CICERO: A Dataset for Contextualized Commonsense Inference in Dialogues

Deepanway Ghosal [DeCLaRe] **Siqi Shen** [M] **Navonil Majumder** [DeCLaRe]
Rada Mihalcea [M] **Soujanya Poria** [DeCLaRe]

[DeCLaRe] DeCLaRe Lab, Singapore University of Technology and Design, Singapore
[M] University of Michigan, USA
{deepanway_ghosal@mymail.,navonil_majumder@,sporia@}sutd.edu.sg
{shensq,mihalcea}@umich.edu

CICERO is available at: https://declare-lab.github.io/CICERO

Abstract

This paper addresses the problem of dialogue reasoning with contextualized commonsense inference. We curate CICERO, a dataset of dyadic conversations with five types of utterance-level reasoning-based inferences: cause, subsequent event, prerequisite, motivation, and emotional reaction. The dataset contains 53,105 of such inferences from 5,672 dialogues. We use this dataset to solve relevant generative and discriminative tasks: generation of cause and subsequent event; generation of prerequisite, motivation, and listener's emotional reaction; and selection of plausible alternatives. Our results ascertain the value of such dialogue-centric commonsense knowledge datasets. It is our hope that CICERO will open new research avenues into commonsense-based dialogue reasoning.

1 Introduction

Conversational content on the internet is quickly growing, and such content holds valuable knowledge about how information exchange takes place among speakers. A key step towards understanding such dialogues is gaining the ability to reason with the information shared in the dialogue. To this end, we curate a dataset of dyadic conversations named CICERO (**C**ontextual**I**zed **C**ommons**E**nse Infe**R**ence in dial**O**gues) , which contains inferences around the utterances in the dialogues. The dataset focuses on five types of reasoning-based inferences for a given utterance in a dialogue: cause, subsequent event, prerequisite, motivation, and emotional reaction.

Arguably, making such reasoning-based inferences often demands commonsense knowledge, especially when the inference is implicit. Fig. 1a shows such a case where the cause behind the target utterance is not explicit in the context. However, applying the commonsense knowledge `worn gloves` $\xrightarrow{motivates}$ `buy new pair of gloves` allowed the annotator to infer a probable cause of the utterance. On the other hand, commonsense can be crucial in sifting relevant information from the context. Fig. 1b depicts an instance where the cause behind the target utterance is inferred from the context. This inference can be explained by commonsense knowledge (see Fig. 3) such as `repetitive consumption of the same food` $\xrightarrow{causes}$ `boredom` $\xrightarrow{dispelled\ by}$ `changing food` $\xrightarrow{achieved\ by}$ `eating at McDonald's`. Thus, it is reasonable to posit that such knowledge could aid to bridge the gap between the input and the target inference.

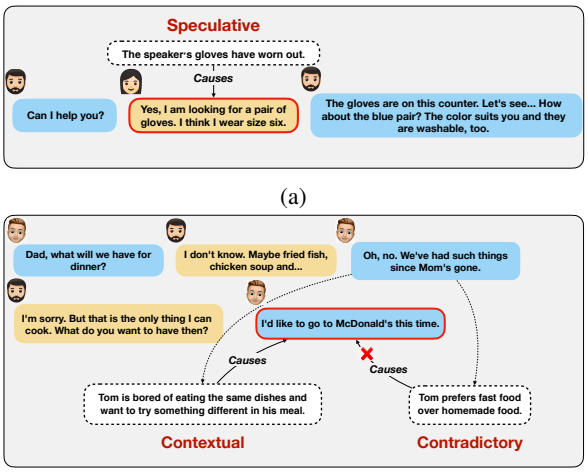


Figure 1: Illustration of (a) contextualized commonsense speculation and (b) contradictory inferences in dialogues.

ATOMIC (Sap et al., 2019; Hwang et al., 2020) is one such dataset for commonsense reasoning-based inference, allowing for a large set of inference types. However, ATOMIC is context-free, as it only provides inferences on short phrases, ignoring the broader context around them. Making an inference on an entire utterance, on the other hand, requires understanding the context around it. As per Grice's maxim (Grice, 1975), in conversations, the interlocutors provide any piece of

information as is needed, and no more. Thus, much of the information required to understand an utterance is likely interspersed along the dialogue, and not necessarily localized in the given utterance. For instance, in the example in Figure 1b, understanding the cause for one of the speakers' desire to go to McDonald's requires the context of the previous utterances. ATOMIC is thus not ideal for commonsense reasoning-based inferences on dialogues, where context is critical for understanding an utterance's implications. We confirm this with our experiments in the subsequent sections (§4).

GLUCOSE (Mostafazadeh et al., 2020) exclusively curates causal inferences — *cause, enable,* and *result in* – from monologues. Thus, it is not ideal for making context-consonant inferences on the dialogues. Also, dialogue-specific dimensions like *motivation* and *reaction* are beyond its scope.

On the other hand, CIDER (Ghosal et al., 2021a) does provide a dataset for commonsense-based inference on dialogues, but it is limited to inferences explicitly observable in the dialogues. As such, systems based on CIDER cannot effectively speculate around the dialogue for implicit inference.

CICERO strives to bring the best of these three datasets by creating a dataset that can enable models to effectively operate on a dialogue by considering the context and speculating when the answer is not apparent.

2 Construction of CICERO

We create CICERO – a large dataset of English dyadic conversations annotated with five types of inferences with the help of human annotators, who are instructed with a carefully crafted set of guidelines.

2.1 Annotation Instructions

The annotators are given a dialogue and a `target` utterance, as exemplified in Fig. 2. The annotators are then asked to make an inference, posed as a question, about the `target` utterance. They write a one-sentence answer that is grammatically correct, concise, and consistent with the dialogue. The answer may contain both *overt* and *speculative* scenarios. An overt scenario is explicitly or implicitly present in the dialogue context. If such contextual scenarios answer the question, the annotators write them as a well-formed sentence. However, in many cases, the dialogue may not hold the answer, neither explicitly nor implicitly. In such cases, the annotators are asked to speculate plausible scenarios around the dialogue, using commonsense and world knowledge, to devise answers that do `not contradict` the given dialogue context.

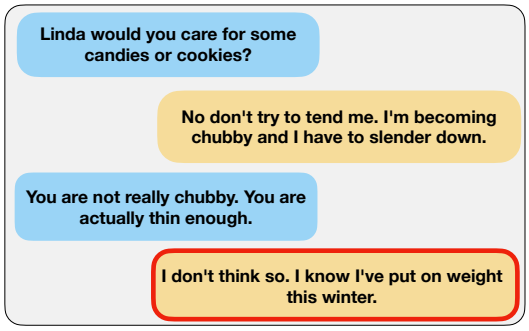


Figure 2: A `dialogue-target` pair. The utterances with red border is the target for this dialogue.

Given the `dialogue-target` pair in Fig. 2, at least one of the following five inferences about the `target` is made by the annotators:

Figure 3: Intermediate commonsense inference steps.

Q1. What is the event that directly causes (overt) or could cause (speculative) `Target`? The annotators consider if any of the events that are or likely to be antecedent to the target can cause the `target`.
Answer: Linda didn't exercise regularly during the winter. **Remark:** The annotators provided possible, speculative answers as the dialogue itself does not provide any reason for Linda's weight gain.

Q2. What subsequent event happens (overt) or could happen (speculative) following the `Target`? The annotators write about the event that happens or could happen following the `target`. Additionally, annotators were told that sometimes, such subsequent events of the `target` are triggered or likely to be triggered by the `target`.
Answer: Linda starts a diet and tries to lose weight.

Admiration	Affection	Afraid	Angry	Annoyed
Anticipating	Anxious	Apprehensive	Ashamed	Awe
Awkwardness	Boredom	Calmness	Caring	Confident
Confusion	Content	Craving	Devastated	Disappointed
Disgusted	Eagerness	Embarrassed	Encouragement	Enthusiasm
Excited	Faithful	Fear	Furious	Grateful
Gratitude	Guilty	Happy	Hopeful	Impressed
Interest	Jealous	Joyful	Lonely	Nostalgic
Prepared	Proud	Relief	Romance	Sad
Satisfaction	Sentimental	Surprised	Terrified	Trusting

Table 1: Possible emotional reactions of the listener.

Remark: The answer is speculative as the dialogue contains no explicit/implicit subsequent event.

Q3. What is (overt) or could be (speculative) the prerequisite of `Target`? Does the `target` have any direct prerequisite or dependency that has to happen or be fulfilled first? (In most cases, prerequisite is the state/event which has to be satisfied before another event causes `target`.) The answer is a state/event which enables the happening of the `target`. In other words, prerequisites are the prior assumptions or background information that the interlocutors agree on about the context.
Answer: Linda was slimmer before the winter.
Remark: Annotators were required to understand the difference between cause and prerequisite clearly before proceeding with the final annotation. Cause of an event X is the event that directly causes X. Prerequisite of an event X is the condition which has to be satisfied in order for X to happen.

Q4. What is an emotion or basic human drive that motivates or could motivate `Target`? Consider the basic human drives, needs (and/or likely emotions) of the speaker of the `target`. Basic human drives and needs are food, water, clothing, warmth, rest, security, safety, intimate relationships, friends, prestige, feeling of accomplishment, self-fulfillment, creative activities, enjoyment, etc. Do any of these human drives/states of mind/emotional feelings motivate the `target`?
Answer: `Not Applicable` for this target.

Q5. What is the possible emotional reaction of the listener: A (or B)? What could be the possible emotional reaction or responses of the listener with respect to the `target`? The annotators capture the appropriate emotion of the listener using the emotion terms listed in Table 1 verbatim or related words (e.g., anxious, confused, interested, etc).
Answer: The listener encourages Linda to maintain her diet.

Additional Guidelines. To ensure the quality and diversity of the samples, we also ask the annotators to adhere to the following guidelines:

- Be creative in speculation. Refrain from rephrasing the *target* and writing low-effort trivial answers. It is recommended to skip a question if rephrasing the *target* is the only possible answer.
- Avoid repeating the same answer for distinct questions on the same *target*.
- The answer must be consistent with the given dialogue.
- It is recommended to base the answer on the most important phrase of the *target* should it contain multiple phrases.

