5.18: The prompt used for state-of-the-art commonsense-augmented open-domain dialogue model, ChatGPT + Doctor. Placeholders “{context}” and “{inferences}” are replaced by the dialogue context and generated rationale, respectively, of a provided example. The turns in the context are prefixed in alternating fashion by nominal tags “A” and “B”, where the last turn in the context is prefixed by “B”.

For generation using ChatGPT-based models (ChatGPT, ChatGPT + CSI-GSR-icl, ChatGPT + Doctor), the latest snapshot at the time of writing is used (`gpt-3.5-turbo-0125`) with a temperature of 0.7. For generation using Llama-based models (Llama2, Llama2 + CSI-GSR-ft), beam search is used with 3 beams. In addition, for Llama+CSI-GSR-ft, training is done with a learning rate of 0.0001, weight decay of 0.001, batch size of 16, and max input length of 1024 for 10 epochs. The LoRA is configured with rank of 32, alpha of 64, and dropout of 0.05. This training configurations is selected based on manual analysis of results on the development set across various configurations to identify the best outputs. Quantization

²<https://huggingface.co/DLI-Lab/DOCTOR>

to `nf4` is used based on manual analysis that finds it improves efficiency while achieving comparable outputs to non-quantized setups.

5.4.2 Test Data

During approach development, it is observed that ChatGPT substantially outperformed CSI-GSR variants. However, it should be noted that the development data is sampled from a dataset of dialogues that are actually generated by ChatGPT in the first place, namely, the SODA dataset. This introduces the possible effect that ChatGPT only performed so well because the dialogues it is being evaluated on are “in-distribution” and represented dialogue situations ChatGPT is capable of handling, thus inflating its performance against the alternative approaches. To conduct a fairer evaluation of the approaches under study, the test data is specifically chosen to be “out-of-distribution” for all models. As such, dialogues are sampled from the Reflect dataset [111]. The Reflect dataset is composed of 600 dialogues from the Commonsense-Dialogues dataset [110]. The dialogues in Commonsense-Dialogues are human-written, where a human annotator is shown a short natural language narrative describing an everyday situation, derived from Atomic commonsense tuples, and asked to write a dialogue between two people that is appropriate for the narrative. The Reflect dataset further modifies the dialogues to terminate at the utterance with the highest cosine similarity to the narrative used as a prompt for each dialogue. 60 dialogues are sampled from Reflect to be used as test data. Each model generates a follow-up response for each test dialogue, resulting in 5 responses to be evaluated for each test dialogue. Table 5.19 presents examples of the generated commonsense and response outputs for each model under study.

5.4.3 Metrics

Three undergraduate students at Emory University are recruited to be human annotators. During evaluation, the human annotators are shown a dialogue context along a response from one of the models. The order of the examples is randomized for each annotator and

Dialogue Context	
<p>Speaker (Other)/B: This story of mine’s been getting a lot of attention on the internet! Listener (You)/A: How many followers do you have now? Speaker (Other)/B: A lot! People are really enjoying reading it.</p>	
ChatGPT	
<p><i>Response:</i> That’s awesome! It must feel great to see your hard work paying off like this.</p>	<p><i>Commonsense:</i> -</p>
ChatGPT+Doctor	
<p><i>Response:</i> Wow, that’s amazing! I’m so happy for you.</p>	<p><i>Commonsense:</i> Subquestion 1: What is the reason for Person A to ask about the number of followers? (xIntent) Subanswer 1: Person A wants to know how many followers Person B has on the internet. Subquestion 2: What is the reaction of Person B to the question about the number of followers? (oReact) Subanswer 2: Person B is surprised by the question and didn’t expect that many people would be interested in their blog. Subquestion 3: What might Person A want to communicate to Person B, given Person B’s surprise at the number of followers? (oWant) Subanswer 3: Person A might want to express their surprise and admiration for Person B’s blog and their ability to attract a large number of followers.</p>
ChatGPT+CSI-GSR-icl	
<p><i>Response:</i> I’m glad to hear that your story is gaining traction! What’s it about? I’m really curious to know more about it.</p>	<p><i>Commonsense:</i> The Listener (You) feels curious about the story and wants to know more about it.</p>
Llama2	
<p><i>Response:</i> That’s great to hear! It’s always exciting to see a story gain traction online. Have you considered sharing it on social media platforms to reach an even wider audience?</p>	<p><i>Commonsense:</i> -</p>
Llama2+CSI-GSR-ft	
<p><i>Response:</i> Wow, that’s awesome! I guess this is no surprise since it is a really interesting story. Do you have any other writing that you want to share with the world?</p>	<p><i>Commonsense:</i> The Speaker (Other) wants to use this success as a platform to promote their other works and gain more exposure.</p>

Table 5.19: Example commonsense outputs, where applicable, and response outputs for the models under study.

annotators are unaware of which model produced which response. 5% of the evaluation data is annotated by all three annotators to calculate inter-annotator agreement.

For each example, the human annotators complete assessments on five different metrics. The metrics are assessed independently of one another. Three behavior metrics are included, targeting specific behaviors that are hypothesized to be improved by commonsense integration for dialogue models. In addition, for overall response quality and informativeness achieved by each model, Likert-style metrics are used to capture these high-level constructs that do not decompose cleanly into binary behaviors.

Quality is measured by a Likert scale on the turn-level, following Qua_t from Chapter 3. The same setup and instructions of Qua_t are used (Section 3.3), in which annotators are instructed to rate the quality of the response given its dialogue context on a scale from 1 (least) to 5 (most). No training is provided due to the subjective nature of the metric.

Informativeness is measured by a Likert scale on the turn-level, following Inf_t from Chapter 3. The same setup and instructions of Inf_t are used (Section 3.3), in which annotators are instructed to rate the level of unique and specific information of the response to its dialogue context on a scale from 1 (least) to 5 (most). No training is provided due to the subjective nature of the metric.

