TIMEDIAL: Temporal Commonsense Reasoning in Dialog

Lianhui Qin◇*   **Aditya Gupta**♠   **Shyam Upadhyay**♠
Luheng He♠   **Yejin Choi**◇   **Manaal Faruqui**♠

♠Google Assistant
◇Paul G. Allen School of Computer Science & Engineering, University of Washington
{gaditya, shyamupa, luheng, mfaruqui}@google.com
{lianhuiq, yejin}@cs.washington.edu

Abstract

Everyday conversations require understanding everyday events, which in turn, requires understanding temporal commonsense concepts interwoven with those events. Despite recent progress with massive pre-trained language models (LMs) such as T5 and GPT-3, their capability of temporal reasoning in dialogs remains largely under-explored. In this paper, we present the first study to investigate pre-trained LMs for their temporal reasoning capabilities in dialogs by introducing a new task and a crowd-sourced English challenge set, TIMEDIAL. We formulate TIMEDIAL as a multiple choice cloze task with over 1.1K carefully curated dialogs. Empirical results demonstrate that even the best performing models struggle on this task compared to humans, with 23 absolute points of gap in accuracy. Furthermore, our analysis reveals that the models fail to reason about dialog context correctly; instead, they rely on shallow cues based on existing temporal patterns in context, motivating future research for modeling temporal concepts in text and robust contextual reasoning about them. The dataset is publicly available at: https://github.com/google-research-datasets/timedial.

1 Introduction

Humans can effortlessly reason about temporal concepts of everyday events such as their duration, frequency, or relative ordering (Allen, 1984; Radvansky and Zacks, 2014) based on rich commonsense knowledge about how the world works, especially in relation to time. However, reasoning about such concepts has been challenging for machines (Kahn and Gorry, 1977; Kozareva and Hovy, 2011) since it requires both understanding the local temporal expressions and reasoning about their global contexts such as their relative ordering and relations

---

A: May we see the wine list please.
B: Sure. Our special wine today is a 1989 Chardonnay.
A: **I'd like a bottle** please.
B: I'll need to see your ID please.
A: Here you go.
B: Sorry about the inconvenience, you look so young. I had to make sure you are over _____.

a) 21 years old ✓   b) 30 years old ✗
c) 4 years old ✗   d) 18 years old ✓

A: Good morning! May I help you?
B: Yes. My wife and I are interested in renting a house for the summer.
A: Very well. How long do you want the house? All summer?
B: No, not all summer. Just for **six weeks**.
A: I am afraid I can only rent it for **two months**.
B: My holiday is only _____, but I think my brother and his family would take it for the other **two weeks**.

a) six decades ✗   b) 45 days ✓
c) six weeks ✓   d) two months ✗

Table 1: Examples from our TIMEDIAL challenge set, demonstrating the need for commonsense knowledge and arithmetic reasoning over the context to infer the correct answers. Key contextual information for reasoning success is highlighted.

(UzZaman et al., 2013; Ning et al., 2018b; Pustejovsky, 2017). The problem becomes even more challenging in dialogs, where explicit and implicit inter-dependencies among temporal concepts can appear across conversation turns.

For instance, for the first dialog in Table 1, one must understand the context, i.e., selling wine, and use world knowledge of minimum legal drinking age in order to reason about correct answers to fill in the blank. Similarly, in the second conversation, commonsense about the durations *summer*, *month*, *week*, *day* and their relations, plus numerical reasoning, are necessary to make the inference.

Although previous works have studied temporal reasoning in natural language, they have either focused on specific time-related concepts in

---

*Work done during an internship at Google.

isolation, such as temporal ordering and relation extraction (Leeuwenberg and Moens, 2018; Ning et al., 2018a), and/or dealt with limited context, such as single-sentence-based question answering (Zhou et al., 2019) and natural language inference (Vashishtha et al., 2020; Mostafazadeh et al., 2016).

In this work, we make the first systematic study of temporal commonsense reasoning in a multi-turn dialog setting. The task involves complex reasoning that requires operations like comparison and arithmetic reasoning over temporal expressions and the need for commonsense and world knowledge.

We design a new task for dialog-based temporal reasoning and present a new challenge set in English, called TIMEDIAL, to evaluate language understanding models on the task. We formulate the problem as a crowd-sourced cloze task with multiple choices based on dialogs in the DailyDialog dataset (Li et al., 2017). Given a dialog with one temporal span masked out, the model is asked to find *all* correct answers from a list of four options to fill in the blank (Table 1).