2.2 Dialogue Selection for CICERO

2.2.1 Source Datasets

To build CICERO, we use the dyadic dialogues of the following three datasets:
DailyDialog (Li et al., 2017) covers dialogues from wide range of topics — life, work, relationships, tourism, finance, etc. The constituent utterances are labelled with emotion and dialogue-act.
MuTual (Cui et al., 2020) is a multi-turn dialogue reasoning dataset. Given a dialogue history, the objective is to predict the next utterance by considering aspects such as intent, attitude, algebraic, multi-fact, and situation reasoning.
DREAM (Sun et al., 2019) is a multiple-choice reading-comprehension dataset collected from exams of English as a foreign language. The dataset presents significant challenges as many answers are non-extractive and require commonsense knowledge and multi-sentence reasoning.

2.2.2 Selection Process

We use the following procedure to select a subset of dialogues from the three datasets:

1. We remove dialogues that are too short or long on either utterance or word level. Dialogues with fewer than five utterances or fewer than six words per utterance on average are removed. Dialogues having more than 15 utterances or more than 275 words in total are also removed.

2. All three source datasets contain dialogues having near identical utterances. We remove these near duplicate dialogues to ensure topical diversity of CICERO. We use a sentence embedding model based on fine-tuned RoBERTa (Gao et al., 2021) to extract dense feature vectors of the dialogues. We remove the duplicates assuming that a pair of duplicate dialogues have at least 0.87 cosine similarity.

2.3 Target Utterance Selection

Given a dialogue D, we select the target utterances as follows:

- We first determine the number of target utterances in D: if D has 1–6 utterances, then we select 2 or 3 targets; if it has 7–12 utterances then we select 3–5 targets; otherwise, we select 4–7 targets if it has more than 12 utterances.
- We divide D into 2–3 segments having roughly equal number of consecutive utterances. We choose roughly an equal number of the top-ranking utterances from each segment. We call this set of utterances x_1. The ranking is performed using a sentence ranking algorithm (Erkan and Radev, 2004; Mihalcea and Tarau, 2004) with sentence-BERT embeddings (Reimers and Gurevych, 2019a).
- We also select the longest utterances in D and the utterances that contain phrases such as *I'm, I'd, I've, I'll* or their expansions. We call this set of utterances x_2. The sets x_1 and x_2 may not be disjoint.
- Set x_3 consisting of the final utterance of D.

We choose the inference type for the target utterances from the sets $x_{1,2,3}$ as follows:

- From $x_1 \cup x_2$:
  - Subsequent Event: 80% of the targets.
  - Both Cause and Prerequisite: 60% of the targets.
  - Exclusively Cause: 28% of the targets.
  - Exclusively Prerequisite: 12% of the targets.
- From x_2: Motivation for all targets.
- From x_3: Reaction of listener for all targets.

2.4 Quality Assurance of CICERO

Dataset quality is ensured with the following steps:

- Initially, we sample 50 random dialogues and manually annotate all the questions (as in §2.1) in those. Each annotator is then evaluated on those dialogues, and is selected for the annotation task if 95% of his/her annotations are approved by us.
- We constantly review and provide feedback to the annotators during the annotation process. Annotators are also instructed to amend their answers.
- Upon completion of the annotation, we employ three additional annotators who manually check the annotated samples and score their acceptability. These annotators reached a consensus for approving 86% of these samples. The samples not bearing majority agreement were removed from the dataset.

A (u_1): Hi, Jenny. Is it true you're moving to London? **B (u_2)**: Yes, it is. **A (u_3)**: What made you decide to do that? **B (u_4)**: Work, mainly. I'm sure I'll be able to find a job there. **A (u_5)**: You're probably right. But where are you going to live? **B (u_6)**: I hope I'll find a flat to share with somebody. That way it will be cheaper. **A (u_7)**: Yes, that's a good idea. Are you taking your dog with you? **B (u_8)**: No, I don't think so. My parents have offered to take care of him, and I don't think he'd be happy in the city. **A (u_9)**: You're probably right. But aren't you afraid of moving to such a big place, especially after living in a small village? **B (u_{10})**: Not really. I think I'll enjoy myself. There's so much to do there; I expect I won't miss the countryside much and I can always come back and visit. **A (u_{11})**: Well, I just hope you'll invite me to stay when you get settled. **B (u_{12})**: Of course I will.
Target - u_6; **Inference: Cause**; **Annotation:** Being an expensive city, it is quite difficult to find an affordable place to live in London.
Target - u_{10}; **Inference: Cause**; **Annotation:** Jinny realizes that a city like London will provide a great quality of life for her.
Target - u_6; **Inference: Subsequent Event**; **Annotation:** The listener gives an idea to Jenny to find the flat on some online portal for searching flatmates as well plenty of cheaper options.
Target - u_{10}; **Inference: Subsequent Event**; **Annotation:** Jenny inquired a social club in London and ask for their membership to utilize her free time.
Target - u_4; **Inference: Prerequisite**; **Annotation:** Jenny has completed her studies.
Target - u_{12}; **Inference: Prerequisite**; **Annotation:** Jenny and the listener are good friends.
Target - u_6; **Inference: Motivation**; **Annotation:** Jenny is optimistic about having someone as her flatmate to save rent.
Target - u_{12}; **Inference: Reaction**; **Annotation:** The listener is happy for Jenny and looks forward to being invited to London by Jenny.

Table 2: Annotated examples in CICERO marked with the target utterance and the inference type. Inference types *Cause, Effect, Prerequisite, Motivation, and Reaction* correspond to questions Q1, Q2, Q3, Q4, and Q5, respectively, in §2.1.

The statistics of the annotated dataset is shown in Table 3. A number of annotated examples from CICERO are also shown in Table 2.

2.5 Features of CICERO

Following Table 3, a majority (~ 59%) of the inferences in CICERO are causal in nature. Again, roughly 80% of the inferences are speculative and context consonant. CICERO is thus much more versatile in terms of its applications as compared to CIDER (Ghosal et al., 2021a) that only contains explicit contextual inferences. CICERO also contains varied commonsense knowledge – from general to physical and social commonsense (see Appendix B for more details).

3 Commonsense Inference on CICERO

We design generative and multi-choice question answering tasks on CICERO to evaluate dialogue-level commonsense-based reasoning capabilities of

Description	# Instances	Percentage
# Dialogues / # Inferences		
DailyDialog	3,280 / 30,509	57.82 / 57.34
MuTual	1,640 / 14,207	28.91 / 26.70
DREAM	753 / 8,488	13.27 / 15.95
Total	5,673 / 53,204	–
# Dialogues with # Inferences		
less than 10	3,140	55.35
between 10-20	2,518	44.39
between 21-30	15	0.26
Avg. # Inferences per Dialogue	9.38	–
Instances with # Correct Answers		
only 1	45759	86.01
only 2	4985	9.37
> 2	2460	4.62
Inference Types in Train / Validation / Test		
Cause	10,386 / 3,060 / 3,071	33.06 / 28.10 / 28.18
Subsequent Event	6,617 / 4,021 / 4,050	21.06 / 36.93 / 37.16
Prerequisite	7,501 / 1,347 / 1,396	23.87 / 12.37 / 12.81
Motivation	4,412 / 1,420 / 1,401	14.04 / 13.04 / 12.86
Reaction	2,502 / 1,040 / 980	7.96 / 9.55 / 8.99

Table 3: Statistics of the annotated CICERO dataset.

language models.

3.1 Task 1: CICERO$_{NLG}$

The objective is to generate the answer to question q, representing one of the five inference types, for a target utterance u_t in a dialogue D. Each inference type has its respective q (illustrated in §4).

Task 1.1: Dialogue Causal Inference. Causality pertains to causes and effects of events and situations. We formulate the dialogue causal inference task as generating the cause or subsequent event of an utterance as an answer to a causal question:

1. **Cause:** Given D, u_t, generate the cause c_t of u_t.

2. **Subsequent Event:** Given D, u_t, generate the subsequent event e_t of u_t.

3. **Subsequent Event Clipped (Subsequent EC):** Given u_t, the dialogue up to u_t: $D_{:u_t}$, generate the subsequent event e_t of u_t.

We consider two different scenarios for *subsequent event*, as the event often appear after the target utterance in the dialogue. Hence, subtask 3 is more challenging to evaluate a model's ability to reason about unobserved effects. We extend subtasks 1, 2 to incorporate longer chains and formulate the chained generation task. We consider utterances u_t in our dataset that has both cause and subsequent event annotated i.e. $c_t \rightarrow u_t \rightarrow e_t$. The causal chain is considered as a triplet, and we formulate tasks where a missing segment has to be generated from the rest of the components:

4. **Chained Cause**: Generate c_t from u_t and e_t.

5. **Chained Subsequent Event (Chained SE)**: Generate e_t from u_t and c_t.

Task 1.2: Prerequisite, Motivation and Reaction Generation. The objective is to generate the prerequisite/motivation/reaction of listener from a given D and u_t. The target u_t is the final utterance of D for reaction generation. Generating the prerequisite (task 1.2.1) requires an understanding of the dependency of events. Generating the motivation (task 1.2.2) and reaction (task 1.2.3) is about learning basic human drives and emotions. Note that, reaction generation is a different problem from dialogue response generation. Responses follow utterance level distributions which are substantially different from emotional reactions.

3.2 Task 2: CICERO$_{MCQ}$

Given dialogue D, target u_t, one of the five questions (inference type) q, true answer a_t, alternate choices $F_t = \{f_{t1}, f_{t2}, f_{t3}, f_{t4}\}$, the CICERO$_{MCQ}$ task aims to select the correct answer a_t (see Fig. 4) and additionally any answer among F_t which might be correct. The alternate choices F_t are created through a combination of automated generation and human supervision as follows:

• We train a `T5` large model on SNLI contradictory pairs (Bowman et al., 2015) and Time-Travel counterfactual pairs (Qin et al., 2019) to generate contradictions/counterfactuals from input sentences. We use this model to generate a pool of alternate answers from the true annotated answers. Alternate answers which have an embedding cosine similarity less than 0.9 with the true answer (from *all-mpnet-base-v2* in Reimers and Gurevych (2019b)) and are contradictory w.r.t the true answer (from *roberta-large-mnli*) are kept, and the rest are discarded. The filtered set is termed $\mathbb{N}$.