Irrelevant is measured by a binary behavior label on the turn-level, following $!Rel$ in ABC-Eval. The setup and instructions of $!Rel$ from ABC-Eval are used, in which annotators are instructed to indicate whether the response is appropriately relevant to its dialogue context. Annotators complete the $!Rel$ training from ABC-Eval before performing this evaluation task, which consists of 3 graded quizzes on the $!Rel$ label with automatic feedback provided for their incorrect labels.

Commonsense Contradictions is measured by a binary behavior label on the turn-level, following *!Com* in ABC-Eval. The setup and instructions of *!Com* from ABC-Eval are used, in which annotators are instructed to indicate whether the response violates commonsense knowledge. Annotators complete the *!Com* training from ABC-Eval before performing this evaluation task, which consists of 3 graded quizzes on the *!Com* label with automatic feedback provided for their incorrect labels.

Conversational Expectations is measured by a binary behavior label on the turn-level, hereby denoted as *!Exp*. This is a new behavior label, which is not present in the ABC-Eval evaluation framework. *!Exp* captures instances where the response fails to meet conversational expectations, such as by focusing on a low salience topic or continuing the conversation in some way that feels out of place. This behavior metric is included based on the prevalence of the error **Suboptimal Follow-up Topics** in the conducted error analysis of CSI-GSR (Section 5.3.2). Following the development of the ABC-Eval framework, a set of instructions and examples are created to describe the Conversational Expectations label. These examples are taken from the responses in the Error Analysis that are categorized as **Suboptimal Follow-up Topics**. Tables 5.20 and 5.21 provide the instructions and examples, respectively, given to annotators for this evaluation label. In addition, 3 training quizzes with automatic feedback are developed to teach annotators how to perform the label of *!Exp*. Positive and negative examples of *!Exp* are taken from the responses in the Error Analysis, excluding those responses already used as examples for the instructions. Each training quiz contains 5 positive and 5 negative examples of *!Exp*. Annotators complete the *!Exp* training before performing this evaluation task.

Instructions

Goal:

Identify if each response option for Speaker A fails to follow conversational expectations/norms.

Details:

In a conversation, there are usually expected topics or actions that follow. If Speaker A's response introduces a topic or action that is unexpected or unlikely based on the conversation so far, it may feel weird or wrong.

This is different from relevance. Even if a response is relevant to the topic, it can still fail to follow expectations/norms if it focuses on something unimportant or unlikely given the context.

One sign of a response that exhibits expectation failure is if there is an obvious preferred action or follow-up point that always fits the conversational situation.

Consider the following for each turn:

Does Speaker A's response fail to follow conversational expectations/norms because it focuses on an event or action that is unlikely or unimportant for the current discussion?

* If Yes, the response focuses on something unlikely or unimportant, or there is an obvious alternative response that better fits the conversational situation.

* If No, the response focuses on the most important and expected follow-up to the current discussion.

Table 5.20: Instructions for the Conversational Expectations evaluation label.

Examples

<p>B: Oh, please. You're the one who's always starting arguments with me.</p> <p>A: No, I'm not. You're the one who's always trying to start something.</p> <p>B: That's because you're always doing something that deserves to be called out.</p> <p>A: Like what?</p> <p>B: Well, like the way you're always putting yourself first and expecting everyone else to just deal with it.</p> <p>A: That's not true. I don't do that.</p> <p>B: Yes, you do. And it's really annoying.</p> <p>A: I'm sorry if it seems like I'm always starting arguments with you. That's not my intention at all. Maybe we just have different perspectives on things. Have you considered seeking advice from a trusted friend or counselor? They might be able to help you navigate this situation and find a resolution that works for both of us. It could be beneficial to have an outside look.</p>	<p><i>Label: 1</i></p> <p>Yes, the response focuses on something unlikely or unimportant, or there is an obvious alternative response that better fits the conversational situation:</p> <p>A and B are having an argument about how they treat one another. Although suggesting to talk about this issue with an external person is relevant, the way that A proposes this solution is not expected since they remove their own participation from the suggested action. This is an example of an unlikely action for A to take and it is easy to think of a different manner in which this could have played out that would be more expected (e.g. A asking B to come with them to talk about their fight with an external mediator).</p>
<p>A: Here, you dropped your pencil. Whatcha doing?</p> <p>B: Just some homework for my math class. It's due tomorrow and I'm still struggling with it.</p> <p>A: Let me take a look. I'm pretty good at math and maybe I can help you understand it better.</p> <p>B: That would be great! Thank you so much.</p> <p>A: You're welcome, John. I'm happy to help. Before we dive into the math homework, though, how about we take a little break and grab a snack? I find that taking a break and refueling with some food can help clear our minds and make it easier to focus. Plus, it's always nice to have a little snack to keep us going.</p>	<p><i>Label: 1</i></p> <p>Yes, the response focuses on something unlikely or unimportant, or there is an obvious alternative response that better fits the conversational situation:</p> <p>A and B just started a conversation, in which A offers to help B with their homework and proposes a snack break first. Although A proposes this activity in such a way that justifies its relevance to the current discussion, it is an unlikely event for A to immediately ask B if they want to take a break and grab a snack, before A helps B at all. This response is a deviation from the expected trajectory of this dialogue.</p>
<p>A: Hey coach, thanks for helping me out these past few months. I really appreciate it.</p> <p>B: No problem, Yusuf. You've worked hard and it's paid off. Your tennis skills have improved a lot.</p> <p>A: Yeah, I've been practicing a lot lately. But I still feel like there's room for improvement. What do you think I should work on next?</p> <p>B: Well, one thing you could work on is your backhand stroke. It's not as strong as your forehand stroke and needs some improvement.</p> <p>A: Thanks, coach. I really appreciate your guidance and support. I've been putting in a lot of effort, but I agree that there's always room for improvement. I'll definitely focus on my backhand stroke. Do you have any specific drills or exercises that you recommend to help me strengthen it?</p>	<p><i>Label: 0</i></p> <p>No, the response focuses on the most important and expected follow-up to the current discussion: A is talking to their coach B about their tennis skills, in which B just pointed out that A's backhand stroke could be improved. A acknowledges the suggestion and follows up by asking for specific exercises/drills to improve it, which is a likely action to take next. It is difficult to think of a response that would be more suitable than that.</p>

Table 5.21: Examples given to the annotators for the Conversational Expectations evaluation label. The correct label and explanation is shown in the right-hand column for the final “A” response of the dialogue context shown in the left-hand column.