The challenge set requires the models to demonstrate understanding of the context and use temporal commonsense to make right choices. Our final challenge set consists of 1.1K carefully curated dialog instances.

We then study the performance of several state-of-the-art pre-trained language models on TIMEDIAL along several dimensions including modeling paradigms (classification, mask filling, and generation), the scope of dialog contexts, in-domain vs. out-of-domain training, dependence on shallow text matching for reasoning, and the types of reasoning required. Our experiments demonstrate that off-the-shelf, pre-trained language models cannot effectively reason about temporal aspects in a dialog, even with domain-specific finetuning. Our findings indicate that large-scale pre-trained models even after fine-tuning may not be sufficient for robust temporal reasoning in dialogs, and motivate future research toward modeling temporal concepts over diverse everyday events, and contextual reasoning about them.

2 Task: Temporal Reasoning in Dialog

We formulate the dialog-based temporal commonsense reasoning problem as a *cloze* task (Taylor, 1953). Formally, given a multi-turn dialog context of n conversational turns between two speakers A

and B, where a temporal words span within the context is masked out, the task is to predict the suitable temporal expression(s) for the masked-out span from a list of options. That is, we want the conversation model to select all the correct answers from the options based on the dialog context. Following similar cloze-style challenge datasets, we use accuracy as the evaluation metric (Mostafazadeh et al., 2016; Onishi et al., 2016; Mihaylov and Frank, 2018).

Having a non-trivial set of options is crucial to build a challenge set and to avoid accidental spurious biases (Geirhos et al., 2020; Gururangan et al., 2018; Le Bras et al., 2020). We ensure this via the following filtering process. **(1)** For each masked span, there is more than one correct answer in the options. This makes the task more challenging for models since more comprehensive understanding of the context is required to recognize all the correct choices. In our dataset (§3) we guarantee two correct answers for each masked span. **(2)** Some incorrect options are selected to be spuriously correlated with the dialog context. For example, we include temporal spans in the dialog context as negative options, which will challenge models that rely primarily only on shallow pattern matching without correct temporal reasoning. We present more information in §3 about how the negative options were created by human annotators.

3 Dataset: TIMEDIAL

The TIMEDIAL dataset is derived from DailyDialog data (Li et al., 2017), which is a multi-turn dialog corpus containing over 13K English dialogs. Dialogs in this dataset consist of turn-taking between two people on topics over 10 broad categories, ranging from daily lives to financial topics.

3.1 Data Collection

Our data collection process involves two steps: (1) identifying dialogs that are rich in temporal expressions, and (2) asking human annotators to provide correct and incorrect options for cloze instances derived from these dialogs. We now describe these steps in detail.

Temporal expression identification. Here, we select dialogs that are rich with temporal information, in order to focus on complex temporal reasoning that arises in natural dialogs. Temporal expressions are automatically identified with SUTime (Chang and Manning, 2012), an off-the-shelf

Category	Dialog	Options
World Knowledge (5%)	A: May we see the wine list ? B: Sure . Our special wine today is a 1989 Chardonnay . A: That sounds pretty good! How much is it ? B: It's $4.25 cents by the glass . The whole bottle is $22.25 . A: **I'd like a bottle** please . B: I'll need to **see your ID** please . A: Here you go . B: Sorry about the inconvenience, I had make sure you are over ———— .	✓ 21 years old ✗ 30 years old ✗ 4 years old ✓ 18 years old
Comparison (24%)	A: Yes , sir. May I help you? B: Please I'd like a ticket to New York. A: For today? B: No, **early Saturday morning** . A: We have a flight that we'll put you there at ————. Is that ok? B: **Nothing earlier? I prefer flight at 9 thirty.** A: I'm afraid not , unless you want a night flight. B: No, exactly not.	✓ ten AM ✗ 9:30 PM ✓ eleven AM ✗ four AM
Arithmetic (5%)	A: How long do you want the house ? All summer ? B: No , just for six weeks. A: I'm afraid I can only **rent it for two months** . B: My holiday is only ————, but I think my brother and his family would take it for the **other two weeks** .	✗ six decades ✓ 45 days ✓ six weeks ✗ two months
General Commonsense (60%)	A: Do you get up early every morning ? B: About 6 in the morning. **I like to walk to the office** . A: Good habit. How long does it take ? B: ————. Do you live alone ? A: No , my little sister lives with me …	✓ 20 minutes ✗ 10 seconds ✓ 15 minutes ✗ 20 hours
Others (6%)	A: How long does a facial service take? B: We have half-hour and one-hour treatments. A: What's the regular price? B: Well , **the half-hour** facial costs $50 and **the one-hour** costs $80. A: Good , I will take ———— facial. B: That's fine , madam.	✓ the one hour ✗ the 20 hour ✗ the 80 second ✓ the half hour