• We use the adversarial filtering (AF) algorithm (Zellers et al., 2018) to select the four alternate answers F_t from $\mathbb{N}$. For multi-choice QA tasks, AF is an effective method to detect easily identifiable alternate answers and replace them with more difficult candidates by detecting and reducing stylistic artifacts. The algorithm is as follows:

(i) We start with annotated true answer a_t and any four choices $\hat{F}_t$ from $\mathbb{N}$ for all instances in our dataset to create $\hat{\mathbb{D}}$. We randomly split $\hat{\mathbb{D}}$ into $\hat{\mathbb{D}}_{train}$ (80%) and $\hat{\mathbb{D}}_{test}$ (20%) according to dialogue IDs.
(ii) A multi-choice QA model (discriminator) is trained on $\hat{\mathbb{D}}_{train}$ that scores all five choices for all instances in $\hat{\mathbb{D}}_{test}$. The highest scoring choice is considered as the predicted answer. For a particular test instance, choices in $\hat{F}_t$ that have lower scores than a_t are replaced with other high scoring choices in $\mathbb{N} - \hat{F}_t$. Answers in $\hat{F}_t$ which are being replaced

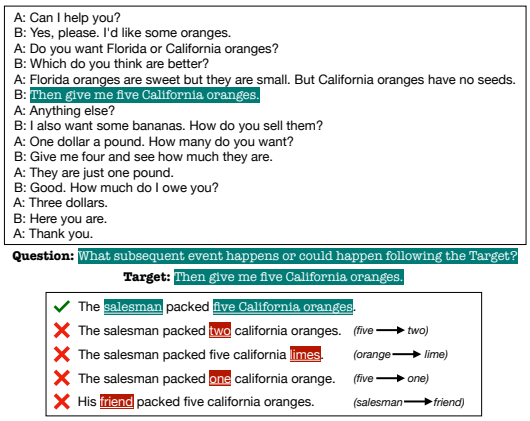


Figure 4: A data sample of CICERO for the Plausible Alternative Selection task. Here, commonsense is required to infer – a salesman packs the items that buyers want to purchase. In this particular dialogue, the buyer wants to purchase five California oranges and four bananas which can be inferred from the context.

are removed from $\mathbb{N}$.

(iii) $\hat{F}_t$ now consists of relatively more difficult choices. A new random split $\hat{\mathbb{D}}_{train}$ and $\hat{\mathbb{D}}_{test}$ is created, and we go back to step **(ii)**. The algorithm is terminated when the accuracy in successive $\hat{\mathbb{D}}_{test}$ reaches a convergence. The final alternate choice set is termed as F_t.

The AF algorithm ensures a robust final dataset $\mathbb{D}$ irrespective of the final train, validation, and test split. We use a new *roberta-large* model to initialize the discriminator and train for 3 epochs before scoring and replacement in step **(ii)**. 14 iterations were required for convergence in $\mathbb{D}_{test}$.

• Annotators perform manual checking on the final AF selected choices F_t. They mark each of the alternate choices in F_t in $\mathbb{D}$ to be speculatively correct or incorrect given the context. Hence, instances might have correct answers in F_t in addition to the originally annotated correct answer a_t. The final dataset statistics after this step are given in Table 3.

Task 2.1: Single Answer Selection. Consider instances where F_t doesn't contain any correct answer. The task is to select the correct answer a_t among the five choices given D, u_t, and q.

Task 2.2: All Answers Selection. This task is performed on the entire dataset (including the subset of data which is used in Task 2.1. There might be one or more correct answers for a particular instance resulting from the AF algorithm. The task is to select all the correct answer(s) (including a_t) among the five choices given D, u_t, and q.

4 CICERO Tasks: Experimental Results

We split our dataset in dialogue level where the training, validation and test instances are obtained from a total of 3477, 1097, 1098 distinct dialogues respectively. This results in a 60:20:20 proportion of total annotation instances. The three sets have 17365, 5370, and 5331 unique target utterances respectively. We tune on the validation dataset and report results on the test dataset (average of 5 runs). For the sake of brevity, the detailed hyperparameters are given in the supplementary material.

We use the following questions (q) for the five inference types for all the tasks: **Cause**: *What is or could be the cause of target?* **Subsequent Event**: *What subsequent event happens or could happen following the target?* **Prerequisite**: *What is or could be the prerequisite of target?* **Motivation**: *What is or could be the motivation of target?* **Reaction**: *What is the possible emotional reaction of the listener in response to target?*

4.1 Baseline Models

CICERO$_{NLG}$ **— (1.1–1.2).** We use large versions of `T5` (Raffel et al., 2020) and `GLUCOSE-T5` (Mostafazadeh et al., 2020) as our models. `GLUCOSE-T5` is a T5 large model that is pre-trained on the GLUCOSE dataset. We concatenate q, u_t, and the context c with separators to form the input to the model: `q` <sep> u_t <sep> `c`. The context c is formed by concatenating utterances of $D_{:u_t}$ (subsequent event clipped) or D (all other tasks). For the chained generation task, we additionally provide the cause/subsequent event as input. The inputs are `q` <sep> u_t <sep> `subsequent event:` e_t <sep> `c` and `q` <sep> u_t <sep> `cause:` c_t <sep> `c` for cause and subsequent event generation, respectively. The objective is to generate the answer as output in the sequence-to-sequence setup. We use teacher forcing during training and beam search during inference.

CICERO$_{MCQ}$ **— Single Answer Selection (2.1).** We use `RoBERTa-large`, `ELECTRA-large`, `T5-large`, and `Unified QA Large` for this task. The input to the models for `RoBERTa-large`, `ELECTRA-large` is the concatenation of question q, target u_t, dialogue D, and candidate answers $x_j, j \in \{1, ..., 5\}$: <cls> `q` <sep> u_t <sep> `D` <sep> x_j. Each score is predicted from the corresponding <cls> vector and the highest scoring one is selected as the answer. For seq2seq

models `T5-large`, and `Unified QA Large`, we use the following input — `q <sep> 1)` x_1 `2)` x_2 `3)` x_3 `4)` x_4 `5)` x_5 `<sep>` u_t `<sep> D`. The output to be generated is the correct answer – such as x_1 or x_2.

CICERO$_{MCQ}$ — All Answers Selection (2.2). We use seq2seq models `T5-large`, and `Unified QA Large` as they can generate both single and multiple-answers (with separator tokens) as output. The input is `q <sep> 1)` x_1 `2)` x_2 `3)` x_3 `4)` x_4 `5)` x_5 `<sep>` u_t `<sep> D`. The output to be generated are the correct answer(s), such as x_2 (single answer) or x_1 `<sep>` x_3 `<sep>` x_4 (multiple answers). Here, $x_1 - x_5$ denotes the five possible choices shuffled randomly.

4.2 Results of the CICERO$_{NLG}$ Task

Automatic Evaluation Metrics. For generative tasks, we report the following metrics: **BLEU** (Papineni et al., 2002), **METEOR** (Banerjee and Lavie, 2005), **ROUGE** (Lin, 2004), **CIDEr** (Vedantam et al., 2015), and **Sem-Sim** which computes the semantic cosine similarity of two sentences using the supervised `RoBERTa-large` sentence embedding model (Gao et al., 2021). All scores are reported in the range of 0-1.

Human Evaluation Metrics. Due to significant dissonance with human evaluation, automatic evaluation metrics are often considered not reliable for generation quality evaluation in literature. Hence, we resort to human evaluation metrics. The human annotators rate on an integer scale from 1 (worst) to 5 (best) on three coarse attributes: **Creativity**: As the majority of the inferences require speculation, this metric measures how creative the models and the annotators are. **Contextuality**: Whether the generated or annotated inferences fit the context. **Fluency**: Whether the generated or annotated inferences are grammatically correct.

Results of Automatic Evaluation. The results for the generative tasks are reported in Table 4 and Table 5. We observe that the fine-tuned models perform quite similarly across various metrics in Table 4. The `T5` model achieves the best performance in most of the experimental settings. The results indicate that the *causal* types are more challenging to infer than the *Motivation*, and *Reaction*. However, the models are posed to the most challenging instances in the case of *Prerequisite* type as inferring this type requires rich commonsense and background knowledge. Hence, for this category, the models achieve a low score compared to rest of the inference categories. We also notice that exposing the future utterances to the models help in attaining better inference performance for the relation type *Subsequent Event*. The trained models perform worse when the future utterances are not available in the input as seen in the *Subsequent Event Clipped* task. A significant drop of performance is noticed in the CIDEr metric. For the chained generation tasks (1.1.4 and 1.1.5), we notice (refer to Table 5) a very similar trend in models' performance i.e., the models tend to perform better for these two experimental settings compared to only *Cause* (1.1.1) and *Subsequent Event* (1.1.2) predictions. We can surmise that the additional cues from the available annotations of *Subsequent Event* type in the Chained Cause setting, and the *Cause* type in the Chained Subsequent Event setting are the key

	Model	BLEU2	METEOR	ROUGE	CIDEr	Sem-Sim
(1.1.1) Cause	T5	0.1493	0.1630	0.2626	0.4560	0.6278
	GLUCOSE-T5	**0.1563**	**0.1634**	**0.2707**	**0.4915**	**0.6305**
	T5*	0.0042	0.0200	0.0266	0.0237	0.3735
	GLUCOSE-T5*	0.0287	0.0560	0.0827	0.1332	0.4442
(1.1.2) SE	T5	**0.1619**	**0.1662**	**0.2760**	**0.4119**	0.6276
	GLUCOSE-T5	0.1611	0.1628	0.2778	0.4430	**0.6297**
	T5*	0.0045	0.0191	0.0264	0.0241	0.3865
	GLUCOSE-T5*	0.0001	0.0070	0.0024	0.0032	0.3073
(1.1.3) SE Clipped	T5	0.1448	**0.1549**	0.2618	0.3099	**0.6123**
	GLUCOSE-T5	0.1461	0.1523	0.2645	**0.3238**	0.6094
	T5*	0.0199	0.0439	0.0564	0.0762	0.4549
	GLUCOSE-T5*	0.0001	0.0066	0.0025	0.0034	0.3063
(1.2.1) Prerequisite	T5	0.1002	0.1282	**0.2176**	**0.3357**	**0.5902**
	GLUCOSE-T5	0.1001	**0.1299**	0.2197	0.3144	0.5896
	T5*	0.0043	0.0222	0.0279	0.0225	0.3541
	GLUCOSE-T5*	0.0108	0.0394	0.0625	0.0889	0.4392
(1.2.2) Motivation	T5	0.2503	0.1998	0.3781	0.7109	0.6973
	GLUCOSE-T5	**0.2582**	**0.2037**	**0.3840**	**0.7499**	**0.7048**
	T5*	0.0033	0.0183	0.0257	0.0181	0.4038
	GLUCOSE-T5*	0.0174	0.0434	0.0632	0.0696	0.4053
(1.2.3) Reaction	T5	**0.2397**	**0.1939**	**0.3720**	0.5177	**0.6665**
	GLUCOSE-T5	0.2318	0.1903	0.3716	**0.5364**	0.6653
	T5*	0.0037	0.0201	0.0239	0.0167	0.3899
	GLUCOSE-T5*	0.0213	0.0459	0.0759	0.0719	0.4125

Table 4: Results of the CICERO$_{NLG}$ task. `T5*` and `GLUCOSE-T5*` are not fine-tuned on our dataset. All models are Large models. SE denotes Subsequent Event.