5.4.4 Metric Reliability

Interannotator agreement is calculated using Krippendorff’s alpha for each metric. Annotator agreement and reliability results are shown in Table 5.22. It is seen that the human annotators produce similar agreement to that achieved during the original ABC-Eval study, achieving moderate agreement for most metrics. Although the agreement on Qua_t is low, this is expected due to the highly subjective nature of rating the quality of a dialogue and is seen in other studies, including ABC-Eval. In fact, ABC-Eval verifies that Qua_t is capable of detecting significant differences between different dialogue models even with a low interannotator agreement on the turn annotations, thus supporting the usage of Qua_t for this study despite the low interannotator agreement. In addition, interannotator reliability on annotating the behaviors ($!Com$, $!Rel$, and $!Exp$) is also confirmed from the annotators’ performance on the training quizzes, reaching nearly 80% accuracy on average for the training quizzes for each behavior metric.

Metric	Krippendorff’s α	Training %			Training Avg.
Quality	0.21	–	–	–	–
Informativeness	0.53	–	–	–	–
Irrelevant	0.56	69	88	83	80
Commonsense Contradictions	0.46	76	81	74	77
Conversational Expectations	0.41	73	77	77	76

Table 5.22: Measure of annotator reliability for each metric under study. The inter-annotator agreement (**Krippendorff’s α**), average performance on each training quiz (**Training %**), and average overall training performance (**Training Avg.**) are shown. Averages are calculated across all annotators and **Training %** is formatted as *quiz #1 | quiz #2 | quiz #3*.

5.4.5 Results

Informativeness It can be seen in Figure 5.3 that the CSI-GSR approach to response generation successfully and substantially improves the amount of specific detail that is expressed in the generated dialogue responses, for both few-shot strategies explored in this study. Indeed, both ChatGPT+CSI-GSR-icl and Llama2+CSI-GSR-ft demonstrate statistically

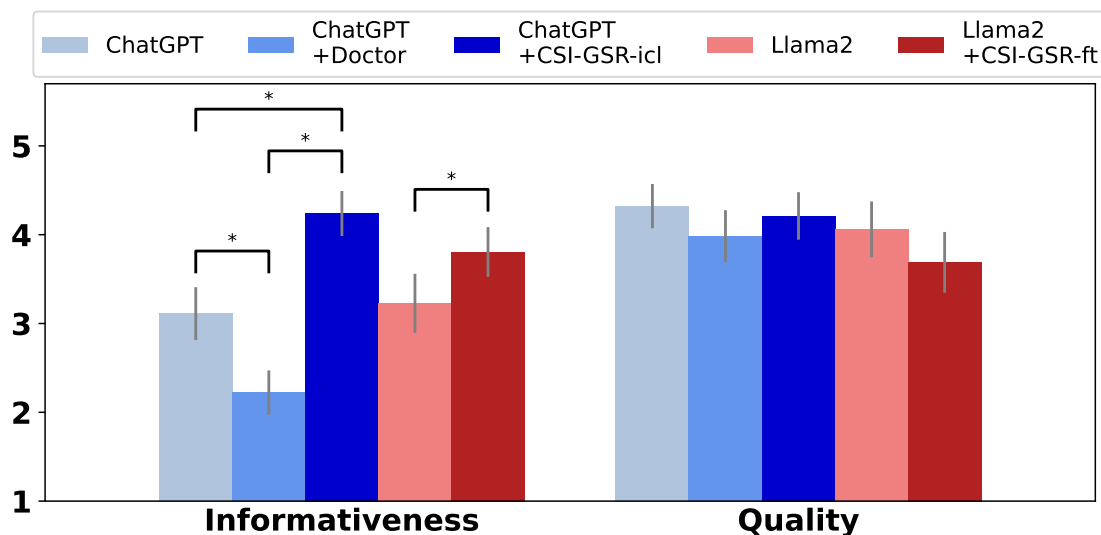


Figure 5.3: Average scores achieved by each dialogue model on the Likert-style metrics of Informativeness and Quality, where **higher is better**. Error bars denote 95% Student’s t confidence intervals. * denotes statistically significant differences (paired t-test, $\alpha < 0.01$).

significant improvements in Informativeness over their counterparts using the same base model. Furthermore, it can be seen that ChatGPT+CSI-GSR-icl model vastly outperforms the SOTA commonsense dialogue model ChatGPT+Doctor on the Informativeness of the responses. Consequently, the results shown for CSI-GSR demonstrate the power of this approach to substantially reduce the occurrence of generic or non-contextualized responses, far outperforming the capabilities of the previous state-of-the-art in commonsense-augmented response generation.

The observed results on the low informativeness of ChatGPT+Doctor responses runs contrary to the original results from Chae et al. [8], which showed that ChatGPT+Doctor improved the specificity (i.e. informativeness) of responses against native ChatGPT responses. Examination of a sample of responses from the native ChatGPT prompt used by Chae et al. [8]³ suggest that this prompt resulted in generic responses with little detail. On the other hand, the current study uses a prompt for native ChatGPT response generation that does not exhibit this same response genericness, which is accomplished through targeted effort spent

³It should be noted that Chae et al. [8] did not provide their exact native ChatGPT prompt; based on the details in the paper, I take the prompt from Chae et al. [8] and remove the rationale component to stand in as the native ChatGPT prompt used in that work.

optimizing native ChatGPT to produce high quality responses. Consequently, the ChatGPT baseline used by Chae et al. [8] seems to be an underestimate of the power of native ChatGPT at response generation, and the prompt used in the current study is a fairer representation of the capabilities of ChatGPT, thereby serving as a more appropriate and stronger baseline against which to measure improvements in specificity. Table 5.23 shows examples of responses produced by native ChatGPT from the current study, native ChatGPT from Chae et al. [8], and SOTA ChatGPT+Doctor, which illustrates the discrepancy in baseline genericness.