Table 2: Example dialogs and answer options from the TIMEDIAL dataset, categorized by the nature of reasoning required to correctly answer them, along with the percentage of each reasoning category in the set of 100 sampled examples. The relevant key information in the dialog context is highlighted.

temporal expression detector.[1] We keep only the dialogs with more than 3 temporal expressions and at least one expression that contains **numerals** like "*two weeks*" (as opposed to non-numeric spans, like "*summer*", "*right now*", and "*later*"). In our initial experiment, we observe that language models can often correctly predict these non-numerical temporal phrases.

We note that temporal expressions containing numerals serve as more challenging sets of options than non-numerical ones. This filtering step results in 1,127 unique dialogs for further processing.

Human annotated options. Next, we make spans in the dialogs. For a dialog, we mask out each temporal expression that contains numerals, each resulting in a cloze question that is then sent for human annotation.

This resulted in 1,526 instances for annotation. For each masked span in each dialog, we obtain human annotation to derive a fixed set of correct and incorrect options given the context. Concretely, given a masked dialog and a seed correct answer (i.e., the original text) for the masked span, the

annotators[2] were asked to (1) come up with *an alternative correct answer* that makes sense in the dialog adhering to commonsense, and (2) formulate *two incorrect answers* that have no possibility of making sense in the dialog context. We highlight all time expressions in the context to make it easier for annotators to select reasonable time expressions.

To ensure that the annotated incorrect options are not too trivially distinguishable by the models (as discussed in §2), we define three rules for the annotators to follow.

- **Rule 1: Phrase Matching.** The rater should first try to pick another temporal span from the dialog context that makes syntactic/semantic sense (e.g., when the span is of the appropriate type, such as duration, for the masked span) but is still incorrect according to commonsense.
- **Rule 2: Numeral Matching.** If Rule 1 does not apply, raters should follow a relaxed version of Rule 1, whereby the incorrect option should contain any numeral occurring in the dialog context.

# Dialog instances	1, 104
# Temporal Expressions	1, 985
# Avg. Turns Per Dialog	11.7
# Avg. Words Per Turn	16.5
# Avg. Time Spans Per Dialog	3.0
Incorrect Options	
% Phrase Matching	16.3 %
% Numeral Matching	49.6 %
% Open-ended	45.4 %

Table 3: Statistics of our TIMEDIAL challenge set.

- **Rule 3: Open-ended.** If neither of the above rules is applicable, then raters can come up with an incorrect option using their own judgment. The two incorrect options are required to differ from each other as much as possible.

Rules-1&2 are designed to confuse models that rely on shallow pattern matching. Finally, to ensure the quality of the human-annotated options, we perform a subsequent round of human validation on the gathered data. The validators identify and fix issues such as duplicate options, unreasonable or obscure annotations w.r.t natural usage, or ungrammatical annotations that do not fit in the context.

3.2 Properties of TIMEDIAL

Table 3 shows statistics of TIMEDIAL. The dataset contains over 1.1K test instances. Each dialog contains 11.7 turns and 3 temporal expressions on average, presenting richer and more complex context compared to the recent single-sentence-based temporal question answering benchmarks (e.g., Zhou et al., 2019; Vashishtha et al., 2020). As above, each test instance contains two correct answers and two incorrect ones.[3] Over half of the incorrect options are annotated based on phrase and numeral matching from context, which pose a significant challenge for models relying on shallow text matching, as we show in our experimental analysis (§5).