Model	BLEU2	METEOR	ROUGE	CIDEr	Sem-Sim
(1.1.4) Chained Cause					
T5	0.1566	0.1675	0.2757	0.5303	0.6518
GLUCOSE-T5	0.1600	**0.1697**	**0.2796**	**0.5633**	**0.6557**
(1.1.1)* Cause					
T5	0.1503	0.1635	0.2634	0.4591	0.6284
GLUCOSE-T5	**0.1564**	**0.1636**	**0.2709**	**0.4915**	**0.6310**
(1.1.5) Chained SE					
T5	**0.1813**	**0.1784**	0.2940	0.5136	0.6469
GLUCOSE-T5	0.1789	0.1776	**0.2943**	**0.5218**	**0.6516**
(1.1.2)* SE					
T5	**0.1622**	0.0841	0.2764	0.4167	0.6279
GLUCOSE-T5	0.1612	**0.1628**	**0.2778**	**0.4471**	**0.6294**

Table 5: Results of the CICERO$_{NLG}$ subtasks – chained cause and subsequent event generation. (1.1.1)* and (1.1.2)* indicates results from Task 1.1.1 and 1.1.2 (as in Table 4), but only for targets which have both cause and effect annotated, ensuring a fair comparison with (1.1.4) and (1.1.5). SE denotes Subsequent Event.

Metric	Gold	T5	GLUCOSE	T5*	GLUCOSE*
Creativity	4.7	3.8	3.9	2.4	1.9
Contextuality	4.8	4.1	4.3	2.1	2.1
Fluency	5.0	4.8	4.9	1.9	2.9

Table 6: Results of the human evaluation for CICERO$_{NLG}$. T5* and GLUCOSE-T5* represent non fine-tuned versions.

A (u_1): I'm hungry, let's order up something to eat. **B (u_2)**: Ok, maybe we can order a soup and a salad from the restaurant down the street. **A (u_3)**: I was thinking of getting a hamburger, fries and a chocolate sundae. **B (u_4)**: You eat too much junk food. That sort of stuff clogs up your arteries and is very high in cholesterol. **A (u_5)**: Well I never seem to gain weight so I don't mind. **B (u_6)**: It's not only about getting fat or not, it's about being healthy. You could really have some health problems later on. **A (u_7)**: How about pizza or maybe some fried chicken! Better yet, let's order some hot dogs! **B (u_8)**: You are a lost cause.
Target - u_1; Inference: Cause; Answers: Gold - The speaker did not eat breakfast. **T5** - The speaker and the listener are going to a restaurant for lunch. **GLUCOSE-T5** - The speaker and the listener are at a restaurant.
Target - u_2; Inference: Cause; Answers: Gold - The restaurant down the street is famous for its delicious food. **T5** - The listener is hungry and wants to eat something. **GLUCOSE-T5** - The speaker is hungry and wants to order something to eat.
Target - u_3; Inference: Motivation; Answers: Gold - The speaker is craving some savory fast food. **T5** - The speaker is craving for a hamburger, fries and a chocolate sundae. **GLUCOSE-T5** - The speaker is craving for a burger, fries and sundae.
Target - u_6; Inference: Prerequisite; Answers: Gold - The speaker is a fitness freak and keeps track of his daily diet. **T5** - The speaker is a healthy person. **GLUCOSE-T5** - The speaker is a health conscious person.
Target - u_7; Inference: Subsequent Event; Answers: Gold - The listener refused to eat anything that is unhealthy. **T5** - The speaker and the listener decided to order some hot dogs. **GLUCOSE-T5** - The speaker and the listener decided to order some hot dogs.
Target - u_8; Inference: Reaction; Answers: Gold - The listener felt embarrassed by the statement of the speaker. **T5** - The listener is shocked to hear the speaker's comment. **GLUCOSE-T5** - The listener is disappointed with the speaker's decision.

Table 7: Inferences by different models extracted from a sample dialogue for the CICERO$_{NLG}$ task.

to such performance improvement. As depicted in Table 4 (and also Table 6), the non fine-tuned versions of T5 and GLUCOSE-T5 perform poorly as they produce gibberish outputs across all the five inference categories indicating the importance of fine-tuning on CICERO.

Results of Human Evaluation. For each of the five inference types, we randomly sample 40 inferences generated by each model and their corresponding gold inferences. These inferences are then manually rated by three independent annotators based on the human-evaluated metrics. As suggested by Table 6, we observe that most of the fine-tuned models on CICERO perform similarly but fail to reach gold annotation performance. Moreover, as expected, the fine-tuned models significantly outperform their non fine-tuned counterparts. We provide some examples of the generated inferences in Table 7. Inspection of the model generated inferences reveal that usage of keywords from the dialogue without generalizing the events is more frequent. Generated inferences are significantly less diverse and creative than gold annotations.

Performance of GLUCOSE. GLUCOSE contains contextual commonsense inferences on events in monologues. Comparing the results (Table 4, Table 6) of fine-tuned and non fine-tuned checkpoints suggests that pre-training on a monologue-based contextual commonsense inference dataset does not ensure good performance on the same task for dialogues. Akin to the non fine-tuned T5, non fine-tuned GLUCOSE-T5 produces gibberish outputs for all the commonsense inference types but the causal and motivation types. We surmise this happens as these two commonsense types exist in the GLUCOSE dataset. Although the generated text for these two commonsense inference types are grammatically correct and sometimes contain contextual words, they are far from the desired quality, semantically very much dissimilar from the annotated gold instances, and rated low in the qualitative evaluation, as shown in Table 6. We also confirm the efficacy of fine-tuning the models on CICERO through human evaluation, as explained in §4.

4.3 Results of the CICERO$_{MCQ}$ Task

Evaluation Metrics. 1) **RoBERTa** and **ELECTRA**: The accuracy of selecting the correct answer is used to evaluate the performance of these models. 2) **T5** and **Unified QA**: The output is considered as a single answer if it doesn't contain any separator token. Otherwise, the output is segmented at separator tokens to obtain multiple answers. We then follow the method in Khashabi et al. (2020), where match is computed by comparing each of the generated answer(s) with the candidate choices based on their token-level overlap. For each generated answer, the most similar candidate choice is considered as the corresponding output. The prediction is considered as correct if the final output(s) is an exact match (EM) with the gold annotated answer(s).

Single Answer Selection (2.1). We report the results of this setting in Table 8. The reported metric is accuracy of selecting the correct answer. The overall score is 83.28% for RoBERTa and 86.82% for ELECTRA. ELECTRA has an edge over RoBERTa on all the five inference types. This could be a side effect of using RoBERTa as the backbone model for the AF algorithm and subsequently as a solver for

Model	Cause	SE	Prereq.	Motiv.	Emo. Reac.	Avg.
RoBERTa	83.34	83.17	79.48	86.33	84.26	83.28
ELECTRA	**87.09**	**86.09**	**85.15**	**90.31**	**86.11**	**86.82**
T5	95.19	**95.29**	94.93	**96.52**	96.99	95.54
Unified QA	**95.85**	94.99	**95.55**	96.35	**97.22**	**95.70**

Table 8: Accuracy scores for Task 2.1. Models are trained and evaluated on instances with a single correct answer.

Model	Eval On	Cause	SE	Prereq.	Motiv.	Emo. Reac.	Avg.
T5	S + M	**78.18**	74.72	**75.50**	**82.51**	**84.59**	**77.68**
Unified QA		78.12	**74.79**	75.36	81.58	84.08	77.51
T5	S	**93.20**	**91.28**	**91.27**	**95.19**	**95.14**	**92.71**
Unified QA		93.12	91.16	91.00	94.28	94.79	92.45
T5	M	**3.50**	2.77	**3.59**	**3.61**	**6.03**	3.38
Unified QA		**3.50**	**3.69**	**3.98**	2.58	**4.31**	**3.60**

Table 9: Exact match scores for Task 2.2. Models are trained on instances with both single and multiple correct answers, i.e., the entire dataset. SE → Subsequent Event; S → Single-Answer Instances; M → Multi-Answer Instances.

the final CICERO$_{MCQ}$ task. We think, this results expose the model dependency of the AF process. In other words, the negative samples chosen by the backbone model X for the AF algorithm will be difficult to distinguish from the human-annotated true samples using the same model X. These negative samples, however, could be relatively easier to identify using another model Y. The seq2seq models `T5` and `Unified QA` perform significantly better than `RoBERTa` and `ELECTRA` as can be seen in Table 8. While models like `RoBERTa, ELECTRA` encode each candidate answer separately, `T5` and `Unified QA` encode them together. Thanks to this joint encoding of candidate answers, `T5` and `Unified QA` can take advantage of more task-related information that `RoBERTa` and `ELECTRA` might miss due to the separate encoding scheme. We surmise it could be one of the reasons why the seq2seq models have an edge over `RoBERTa` and `ELECTRA` for this particular task. `T5` and `Unified QA` attain almost the same score for single answer selection. This is surprising as `Unified QA` is initialized from the `T5-large` checkpoint and then further trained on other QA datasets. As such, we think, the different fine-tuned domains of `Unified QA` does not help in the CICERO$_{MCQ}$ task.