Dialogue	Response
B: I want to take a nap but I am so hungry. A: Why don't you eat then nap. B: I just got back from the museum with the kids. I think we are all going to order hamburgers. We are starving.	[<i>ChatGPT</i>] That sounds like a perfect post-museum meal! Enjoy those burgers and I hope you all get a good nap afterwards. [<i>ChatGPT [8]</i>] Sounds like a good plan. Enjoy your burgers and nap after! [<i>Doctor [8]</i>] Sounds like a good plan. Enjoy your meal!
B: I wish I didn't have to always use the bus because I never seem to be on time. A: I wonder why your routes are late, I don't have that problem. B: I'm not sure! I waited an entire hour after work to get a ride home.	[<i>ChatGPT</i>] That's frustrating, waiting for an hour must have been really annoying. Have you tried reaching out to the bus company about the delays? [<i>ChatGPT [8]</i>] That sounds frustrating, maybe there was an accident or construction causing delays. [<i>Doctor [8]</i>] Have you thought about trying to carpool with someone instead of taking the bus?

Table 5.23: Examples highlighting discrepancies in informativeness between responses based on Chae et al. [8].

Quality For the open-domain dialogue setting explored in this study, the human-perceived Quality of the responses is not greatly affected by the incorporation of commonsense (Figure 5.3). Instead, each of the commonsense-augmented models achieves similar response Quality to that of their native counterparts.

Irrelevant Responses Most models are quite capable of producing relevant responses, with the rate of irrelevance responses at or below 5% (Figure 5.4). It is suggested that

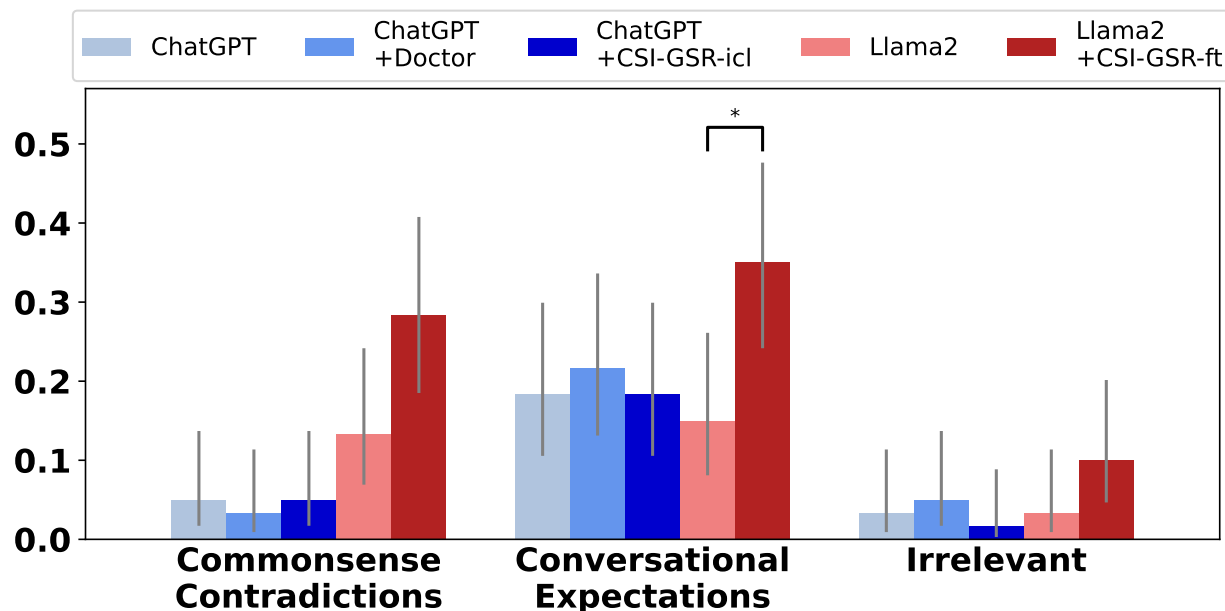


Figure 5.4: Average proportions of turns of each dialogue model expressing the behaviors of Commonsense Contradictions, violations of Conversational Expectations, and Irrelevant, where **lower is better**. Error bars denote 95% Wilson score confidence intervals. * denotes statistically significant differences (McNemar’s test, $\alpha < 0.01$).

Llama2+CSI-GSR-ft struggles with response relevance more, although the difference is not statistically significant. Consequently, incorporating commonsense into the responses seems to have little effect on the overall relevance of the response.

Commonsense Contradictions As shown in Figure 5.4, models based on ChatGPT display comparable rates of Commonsense Contradictions, falling below 5%, with no significant differences observed. Models based on Llama2 perform somewhat worse and seem to produce greater rates of commonsense contradictions. Even so, these results indicate that commonsense contradictions are less of an issue for current generation large language models compared to previous neural dialogue models (Chapter 3). The rates observed for ChatGPT and Llama2 models are a marked improvement over previous dialogue models which outputted Commonsense Contradictions at a rate of approximately 20%. This suggests that recent advancements in large language modeling seem to have mitigated this commonsense challenge without explicit training on dialogue response generation or specialized techniques

for enhancing commonsense understanding. Since LLMs like Llama2 and ChatGPT are becoming the commonplace base model to use for dialogue applications, the direction of correcting commonsense errors specifically for dialogue seems to have diminishing necessity.

Conversational Expectations Out of all of the measured behaviors, the models are most vulnerable to producing violations of conversational expectations in their responses, with Llama2+CSI-GSR-ft expressing the greatest rate of violations (Figure 5.4). For the other commonsense-based approaches, their rate of production does not differ much from their native counterparts, suggesting that current approaches to incorporating commonsense into responses neither helps nor hinders this behavior. Previously unexplored, this new behavior label seeks to identify instances where model responses are relevant to the current discussion, but focus on an aspect of the topic under discussion that is less salient or important than other aspects that come to mind. These results suggest that further work is needed towards improving the capability of dialogue models to conform to conversational expectations and to identify the most compelling follow-up topic. Examples in Table 5.24 underscore instances where responses deviated from discussing the important aspects of the current discussion.