Answering different instances in the dataset requires different types of core reasoning abilities, such as comparison, arithmetic inference, or reasoning based on world knowledge or general commonsense. To facilitate fine-grained analysis, we also annotate the **reasoning categories** for a randomly sampled set of 100 dialogs. Though each

---

[3]We also collected 342 extra instances for which the annotators deem there is only one unique correct answer for the context. Thus, each of those instances contains one correct option and two incorrect ones. We release those instances along with the dataset, though we did not include them in empirical study in this paper.

instance can involve multiple reasoning types, we associate it with one predefined category label that indicates the primary type of reasoning it requires. Table 2 shows the category distribution and examples in each of the category. We observe that the dataset requires general commonsense for 60% of the dialogs, making it the most common reasoning type.

4 Modeling

We consider a broad set of methods and evaluate their performance on our challenge TIMEDIAL dataset. These methods vary in terms of the modeling paradigms, the scope of the dialog contexts, and training settings. In particular, they encompass the major ways pre-trained LMs are currently used in downstream tasks (§4.1) which often outperform earlier specialized non-pretrained models. We also consider different lengths of context used in reasoning, varying by their vicinity to the masked span (§4.2). Finally, we study different training settings, including zero-shot, in-domain, and out-of-domain training (§4.3).

4.1 Modeling Paradigms

We experiment across three major modeling paradigms: (i) Binary Classification, (ii) Mask Filling, and (iii) Generation. Figure 1 shows the different architectures. For each test instance, the model takes as input a pair of (masked dialog context, candidate), and outputs a score measuring how likely the candidate being a correct answer. Based on the prediction scores of all options, the model then chooses the top two positive candidates as the predicted answer for the instance. Each paradigm of models is finetuned using training data from different domains, as discussed in §4.3.

4.1.1 Binary Classification

In this setting, we formulate the task as a binary classification problem, i.e., we use a classifier to measure the probability of the candidate in the (masked dialog context, candidate) pair being a correct answer. Any powerful LM — e.g., BERT (Devlin et al., 2019), ALBERT (Lan et al., 2019), ROBERTA (Liu et al., 2019), etc. can be used to build the classifier.

This method's key challenge is the lack of annotated training data for direct supervision. We generate weak supervision training data as follows. In an unlabeled corpus, we use the SUTime tool
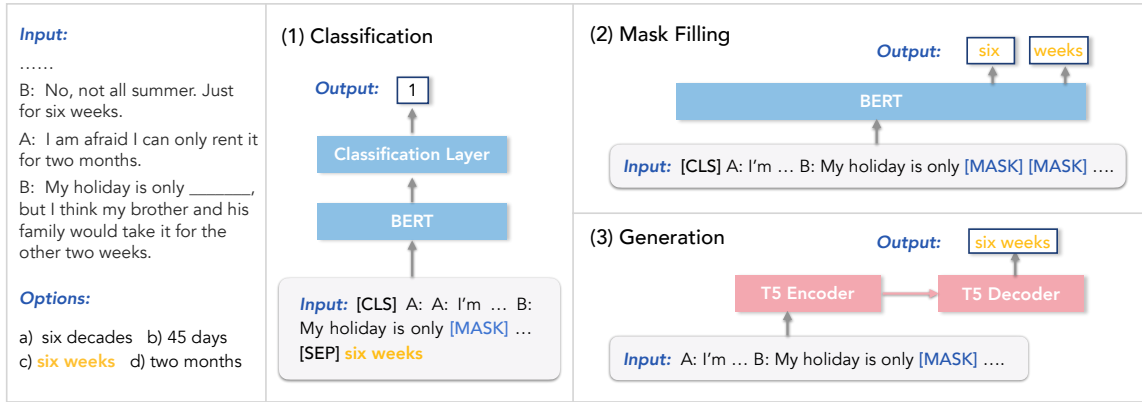
Figure 1: We study three modeling paradigms for the task, based on BERT and T5, including (1) Classification, (2) Mask Filling, and (3) Generation (§4.1). The models are finetuned with various training data, as discussed in §4.3.

to annotate temporal spans. We mask each temporal span in this corpus and use the masked text as one positive example for binary classification. To generate negative example, we randomly sample another temporal span from the dialog context and use it as a negative example for the masked temporal span. The resulting data is noisy because the randomly sampled temporal span can also logically fit in the masked span in the given context; however, we assume the likelihood of that happening is low. We leave drawing harder negative instances using heuristics to future work.