All Answers Selection (2.2). We train and evaluate `T5` and `Unified QA` on the entire dataset of both single and multiple correct answers and report the results in Table 9. Overall, `T5` and `Unified QA` perform similarly. The general performance, across the models, on instances with multiple correct answers is much worse than instances with a single correct answer. We confirm this by reporting the results only on instances with multiple answers in Table 9, where `T5` and `Unified QA` achieve only 3.38% and 3.60% exact match, respectively. This could probably be attributed to the stark data imbalance of ~86/14% between single- and multi-answer instances, respectively (see Table 3).

5 Related Work

Commonsense knowledge has received more attention compared with factual knowledge, as it is usually not mentioned explicitly in the context. It is demonstrated to be essential in open-ended generation tasks, such as story explanation generation (Mostafazadeh et al., 2020), story end generation (Guan et al., 2019) and abductive reasoning (Bhagavatula et al., 2019). To infuse commonsense knowledge in NLP models, several approaches to tasks like sentence ordering (Ghosal et al., 2021b), emotion recognition (Ghosal et al., 2020), story generation (Guan et al., 2020; Xu et al., 2020) and dialogue generation (Zhou et al., 2018) use prevalent commonsense knowledge bases (CSKB) like ConceptNet (Speer et al., 2017) or ATOMIC (Sap et al., 2019). However, ConceptNet is context-free, meaning that they only capture relationships around a selected set of entities, without paying attention to the context where the entity occurs. Moreover, inference is often needed in discourse level, which do not always align with the entities in knowledge bases. Knowledge models such as COMET (Bosselut et al., 2019) is a way to circumvent this issue and make inferences on an utterance (sentence) level. But the generated knowledge still lacks the detail from the dialogue, as it is trained on the aforementioned knowledge base. Our approach, instead, centers on the dialogue dataset and provides more detailed commonsense inference at an utterance level.

6 Conclusion

We introduced CICERO, a new dataset for dialogue reasoning with contextualized commonsense inference. It contains ~53K inferences for five commonsense dimensions – cause, subsequent event, prerequisite, motivation, and emotional reaction – collected from ~5.6K dialogues. To show the usefulness of CICERO for dialogue reasoning, we design several challenging generative and multi-choice answer selection tasks for state-of-the-art NLP models to solve.

Acknowledgements

This work is supported by the A*STAR under its RIE 2020 AME programmatic grant RGAST2003 and project T2MOE2008 awarded by Singapore's MoE under its Tier-2 grant scheme.

Ethics Statement

The annotators for CICERO were hired through a data annotation service. The compensation was derived based on the country of residence of the annotators, as deemed by the company. The study has been categorized as "exempt" by the IRB. Annotators were strictly asked not to write any toxic content (hateful or offensive toward any gender, race, sex, religion). They were asked to consider gender-neutral settings in dialogues whenever possible.

The source dialogue datasets – DailyDialog, MuTual, and DREAM are high quality multi-turn dialogue datasets manually annotated by experts in dialogue, communication theory and linguistics. All three datasets have been extensively used and studied in the natural language processing literature. The three source datasets and our annotations in CICERO do not contain any personal data or any information that can uniquely identify individual people or groups.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 632–642.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021a. 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, Singapore and Online. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021b. Stack: Sentence ordering with temporal commonsense knowledge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8676–8686.

Herbert P. Grice. 1975. Logic and conversation. *Speech acts*, pages 41–58.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, 8:93–108.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs. *CoRR*, abs/2010.05953.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

D. Khashabi, S. Min, T. Khot, A. Sabhwaral, O. Tafjord, P. Clark, and H. Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *EMNLP - findings*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Nils Reimers and Iryna Gurevych. 2019a. 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.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Peng Xu, Mostofa Ali Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. Controllable story generation with external knowledge using large-scale language models. In *EMNLP*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *EMNLP*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*.

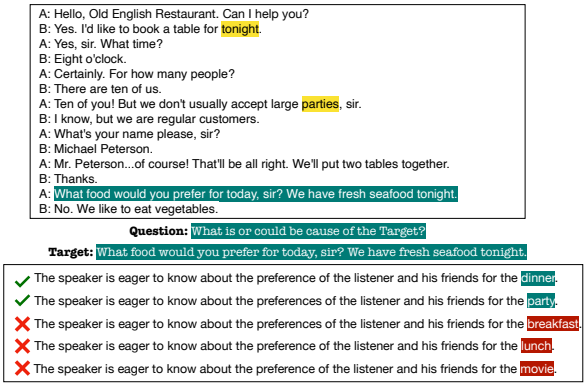
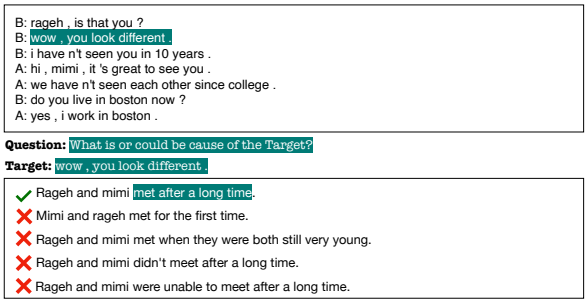
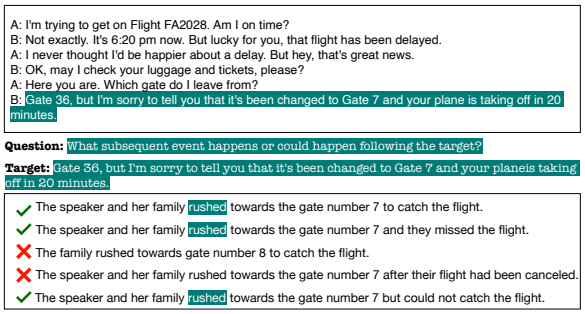


Figure 5: A data sample of CICERO for the CICERO$_{MCQ}$ task. Here, commonsense is required to infer the following events – booking a table at night implies the intention of having dinner.



(a)



(b)

Figure 6: Instances of `temporal commonsense` in CICERO.

A Additional Details on CICERO

The total compensation for the complete annotation process of CICERO including all the manual labeling (§2), and verification stages in AF (§3.2) was USD 13,500. The annotators were hired through a data annotation company. The total compensation was derived based on the country of residence of the annotators, as deemed by the company.

Being a dialogue-centric dataset, CICERO encompasses various aspects of human to human conversations such as temporal commonsense awareness in Fig. 5, Fig. 6, physical commonsense in Fig. 7, general commonsense in Fig. 8, and social commonsense in Fig. 10. In Fig. 6a, commonsense is required to infer that a familiar face may look different to us if we meet that person after a long time. There could be other potential reasons why a person might look different to his/her friends such as facial surgery, sickness, makeup, etc. However, in this particular dialogue context, the most appropriate speculative cause of the target is meeting the person after a long time. Similarly in Fig. 6b, the person hurries to the boarding gate as only 20 minutes is left before the flight takes off. Leveraging commonsense inference, we can infer that going to a place in a very short period requires us to rush. In Fig. 7, `physical commonsense knowledge` is required to infer — touching a hot element can burn our fingers and pans or microwaves are used for cooking.

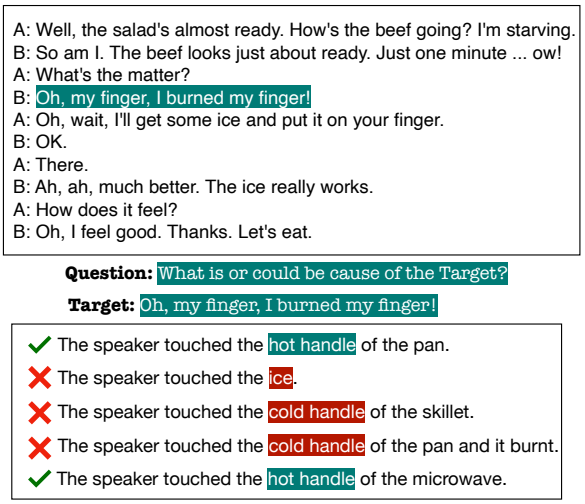


Figure 7: A data sample of CICERO where physical commonsense inference is prevalent.

B CICERO$_{NLG}$ Task: Extended Results

We report BLEU1 scores (Papineni et al., 2002) in addition to the automatic evaluation metrics described in §4.2. We also report results for generative tasks with the `BART-large` (Lewis et al., 2020), and `COMET` (Hwang et al., 2021) model. `COMET` is a commonsense generation model from free text input. It is a pre-trained `BART-large` model fine-tuned on the ATOMIC dataset (Hwang et al., 2021). In our work, we have used all the models in two distinct ways – i) with fine-tuning and ii) without fine-tuning on CICERO. The results are shown in Table 10, and Table 11. Surprisingly, despite being pre-trained on a large commonsense in-