Overall, Llama2+CSI-GSR-ft fares the worst on all metrics out of all approaches under study. This observation suggests that the limited volume of few-shot training data utilized in this investigation may not adequately support the effective mastery of the CSI-GSR task. Conversely, the CSI-GSR-icl methodology demonstrates notable efficacy, underscoring the value of the introduced commonsense selection and integration strategy. The evaluation outcomes, in conjunction with previously noted high levels of response faithfulness (Section 5.3.4), endorse CSI-GSR-icl’s proficiency in seamlessly incorporating rich and contextually pertinent commonsense information within dialogue responses. This enhancement distinctly bolsters the specificity of interactions relative to the given context. Importantly, this is achieved without compromising response quality or coherence of the overall dialogue flow.

Dialogue	Response	Explanation
B: You know that I have a huge crush on Kendall. I got her to go to the movies with me.	A: That’s awesome! How was the movie? source: ChatGPT	The salient topic is the interpersonal relationship between B and Kendall, not details about the movie.
B: I couldn’t hear anything in here this morning. A: I know. I asked your brother to turn the music down but he didn’t want to. B: I know, when he does that you can only try talking louder.	A: I guess we’ll just have to invest in some noise-canceling headphones. source: ChatGPT	The goal of the discussion is how to hear better given the large amount of noise from the brother, whereas noise-canceling headphones would block all sound.
B: Kendall just rubbed my chin. A: Why did he do that? B: He said it is a good luck but I find it weird.	A: He probably meant well, but I can see why you found it weird. It’s kind of a quirky thing to do for good luck. So, how do you usually deal with situations that catch you off guard like that? source: ChatGPT+CSI-GSR-icl	The phrasing of the question is too analytical and therapeutical, which is out-of-place in the casual nature of the dialogue.

Table 5.24: Examples of responses that are positive examples of violations of Conversational Expectations, along with explanations of the violation.

5.5 Conclusion

This study introduced the Commonsense Inference Generate-Select-Respond (CSI-GSR) model, a novel approach to enhancing the capabilities of large language models (LLMs) in generating more informative and context-specific dialogue responses. By harnessing a diverse pool of commonsense inferences, CSI-GSR not only surpasses existing models, including ChatGPT3.5, Llama2, and ChatGPT + Doctor, in producing responses with greater detail and specificity to the dialogue context but also upholds the natural flow and coherence of conversation. Moreover, the insights gained from this research indicate that commonsense errors are less common in the current era of large language models for dialogue applications

and point toward an imperative future direction in LLM-based dialogue development: the refinement of models to better discern and emphasize the most pertinent aspects of a conversation, thereby generating follow-up responses that not engage on the most important discussion topics.

Chapter 6

Conclusion

6.1 Research Contributions

This dissertation focuses on the pivotal role of social commonsense knowledge in enriching open-domain dialogue, with a focus on social commonsense that is multi-faceted (covering a diverse array of commonsense types) and contextually novel (produces new information that is predictive in nature to the dialogue context). The findings of the work in this dissertation provide a verified framework for measuring commonsense capability of dialogue models using a behavior-based evaluation design, present the first successful methods for obtaining multi-faceted and contextually novel social commonsense for dialogue, and demonstrate that incorporating such social commonsense into dialogue responses significantly enhances their specificity to their dialogue contexts. In summary, the findings of the three covered research questions are as follows.

Does a binary behavior-based evaluation framework result in more reliable and informative evaluation of dialogue model characteristics?

Thorough statistical analyses demonstrate that the behavior labels covered in the Annotation of Behaviors in Chat Evaluation (ABC-Eval) are more interpretable, predictive of overall

dialogue quality, and capable of identifying differences between dialogue models than existing dialogue characteristic evaluation methods. Thus, these results confirm that a binary behavior-based evaluation framework does result in more reliable and informative fine-grained dialogue model evaluations.

Is it possible to reliably obtain multi-faceted contextually novel social commonsense inferences for dialogue?

Generative commonsense inference models trained on ConvoSense are shown to produce highly reasonable, detailed, and contextually novel commonsense inferences for dialogue. This is observed for ten different commonsense types across a variety of dialogue situations, thus verifying that it is possible to reliably obtain multi-faceted and contextually novel social commonsense inferences for dialogue.

Does access to a multi-faceted and contextually novel social commonsense pool improve dialogue responses?

The Commonsense Inference Generate-Select-Respond (CSI-GSR) dialogue approach achieves substantial improvement in response specificity and informativeness, while maintaining the high response quality observed from LLMs, thereby surpassing the performance of the state-of-the-art commonsense-augmented dialogue model. These results demonstrate the benefits of providing dialogue models with a multi-faceted and contextually novel commonsense pool.

6.2 Future Work

Building upon the work presented in this dissertation, the following avenues of future work are identified.

6.2.1 Behavioral Evaluation of LLM-based Dialogue Models

The ABC-Eval framework, detailed in Chapter 3, has been tested and applied to a previous generation of neural dialogue models, which have now been superseded by LLM-based dialogue models. The findings outlined in Chapter 5 suggest that the behavioral patterns observed in previous dialogue models may not directly apply to LLM-based models, as exemplified by the change in frequency of expressing commonsense contradictions. Therefore, it is crucial to conduct thorough behavioral evaluations on LLM-based dialogue models to gain a nuanced understanding of their current limitations, rather than relying on insights from outdated models. Moreover, there are likely aspects crucial for assessing the performance of LLM-based dialogue models that are not covered by ABC-Eval, such as the analysis of conversational expectation violations as demonstrated in Chapter 5. Adapting ABC-Eval to capture the state of LLMs represents a valuable research endeavor, as it can pinpoint the weaknesses of these models that will help to direct and propel dialogue model advancement in the most efficient manner.

6.2.2 Commonsense-augmented Dialogue Models “In-the-Wild”

Most previous works on commonsense-augmented dialogue models, including that described in this dissertation in Chapter 5, rely on evaluations of single responses outputted to static dialogue contexts. This is due to the high cost of “in-the-wild” evaluation, where such dialogue models would be deployed in an interactive multi-turn setting with human users to test the capabilities of the systems in full conversations. However, these “in-the-wild” evaluations are often better indicators of the true performance of dialogue systems as they represent the realistic deployment setting. It is imperative that future work on commonsense-augmented dialogue models include “in-the-wild” evaluations in order to gain a better understanding of their real-world applicability.

6.2.3 Improving Social Commonsense Integration for Open-Domain Dialogue

Certain design choices of CSI-GSR dialogue approach in Chapter 5 are likely to benefit from further experimentation, especially given the observed limitations of the current implementation. For one, strategies to mitigate the observed trade-off between response long-windedness/coherence and inference diversity can serve to make multi-inference augmentation more successful, which faced much difficulty in the work in this dissertation, perhaps through alternative implementations of the Inference Generation module. In addition, as discussed in Section 5.3.3, there can be multiple ways of expressing the same commonsense inference content, which may have an effect on the overall quality or coherence of the response in its context. The work in this dissertation did not experiment with different presentation styles or strategies, which leaves room for future work to build upon. Finally, as discussed in Section 4.3.1, the work in this dissertation focused on commonsense inferences for dialogue situations that are rich in information-sharing from the human user. Utterances which focus less on information-sharing (e.g. acknowledgment, question-asking, etc.) may have different requirements or integrations with commonsense reasoning, which should be examined in future work.

6.2.4 Alternative Commonsense Integrations

Depending on the dialogue application, the role of commonsense could differ. Whereas open-domain dialogue tends to involve discussions on life experiences and events thus promoting the need for social commonsense, other dialogue applications may benefit from other types of commonsense, such as that related to frame semantics or physical properties of objects. Further research can extend into additional commonsense categories and also into additional settings of dialogue, beyond the open-domain setting.

In addition, there are other possible strategies to integrating commonsense that may

be more suitable for other dialogue applications. This dissertation found success for open-domain dialogue models when treating commonsense inferences as response content candidates, but other applications could benefit from a more structured approach. Further research investigating the combination of commonsense and structured text representations for dialogue, such as dialogue acts/intents or rhetorical structure theory [53], are worth exploring, especially for dialogue applications that are more goal-oriented in nature, such as virtual assistants, tutors, or persuasive systems.

6.2.5 Improving Open-source LLMs with Commonsense

Future efforts should include extending commonsense-augmentation to open-source LLMs, like Llama2, by employing more robust datasets compared to the few-shot methodology utilized in this study. Enhancing these models with commonsense knowledge could reduce the observed error rates and increase the informativeness of responses, making them more competitive with models like ChatGPT.

6.2.6 Enhancing Topical Saliency in Dialogue Models

There is a promising avenue for enhancing Large Language Model (LLM)-based dialogue systems by improving their ability to recognize and focus on the most compelling follow-up points in a conversation. By refining the models to identify and prioritize key aspects of the dialogue, these systems can become more effective in maintaining the flow of conversation and adhering to human conversational patterns. This improvement is essential as it addresses the observed tendency of LLM-based models to focus on less critical details (Section 5.4.5).

6.2.7 Refining Response Content Control

Many dialogue applications can benefit from the ability to integrate specific information into the system's responses. Achieving this high degree of control over the response content can

conflict with ensuring the response flows naturally and coherently within the current dialogue. The successful application of the CSI-GSR method for commonsense integration in this study suggests that leveraging LLMs' advanced language capabilities could be a promising approach to achieve better control over response content. Future efforts in this direction could explore the general capability of LLM-based dialogue models for incorporating specific details into responses without disrupting the conversational continuity.

Bibliography

- [1] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.
- [2] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998.
- [3] Hedin Beattie, Lanier Watkins, William H. Robinson, Aviel Rubin, and Shari Watkins. Measuring and Mitigating Bias in AI-Chatbots. In *2022 IEEE International Conference on Assured Autonomy (ICAA)*, pages 117–123, March 2022. doi: 10.1109/ICAA52185.2022.00023.
- [4] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, 2019.
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

- [6] Rainer E Burkard and Eranda Cela. Linear assignment problems and extensions. In *Handbook of combinatorial optimization: Supplement volume A*, pages 75–149. Springer, 1999. URL https://link.springer.com/chapter/10.1007/978-1-4757-3023-4_2.
- [7] Yue Cao and Xiaojun Wan. Divgan: towards diverse paraphrase generation via diversified generative adversarial network. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2411–2421, 2020. URL <https://aclanthology.org/2020.findings-emnlp.218/>.
- [8] Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. Dialogue Chain-of-Thought Distillation for Commonsense-aware Conversational Agents. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.342. URL <https://aclanthology.org/2023.emnlp-main.342>.
- [9] Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. Places: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 814–838, 2023.
- [10] HH Clark. Grounding in communication. *Perspectives on Socially Shared Cognition*, pages 127–149, 1991. doi: <https://doi.org/10.1037/10096-006>.
- [11] Yao Dou, Maxwell Forbes, Ari Holtzman, and Yejin Choi. Multitalk: A highly-branching dialog testbed for diverse conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12760–12767, 2021. doi: <https://doi.org/10.1609/aaai.v35i14.17510>.