	Model	BLEU1	BLEU2	METEOR	ROUGE	CIDEr	Sem-Sim
(1.1.1) Cause	T5	0.2874	0.1493	0.1630	0.2626	0.4560	0.6278
	BART	0.2542	0.1396	0.1527	0.2586	0.4241	0.6224
	COMET	0.2762	0.1518	0.1580	0.2652	0.4486	0.6253
	GLUCOSE-T5	**0.2935**	**0.1563**	**0.1634**	**0.2707**	**0.4915**	**0.6305**
	T5*	0.0137	0.0042	0.0200	0.0266	0.0237	0.3735
	BART*	0.0793	0.0053	0.0347	0.0872	0.0153	0.3181
	COMET*	0.0562	0.0216	0.0474	0.0902	0.0862	0.4402
	GLUCOSE-T5*	0.0654	0.0287	0.0560	0.0827	0.1332	0.4442
(1.1.2) SE	T5	**0.3083**	**0.1619**	**0.1662**	**0.2760**	**0.4119**	0.6276
	BART	0.2926	0.1484	0.1608	0.2670	0.3681	0.6166
	COMET	0.3053	0.1565	0.1588	0.2730	0.3850	0.6211
	GLUCOSE-T5	0.3000	0.1611	0.1628	0.2778	0.4430	**0.6297**
	T5*	0.0133	0.0045	0.0191	0.0264	0.0241	0.3865
	BART*	0.0823	0.0061	0.0345	0.0926	0.0140	0.3243
	COMET*	0.0567	0.0217	0.0472	0.0937	0.0884	0.4523
	GLUCOSE-T5*	0.0003	0.0001	0.0070	0.0024	0.0032	0.3073
(1.1.3) SE Clipped	T5	0.2889	0.1448	**0.1549**	0.2618	0.3099	**0.6123**
	BART	0.2651	0.1272	0.1384	0.2409	0.2765	0.5814
	COMET	**0.3023**	**0.1509**	0.1536	**0.2667**	0.3090	0.6083
	GLUCOSE-T5	0.2870	0.1461	0.1523	0.2645	**0.3238**	0.6094
	T5*	0.0559	0.0199	0.0439	0.0564	0.0762	0.4549
	BART*	0.0931	0.0067	0.0367	0.0869	0.0198	0.3541
	COMET*	0.0577	0.0215	0.0479	0.0953	0.0911	0.4583
	GLUCOSE-T5*	0.0003	0.0001	0.0066	0.0025	0.0034	0.3063
(1.2.3) Reaction	T5	**0.3410**	**0.2397**	**0.1939**	**0.3720**	0.5177	**0.6665**
	BART	0.3320	0.2297	0.1869	0.3531	0.4575	0.6575
	COMET	0.3338	0.2273	0.1815	0.3406	0.2662	0.6520
	GLUCOSE-T5	0.3283	0.2318	0.1903	0.3716	**0.5364**	0.6653
	T5*	0.0116	0.0037	0.0201	0.0239	0.0167	0.3899
	BART*	0.1815	0.0418	0.0913	0.1531	0.0194	0.5353
	COMET*	0.0590	0.0204	0.0454	0.0966	0.0653	0.4299
	GLUCOSE-T5*	0.0534	0.0213	0.0459	0.0759	0.0719	0.4125
(1.2.1) Prerequisite	T5	0.1826	0.1002	0.1282	**0.2176**	0.3357	**0.5902**
	BART	0.1817	0.1020	0.1260	0.2118	**0.3401**	0.5804
	COMET	**0.2115**	**0.1145**	0.1296	0.2168	0.3064	0.5815
	GLUCOSE-T5	0.1812	0.1001	**0.1299**	0.2197	0.3144	0.5896
	T5*	0.0177	0.0043	0.0222	0.0279	0.0225	0.3541
	BART*	0.0779	0.0065	0.0334	0.0827	0.0166	0.2913
	COMET*	0.0517	0.0186	0.0447	0.0782	0.0768	0.4281
	GLUCOSE-T5*	0.0259	0.0108	0.0394	0.0625	0.0889	0.4392
(1.2.2) Motivation	T5	0.3462	0.2503	0.1998	0.3781	0.7109	0.6973
	BART	0.3497	0.2482	0.1961	0.3709	0.6434	0.6914
	COMET	0.3428	0.2381	0.1935	0.3649	0.6286	0.6962
	GLUCOSE-T5	**0.3546**	**0.2582**	**0.2037**	**0.3840**	**0.7499**	**0.7048**
	T5*	0.0134	0.0033	0.0183	0.0257	0.0181	0.4038
	BART*	0.1072	0.0082	0.0416	0.1212	0.0164	0.3497
	COMET*	0.0582	0.0215	0.0475	0.0882	0.0782	0.4516
	GLUCOSE-T5*	0.0504	0.0174	0.0434	0.0632	0.0696	0.4053

Table 10: Results for Task 1. T5*, BART*, COMET* and GLUCOSE-T5* are not fine-tuned on CICERO. SE denotes Subsequent Event.

ference dataset, the fine-tuned COMET model fails to outperform both fine-tuned T5 and BART in most of the experiments. This could be due to catastrophic forgetting triggered by disparate inputs, which are at odds with ATOMIC. Further research is needed to draw any conclusion.

The results of human evaluation of the models

Model	BLEU1	BLEU2	METEOR	ROUGE	CIDEr	Sem-Sim
(1.1.4) Chained Cause						
T5	0.2781	0.1566	0.1675	0.2757	0.5303	0.6518
BART	0.1960	0.1104	0.1382	0.2242	0.4231	0.6074
COMET	**0.2893**	**0.1633**	0.1674	0.2742	0.5247	0.6488
GLUCOSE-T5	0.2820	0.1600	**0.1697**	**0.2796**	**0.5633**	**0.6557**
(1.1.1)* Cause						
T5	0.2884	0.1503	0.1635	0.2634	0.4591	0.6284
BART	0.2548	0.1400	0.1530	0.2590	0.4279	0.6225
COMET	0.2769	0.1522	0.1584	0.2654	0.4510	0.6257
GLUCOSE-T5	**0.2938**	**0.1564**	**0.1636**	**0.2709**	**0.4915**	**0.6310**
(1.1.5) Chained SE						
T5	**0.3322**	**0.1813**	**0.1784**	0.2940	0.5136	0.6469
BART	0.3131	0.1649	0.1672	0.2795	0.4106	0.6314
COMET	0.3057	0.1626	0.1673	0.2742	0.4515	0.6321
GLUCOSE-T5	0.3258	0.1789	0.1776	**0.2943**	**0.5218**	**0.6516**
(1.1.2)* SE						
T5	**0.3088**	**0.1622**	0.0841	0.2764	0.4167	0.6279
BART	0.2919	0.1490	0.1617	0.2667	0.3719	0.6165
COMET	0.3036	0.1557	0.1580	0.2727	0.3790	0.6187
GLUCOSE-T5	0.2998	0.1612	**0.1628**	**0.2778**	**0.4471**	**0.6294**

Table 11: Results for chained cause effect generation. (1.1.1)* and (1.1.2)* indicates results from Task 1.1.1, and 1.1.2 (as in Table 10), but only for target instances which have both cause and effect annotated, ensuring a fair comparison with (1.2). SE denotes Subsequent Event.

Model	Creativity	Contextuality	Fluency
Gold	4.7	4.8	5.0
T5	3.8	4.1	**4.9**
BART	3.6	**4.3**	**4.9**
COMET	3.8	4.1	4.8
GLUCOSE-T5	**3.9**	**4.3**	**4.9**
T5*	2.4	2.1	1.9
BART*	2.6	2.5	1.8
COMET*	2.2	2.3	2.5
GLUCOSE-T5*	1.9	2.1	2.9

Table 12: Results of the human evaluation for the CICERO$_{NLG}$ task. T5*, BART*, COMET*, and GLUCOSE-T5* represent non fine-tuned versions.

are illustrated in Table 12. It can be seen that all the models perform almost similarly on CICERO and stand far from reaching human-level performance.

Fine-tuned vs non Fine-tuned Evaluations. All the models perform very poorly when they are not fine-tuned on CICERO. The non fine-tuned models generate gibberish sentences across all five inference categories. The automatic and human evaluation results of these models are also reported in Table 10 and Table 12, respectively. The results confirm that fine-tuning is necessary for dialogue-level commonsense inference thus reaffirming the importance of our curated dataset CICERO. The non fine-tuned COMET produces very short outputs (1–3 words, akin to ATOMIC annotations) that are not readily comparable with CICERO, resulting in poor evaluation scores.

Finally, we provide some additional examples to depict the inference generation quality of the models in Table 13.

C CICERO$_{MCQ}$: Extended Results, Quantitative and Qualitative Analysis

For answer selection with generative models in CICERO$_{MCQ}$, we train T5 and Unified QA models under three distinct settings: 1) **Setting 1:** train models only on instances with a single-answer, 2) **Setting 2:** train models only on instances with multiple-answers, 3) **Setting 3:** train models on the

A (u_1): Hi, Jenny. Is it true you're moving to London? **B (u_2)**: Yes, it is. **A (u_3)**: What made you decide to do that? **B (u_4)**: Work, mainly. I'm sure I'll be able to find a job there. **A (u_5)**: You're probably right. But where are you going to live? **B (u_6)**: I hope I'll find a flat to share with somebody. That way it will be cheaper. **A (u_7)**: Yes, that's a good idea. Are you taking your dog with you? **B (u_8)**: No, I don't think so. My parents have offered to take care of him, and I don't think he'd be happy in the city. **A (u_9)**: You're probably right. But aren't you afraid of moving to such a big place, especially after living in a small village? **B (u_{10})**: Not really. I think I'll enjoy myself. There's so much to do there; I expect I won't miss the countryside much and I can always come back and visit. **A (u_{11})**: Well, I just hope you'll invite me to stay when you get settled. **B (u_{12})**: Of course I will.
Target - u_6; **Inference: Cause**; **Answers: Gold** - Being an expensive city, it is quite difficult to find an affordable place to live in London. **T5** - The listener asked Jenny where she was going to live. **COMET** - The speaker is looking for a flat to live in London. **GLUCOSE-T5-** Jenny has decided to move to London for her job.
Target - u_{10}; **Inference: Cause**; **Answers: Gold** - Jenny realizes that a city like London will provide a great quality of life for her. **T5** - The listener asked Jenny if she was afraid of moving to London after living in a small village. **COMET** - The speaker is moving to London for a job. **GLUCOSE-T5-** The listener asked Jenny if she was afraid of moving to such a big place.
Target - u_6; **Inference: Subsequent Event**; **Answers: Gold** - The listener suggests Jenny to find potential flats or flatmates online. **T5** - The speaker will find a flat to share with a friend. **COMET** - The speaker informed the listener that she will share the flat with someone else. **GLUCOSE-T5-** Jenny will find a flat to share with her friend.
Target - u_{10}; **Inference: Subsequent Event**; **Answers: Gold** - Jenny inquired a social club in London and ask for their membership to utilize her free time. **T5** - The speaker told the listener that he would love to visit London. **COMET** - The speaker informed the listener that he will miss the countryside very much. **GLUCOSE-T5-** The speaker informed the listener that he would love to come back to London.
Target - u_4; **Inference: Prerequisite**; **Answers: Gold** - Jenny has completed her studies. **T5** - The speaker has a job in London. **COMET** - Jenny has applied for a job in London. **GLUCOSE-T5-** The speaker has a job in London.
Target - u_{12}; **Inference: Prerequisite**; **Answers: Gold** - Jenny and the listener are good friends. **T5** - Jenny has invited her friend to stay with her in London. **COMET** - Jenny has a place to stay in London. **GLUCOSE-T5-**The listener invited Jenny to stay in London.
Target - u_6; **Inference: Motivation**; **Answers: Gold** - Jenny is optimistic about having someone as her flatmate to save on rent. **T5** - Jenny is hopeful of finding a flat to share with someone. **COMET** - Jenny is optimistic about having someone as her flatmate. **GLUCOSE-T5-** Jenny is hopeful that she will find a flat to share with somebody.
Target - u_{12}; **Inference: Reaction**; **Answers: Gold** - The listener is happy for Jenny and looks forward to being invited to London by Jenny. **T5** - The listener is happy for Jenny. **COMET** - The listener is happy to know that the speaker is moving to London. **GLUCOSE-T5-** The listener is excited to meet Jenny in London.