- [12] Christiane Fellbaum. Wordnet and wordnets. In Alex Barber, editor, *Encyclopedia of Language and Linguistics*, pages 2–665. Elsevier, 2005.
- [13] Sarah Fillwock and David Traum. Identification of personal information shared in chat-oriented dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018. URL <https://aclanthology.org/L18-1629.pdf>.
- [14] Sarah E Finch and Jinho D Choi. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 236, 2020.
- [15] Sarah E. Finch and Jinho D. Choi. ConvoSense: Overcoming monotonous commonsense inferences for conversational AI. *Transactions of the Association for Computational Linguistics*, 2024.
- [16] Sarah E. Finch, James D. Finch, Ali Ahmadvand, Ingyu, Choi, Xiangjue Dong, Ruixiang Qi, Harshita Sahijwani, Sergey Volokhin, Zihan Wang, Zihao Wang, and Jinho D. Choi. Emora: An Inquisitive Social Chatbot Who Cares For You, September 2020. URL <https://arxiv.org/abs/2009.04617v1>.
- [17] Sarah E. Finch, James D. Finch, and Jinho D. Choi. 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, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.839. URL <https://aclanthology.org/2023.acl-long.839>.
- [18] Yahui Fu, Koji Inoue, Chenhui Chu, and Tatsuya Kawahara. Reasoning before responding: Integrating commonsense-based causality explanation for empathetic response

- generation. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 645–656, 2023.
- [19] Silin Gao, Jena D. Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 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, December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.findings-emnlp.120>.
- [20] Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 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, 2021.
- [21] Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 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, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.344. URL <https://aclanthology.org/2022.acl-long.344>.
- [22] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowdworkers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.
- [23] Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. A systematic exploration of diversity in machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111, 2013. URL <https://aclanthology.org/D13-1111.pdf>.
- [24] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-

- Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Interspeech 2019*, pages 1891–1895. ISCA, September 2019. doi: 10.21437/Interspeech.2019-3079. URL https://www.isca-speech.org/archive/interspeech_2019/gopalakrishnan19_interspeech.html.
- [25] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022. URL <https://arxiv.org/abs/2203.05794>.
- [26] Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. DialFact: A Benchmark for Fact-Checking in Dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.263. URL <https://aclanthology.org/2022.acl-long.263>.
- [27] KL Gwet. Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-Rater Reliability Assessment*, 1, 2002. URL https://www.agreestat.com/papers/kappa_statistic_is_not_satisfactory.pdf.
- [28] Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 89–98, Singapore and Online, July 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.sigdial-1.10>.
- [29] Or Honovich, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. Q2: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and

- Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.619. URL <https://aclanthology.org/2021.emnlp-main.619>.
- [30] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
- [31] Yu-Cian Huang, Edwinn Gamborino, Yan-Jia Huang, Xiaobei Qian, Li-Chen Fu, and Su-Ling Yeh. Inferring stressors from conversation: Towards an emotional support robot companion. *International Journal of Social Robotics*, 14(7):1657–1671, 2022.
- [32] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6384–6392, 2021. URL <https://ojs.aaai.org/index.php/AAAI/article/view/16792>.
- [33] Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 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, 2019. URL <https://aclanthology.org/P19-1365/>.
- [34] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. 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, 2013. URL <https://ieeexplore.ieee.org/abstract/document/6681438>.
- [35] Di Jin, Sijia Liu, Yang Liu, and Dilek Hakkani-Tur. 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, 2022. URL <https://aclanthology.org/2022.sigdial-1.56/>.
- [36] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021.
- [37] Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. Soda: Million-scale dialogue distillation with social commonsense contextualization, 2022.
- [38] Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. Soda: Million-scale dialogue distillation with social commonsense contextualization. *ArXiv*, abs/2212.10465, 2022.
- [39] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*, 2023.
- [40] Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. Prompted llms as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.
- [41] Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683, 2022.
- [42] Margaret Li, Jason Weston, and Stephen Roller. ACUTE-EVAL: Improved Dialogue

- Evaluation with Optimized Questions and Multi-turn Comparisons, September 2019. URL <https://arxiv.org/abs/1909.03087v1>.
- [43] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10993–11001, 2022.
- [44] Siheng Li, Wangjie Jiang, Pengda Si, Cheng Yang, Qiu Yao, Jinchao Zhang, Jie Zhou, and Yujiu Yang. Enhancing dialogue generation with conversational concept flows. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1514–1525, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.112. URL <https://aclanthology.org/2023.findings-eacl.112>.
- [45] Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. Addressing Inquiries about History: An Efficient and Practical Framework for Evaluating Open-domain Chatbot Consistency. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1057–1067, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.91. URL <https://aclanthology.org/2021.findings-acl.91>.
- [46] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.
- [47] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.
- [48] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong

- Jiang, and Minlie Huang. Towards emotional support dialog systems. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.269. URL <https://aclanthology.org/2021.acl-long.269>.
- [49] Yiren Liu and Halil Kilicoglu. Commonsense-aware prompting for controllable empathetic dialogue generation. In *Proceedings of the Workshop on Knowledge Augmented Methods for Natural Language Processing*, 2023.
- [50] Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*, 2023.
- [51] Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*, 2021.
- [52] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. MIME: MIMicking Emotions for Empathetic Response Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.721. URL <https://aclanthology.org/2020.emnlp-main.721>.
- [53] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.
- [54] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based

- clustering. *J. Open Source Softw.*, 2(11):205, 2017. URL <https://doi.org/10.21105/joss.00205>.
- [55] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 2018. URL <https://doi.org/10.21105/joss.00861>.
- [56] Shikib Mehri and Maxine Eskenazi. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting, July 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.sigdial-1.28>.
- [57] Shikib Mehri and Maxine Eskenazi. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.64. URL <https://aclanthology.org/2020.acl-main.64>.
- [58] Koh Mitsuda, Ryuichiro Higashinaka, and Yoshihiro Matsuo. 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, 2019. URL https://link.springer.com/chapter/10.1007/978-3-319-92108-2_3.
- [59] Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, 2023.
- [60] Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 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 (Volume 1: Long Papers), pages 1699–1713, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.134. URL <https://aclanthology.org/2021.acl-long.134>.
- [61] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [62] Zhufeng Pan, Kun Bai, Yan Wang, Lianqiang Zhou, and Xiaojiang Liu. 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, 2019. URL <https://aclanthology.org/D19-1191/>.
- [63] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. URL <https://aclanthology.org/P02-1040.pdf>.
- [64] Vitou Phy, Yang Zhao, and Akiko Aizawa. Deconstruct to Reconstruct a Configurable Evaluation Metric for Open-Domain Dialogue Systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.368. URL <https://aclanthology.org/2020.coling-main.368>.
- [65] David Quarfoot and Richard A Levine. How robust are multirater interrater reliability

- indices to changes in frequency distribution? *The American Statistician*, 70(4):373–384, 2016. URL <https://www.tandfonline.com/doi/abs/10.1080/00031305.2016.1141708>.
- [66] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL <http://jmlr.org/papers/v21/20-074.html>.
- [67] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrew. Conversational AI: The Science Behind the Alexa Prize, January 2018. URL <https://arxiv.org/abs/1801.03604v1>.
- [68] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL <https://aclanthology.org/P19-1534>.
- [69] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL <https://aclanthology.org/P19-1534>.
- [70] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features.

- In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.58. URL <https://aclanthology.org/2021.acl-long.58>.
- [71] Nils Reimers and Iryna Gurevych. 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, 2019. URL <https://aclanthology.org/D19-1410/>.
- [72] Susan Robinson, David R Traum, Midhun Ittycheriah, and Joe Henderer. What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. In *LREC*, pages 1–7, 2008.
- [73] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. URL <https://aclanthology.org/2021.eacl-main.24>.
- [74] Sahand Sabour, Chujie Zheng, and Minlie Huang. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237, 2022.
- [75] Manuela Sanguinetti, Alessandro Mazzei, Viviana Patti, Marco Scalerandi, Dario Mana, and Rossana Simeoni. Annotating Errors and Emotions in Human-Chatbot Interactions in Italian. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 148–159,

- Barcelona, Spain, December 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.law-1.14>.
- [76] Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. Rome was built in 1776: A Case Study on Factual Correctness in Knowledge-Grounded Response Generation, October 2021. URL <http://arxiv.org/abs/2110.05456>. arXiv:2110.05456 [cs].
- [77] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3027–3035, 2019. URL <https://ojs.aaai.org/index.php/AAAI/article/view/4160>.
- [78] Siqi Shen, Deepanway Ghosal, Navonil Majumder, Henry Lim, Rada Mihalcea, and Soujanya Poria. Multiview Contextual Commonsense Inference: A New Dataset and Task, November 2022. URL <http://arxiv.org/abs/2210.02890>. arXiv:2210.02890 [cs].
- [79] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.320. URL <https://aclanthology.org/2021.findings-emnlp.320>.
- [80] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can You Put it All Together: Evaluating Conversational Agents’ Ability to Blend Skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online, July 2020. Associ-

- ation for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.183. URL <https://aclanthology.org/2020.acl-main.183>.
- [81] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. URL <https://ojs.aaai.org/index.php/AAAI/article/view/11164>.
- [82] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- [83] Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.308. URL <https://aclanthology.org/2022.findings-acl.308>.
- [84] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [85] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [86] Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 308–319, 2022.

- [87] Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK, August 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.inlg-1.14>.
- [88] Deeksha Varshney, Akshara Prabhakar, and Asif Ekbal. Commonsense and named entity aware knowledge grounded dialogue generation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1335, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.95. URL <https://aclanthology.org/2022.naacl-main.95>.
- [89] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [90] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016. URL <https://arxiv.org/abs/1610.02424>.
- [91] Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. Recursively summarizing enables long-term dialogue memory in large language models. *arXiv preprint arXiv:2308.15022*, 2023.

- [92] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*, 2023.
- [93] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1363. URL <https://aclanthology.org/P19-1363>.
- [94] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.341. URL <https://aclanthology.org/2022.naacl-main.341>.
- [95] Jason Weston and Kurt Shuster. Blender Bot 2.0: An open source chatbot that builds long-term memory and searches the internet, July 2021.
- [96] Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 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, 2013. URL <https://link.springer.com/article/10.1186/1471-2288-13-61>.
- [97] Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness.

- In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5811–5820, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.515. URL <https://aclanthology.org/2020.acl-main.515>.
- [98] Sixing Wu, Ying Li, Ping Xue, Dawei Zhang, and Zhonghai Wu. Section-aware commonsense knowledge-grounded dialogue generation with pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 521–531, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL <https://aclanthology.org/2022.coling-1.43>.
- [99] Sixing Wu, Ying Li, Dawei Zhang, and Zhonghai Wu. Generating rational commonsense knowledge-aware dialogue responses with channel-aware knowledge fusing network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:3230–3239, 2022. doi: 10.1109/TASLP.2022.3199649.
- [100] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Bot-Adversarial Dialogue for Safe Conversational Agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.235. URL <https://aclanthology.org/2021.naacl-main.235>.
- [101] Jing Xu, Arthur Szlam, and Jason Weston. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.356. URL <https://aclanthology.org/2022.acl-long.356>.
- [102] Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. Grounded conversation

- generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.184. URL <https://aclanthology.org/2020.acl-main.184>.
- [103] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/pdf?id=SkeHuCVFDr>.
- [104] Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. CARE: Commonsense-Aware Emotional Response Generation with Latent Concepts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14577–14585, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i16.17713. URL <https://ojs.aaai.org/index.php/AAAI/article/view/17713>. Number: 16.
- [105] Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. Care: commonsense-aware emotional response generation with latent concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14577–14585, 2021.
- [106] Wanjun Zhong, Lianghong Guo, Qiqi Gao, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023.
- [107] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629, 2018.
- [108] Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. Commonsense-Focused Dialogues for Response Generation: An Empirical Study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–132,

- Singapore and Online, July 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.sigdial-1.13>.
- [109] Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–132, 2021.
- [110] Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. Commonsense-focused dialogues for response generation: An empirical study. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 121–132, 2021.
- [111] Pei Zhou, Hyundong J. Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. Reflect not reflex: Inference-based common ground improves dialogue response quality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. URL <https://aclanthology.org/2022.emnlp-main.714/>.
- [112] Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 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, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.88. URL <https://aclanthology.org/2022.acl-long.88>.
- [113] Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 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, 2022. URL <https://aclanthology.org/2022.acl-long.88/>.

- [114] Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*, 2023.