Table 13: Inferences extracted from a sample dialogue.

entire dataset comprising both single and multiple-answers.

The performances of both the generative models `T5` and `Unified QA` on instances with multiple answers are very poor (see Table 14, Table 15 and Fig. 11a, Fig. 11b). Further, we can also see instances where the predicted answers by these models contradict (see Fig. 11b). While `T5` surpasses `Unified QA` for Setting 3, `Unified QA` shines over `T5` for the other two settings.

Performance of `ELECTRA` vs `RoBERTa`. We also extend upon the results reported earlier for

Model	Trained On	Evaluated On	Cause	SE	Prereq.	Motiv.	Emo. Reac.	Avg.
RoBERTa	Single	Single	83.34	83.17	79.48	86.33	84.26	83.28
ELECTRA	Single	Single	**87.09**	**86.09**	**85.15**	**90.31**	**86.11**	**86.82**
T5	Single	Single	95.19	**95.29**	94.93	**96.52**	96.99	95.54
Unified QA	Single	Single	**95.85**	94.99	**95.55**	96.35	**97.22**	**95.70**
T5	Multiple	Multiple	20.04	20.45	15.94	25.26	26.72	20.62
Unified QA	Multiple	Multiple	**25.68**	**21.64**	**21.51**	**30.93**	**31.03**	**24.33**
T5	Single & Multiple	Single & Multiple	**78.18**	74.72	**75.50**	**82.51**	**84.59**	**77.68**
Unified QA	Single & Multiple	Single & Multiple	78.12	**74.79**	75.36	81.58	84.08	77.51
T5	Single & Multiple	Single	**93.20**	**91.28**	**91.27**	**95.19**	**95.14**	**92.71**
Unified QA	Single & Multiple	Single	93.12	91.16	91.00	94.28	94.79	92.45
T5	Single & Multiple	Multiple	**3.50**	2.77	3.59	**3.61**	**6.03**	3.38
Unified QA	Single & Multiple	Multiple	**3.50**	**3.69**	**3.98**	2.58	4.31	**3.60**

Table 14: Results of the CICERO$_{MCQ}$ task. SE denotes subsequent event. Single → Instances with single answer. Multiple → Instances with multiple answers.

Model	Trained On	Evaluated On	Cause	SE	Prereq.	Motiv.	Emo. Reac.	Avg.
RoBERTa	Single	Single	-	78.31	-	80.94	-	79.02
ELECTRA	Single	Single	-	**82.02**	-	**87.41**	-	**83.46**
T5	Single	Single	-	94.23	-	95.61	-	94.60
Unified QA	Single	Single	-	**94.38**	-	**96.19**	-	**94.87**
T5	Multiple	Multiple	-	16.49	-	24.23	-	18.07
Unified QA	Multiple	Multiple	-	**19.79**	-	**24.74**	-	**20.80**
T5	Single & Multiple	Single & Multiple	-	**74.99**	-	80.73	-	**76.46**
Unified QA	Single & Multiple	Single & Multiple	-	74.67	-	**80.80**	-	76.24
T5	Single & Multiple	Single	-	**91.95**	-	93.29	-	**92.31**
Unified QA	Single & Multiple	Single	-	91.43	-	**93.37**	-	91.95
T5	Single & Multiple	Multiple	-	1.32	-	**2.58**	-	1.58
Unified QA	Single & Multiple	Multiple	-	**1.85**	-	**2.58**	-	**2.00**

Table 15: Results of the CICERO$_{MCQ}$ task under the zero-shot setting. SE denotes subsequent event. Instance corresponding to cause, prerequisite, and emotional reaction are used for training. Instance corresponding to subsequent event and motivation are used for evaluation. Single → Instances with single answer. Multiple → Instances with multiple answers.

ELECTRA and RoBERTa in §4.3 for the single answer selection (Task 2.1) in CICERO$_{MCQ}$. The performance of ELECTRA is notably better than RoBERTa on this task. We reckon this could be due to the fact that we train our adversarial filtering (AF) method using RoBERTa. As such the efficacy of AF to prevent exposing stylistic artifacts to the discriminators is lesser for ELECTRA compared to RoBERTa. In other words, ELECTRA is more efficient than RoBERTa for the CICERO$_{MCQ}$ task due to its ability to better discriminate machine-generated negative answers from human-annotated true answers by leveraging stylistic artifacts as observed in Zellers et al. (2018).

Despite performing decently on the single answer selection task for CICERO$_{MCQ}$, RoBERTa does make mistakes in understanding some very interesting commonsense-based inferences such as the ones illustrated in Fig. 12. In these two examples, commonsense inference is required to detect the bluff by Tim Smith. Among other kinds of errors, we find RoBERTa failing to capture contextual commonsense cues such as in Fig. 9 — if a person wanting to buy new batteries is informed about the availability of batteries at photocopy stores, that person will search for photocopy stores instead of ad stores.

Zero-shot Setting. We also set up a zero-shot setting for Task 2.1 – Single Answer Selection and Task 2.2 – All Answers Selection. Under this setting, we only keep instances pertaining to cause, prerequisite, and emotional reaction in the train, validation data while instances with subsequent

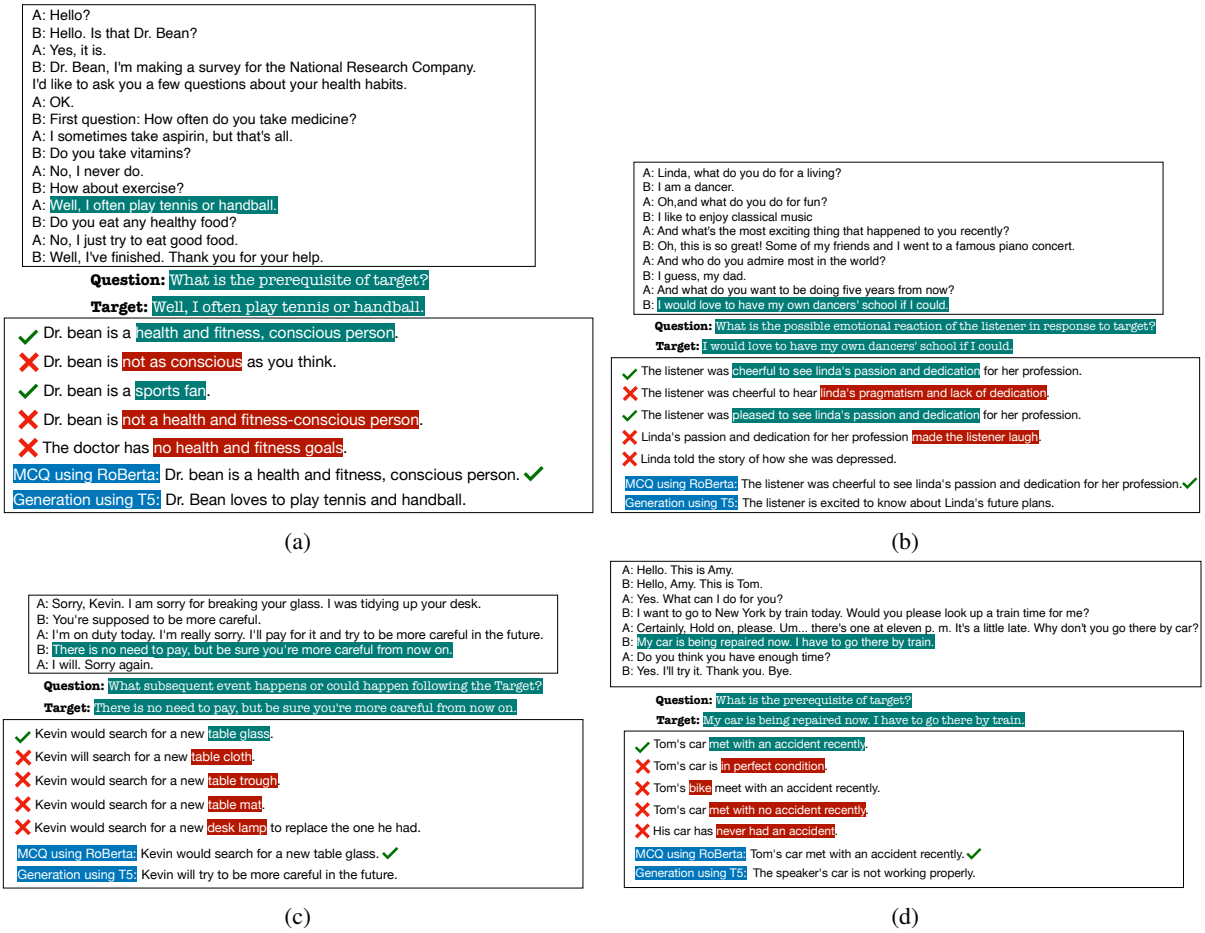


Figure 8: Instances of `general commonsense` in CICERO.

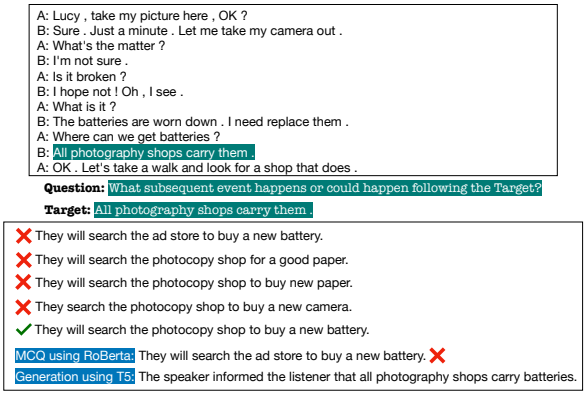
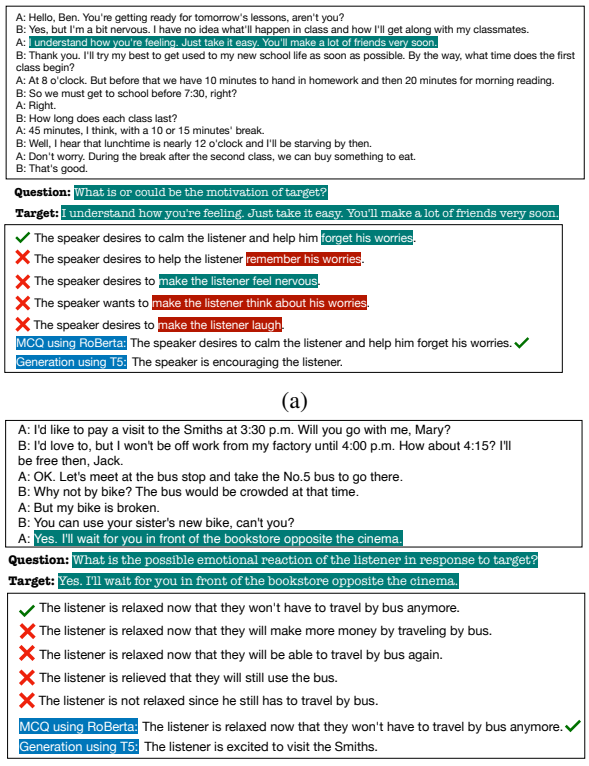


Figure 9: An instance where `RoBERTa` fails to capture the contextual commonsense cue.

event, and motivation are kept in the test data. All the models underperform in the zero-shot setting, as can be seen in Table 15. Like the all and single answer(s) prediction, `T5` and `Unified QA` perform similarly. On the other hand, `ELECTRA`'s zero-shot performance surpasses that of `RoBERTa`. Notably, performance of `T5` and `Unified QA` only drop around 1% in this setting, as compared to 3% drop observed for `RoBERTa` and `ELECTRA`. Hence, it is fair to conclude that for the CICERO$_{MCQ}$ task, `T5` and `Unified QA` are more robust to zero-shot scenarios than `RoBERTa` and `ELECTRA`. In the case of zero-shot single answer prediction, the best model is `Unified QA` which outperforms `RoBERTa` and `ELECTRA` by 11% and 15% respectively.

Performance on Single- vs Multi-answer Instances. It is evident from Tables 14 and 15, that in both regular and zero-shot settings, all the models exclusively trained on single- and multi-answer instances perform better on single- and multi-answer test instances, respectively, as compared to models trained on both types of instances. This is likely a side-effect of the data imbalance between the single- and multi-answer instances (~86/14%) in the training set which causes the scarce multi-answer instances to have confounding effect on the training process, degrading the performance on both types of test instances.

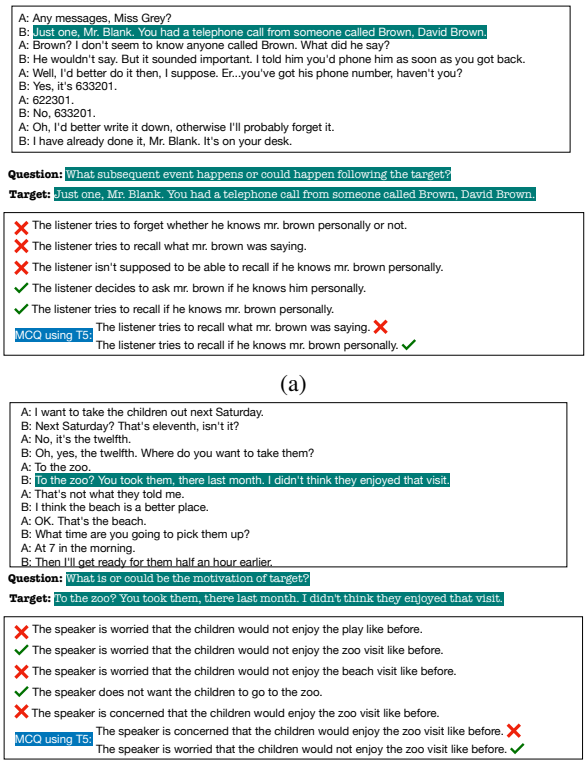
Performance of CICERO$_{NLG}$ vs CICERO$_{MCQ}$. We present the qualitative analysis for gen-



(a)

(b)

Figure 10: Instances of `social commonsense` in CICERO.



(a)

(b)

Figure 11: Multiple-answer predictions by `T5` for the CICERO$_{MCQ}$ task.

erative (CICERO$_{NLG}$) and discriminative (CICERO$_{MCQ}$) experiments in Fig. 8a, Fig. 8b, Fig. 8c, Fig. 8d, Fig. 9, Fig. 10a, and Fig. 10b. Except for Fig. 9, `RoBERTa` provides the accurate answer on all instances. Contrary to this, the performance of `T5` is far from being sublime on those samples for the CICERO$_{NLG}$ task. This depicts that the commonsense-based generative task CICERO$_{NLG}$ poses more challenge than the commonsense-based discriminative task CICERO$_{MCQ}$. We surmise this could happen due to two potential reasons —

1. Machine-generated negative answers may carry stylistic biases (Zellers et al., 2018), thus making the task of discriminators easier.

2. We collate the negative answers by generating counterfactual and contradictory sentences from the annotated true inferences. As a result, the generated negative answers are lexically very similar to the annotated sentences resulting in less diversity in the dataset.

Dataset	RoBERTa-Large
Swag	89.92
HellaSwag	85.20
α-NLI	83.91
Cosmos QA	82.25
Physical IQA	79.40
Social IQA	77.12
CICERO	83.28

Table 16: Results of baseline models in other CSK datasets.

D CICERO vs Other Commonsense Datasets

The key differences that set CICERO apart from the rest of the commonsense datasets are following:

- To the best of our knowledge, CICERO is the only publicly available dialogue-centric commonsense inference dataset.

- The speculative nature of the questions posed to the annotators enforces employment of rich commonsense knowledge in the inferences,

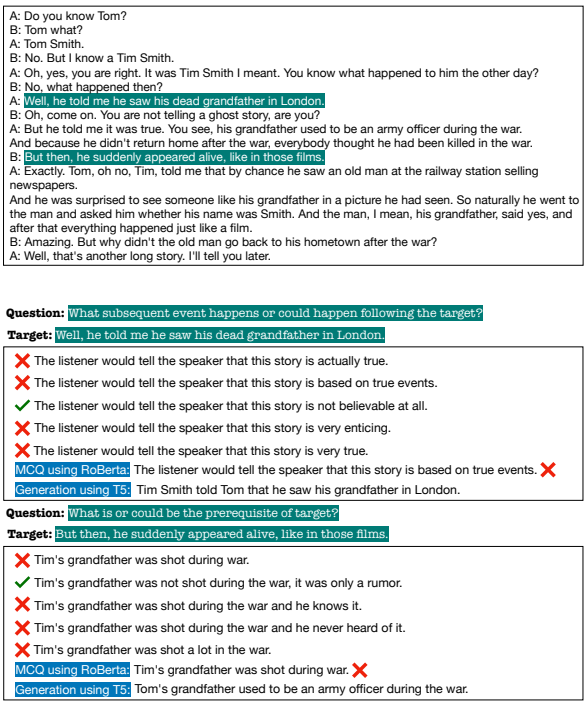


Figure 12: Examples of some incorrect predictions by RoBERTa for the CICERO$_{MCQ}$ task.

thereby, making CICERO commonsense-rich and, thus, difficult inferences for models without relevant commonsense knowledge.

- While the performance of the strong baseline models on CICERO for CICERO$_{MCQ}$ task are comparable (see Table 16) with the performance on other available commonsense-based question-answering datasets, unlike the others, around 14% of the instances in CICERO contain multiple correct inferences/answers. These are more challenging to the baselines, as can be seen in Table 14.

- Dialogue-centric commonsense inference/answer generation task, i.e., CICERO$_{NLG}$ is novel and hard to solve. Strong baselines, such as, T5, BART, and their checkpoints pre-trained on large external commonsense datasets, such as, ATOMIC and GLUCOSE, perform poorly at this task.

E Hyperparameter Details

All models for the CICERO$_{NLG}$ generative tasks were trained with the Adafactor optimizer (Shazeer and Stern, 2018) with a learning rate of 5e-6. The models CICERO$_{MCQ}$ alternative selection were trained with the AdamW (Loshchilov and Hutter, 2018) optimizer with a learning rate of 1e-5. We used a batch size of 4 for all our experiments.

F Computational Resources

The T5 Large and GLUCOSE-T5 Large have 770M parameters each. The RoBERTa-Large and ELECTRA-Large have 355M and 335M parameters, respectively. We also use a BART-Large and COMET-Large models for more extensive experiments (Appendix B). Both the models have 406M parameters. We use a single RTX 8000 GPU for our experiments. All models were trained for 5 epochs. Training and inference for the generative tasks i.e., CICERO$_{NLG}$ require between 1.5-6 hours in this GPU. Training and inference for the alternative selection task i.e., CICERO$_{MCQ}$ require a total of 15 hours. Training and inference times are 40% less for zero-shot setting experiments.