4.1.2 Mask Filling

We also use the mask filling approach of BERT-like mask language models (MLMs). For each dialog context and a candidate temporal span of m tokens, we replace the blank in the dialog context with m masked tokens. We then evaluate the likelihood of predicting the temporal span tokens for those masked positions, and make average across the positions. A key advantage of this method is that we can directly apply a BERT model in the *zero-shot* manner since the model was pre-trained in the same way, as for accommodating for [MASK] fillings. Additionally, we also finetune BERT's MLM for learning task specific properties.

4.1.3 Generation

The third method is a fully generative approach using the *text-to-text* paradigm of T5 (Raffel et al., 2020). Given a masked dialog context, the model is trained to generate the masked text in an encoder-decoder framework. As a result, evaluating the likelihood of generating the given temporal span (normalized with the length of the span) is used as the probability of it being correct. Similar to mask

filling, we use T5 either in a zero-shot manner or with additional fine-tuning.

4.2 Dialog Context

We aim to study the influence of context on a model's temporal reasoning in dialog by incorporating varying scopes of dialog context based on their vicinity to the target span. Since the dialogs in TIMEDIAL are rich in temporal concepts, we want to evaluate LMs' dependence on shallow text matching vs. the ability to accurately understand the causal relations between those concepts (see Table 6). We use the following three settings:

- **Full** context, where the model is presented with the complete available dialog to reason on. Due to our design of challenging negatives, the full context can often confuse models that rely on shallow cues.
- **Local** context, where we provide only with the utterances that immediately precede and follow the target utterance.
- **Target** context, where the context is restricted to only the particular utterance that contains the masked span.

4.3 Training Details

For all models, we consider two common training settings, e.g., in-domain data, which is typically small, and out-of-domain training where a large amount of data is available. Table 4 shows training data statistics. For mask-filling and generation, we also evaluate in a zero-shot setup with no finetuning.

In-domain training. Our challenge TIMEDIAL test set is derived from contextually rich dialogs

Mask Filling and Generation		
	# Train	# Dev
In-domain (Daily)	14.5K	2.4K
Out-domain (Meena)	1.26M	23K
Classification			
	# Train	# Dev	# Spans
In-domain (Daily)	58.0K	9.6K	2,153
Out-domain (Meena)	5.04M	92K	38,750

Table 4: Number of training and development instances for different settings. An instance is derived by masking one temporal span of a dialog. For classification, we draw 3 negative samples per positive sample. "# Spans" is the size of temporal span pool from which negative samples are drawn for weak supervision.

from the DailyDialog dataset, based on the number of temporal spans. However, this still leaves remaining data with less than 3 temporal spans or with no numeric span. By masking each temporal span in each dialog, we obtain 14.5K training instances to use in our domain specific fine-tuning.

Out-of-domain training. In this setting, we consider a much larger corpus from a general domain. Specifically, we use the large scale training set based on the Meena dataset Adiwardana et al. (2020), which is mined and filtered from public domain social media conversations over 341GB of text (40B words).[4] Compared to the above in-domain data from DailyDialog which were manually written by human annotators in a clean and consistent way, the dialogs in the Meena corpus tend to be noisy, casual, and usually short. Like our DailyDialog processing, we identify all temporal expressions for dialogs in Meena using SUTime.

5 Experiments and Analyses

Using the proposed TIMEDIAL challenge set, we next conduct extensive experiments and analyses on the different model variants and context settings. We use either 4x4 or 8x8 Cloud TPUs V3 pod slices[5] for fine-tuning and one V100 GPU for inference. We provide more details of the experiment configurations in the appendix.

Evaluation. Since each example of TIMEDIAL contains two correct answers, we report the metric 2-*best accuracy*, which measures whether *both* of the model's top-ranked answers are correct. In

---

[4]We acquired a trimmed down version of the Meena dataset by contacting the authors.
[5]https://cloud.google.com/tpu

SIZE-TRAIN	2-*best Acc (%)*
Classification (BERT)	
BASE-OUT	43.1
BASE-IN	51.1
LARGE-OUT	48.7
LARGE-IN	53.2
Mask Filling (BERT)	
BASE-ZERO	44.8
BASE-OUT	47.4
BASE-IN	67.4
LARGE-ZERO	47.7
LARGE-OUT	54.8
LARGE-IN	70.0
Generation (T5)	
BASE-ZERO	39.8
BASE-OUT	50.6
BASE-IN	59.2
LARGE-ZERO	39.1
LARGE-OUT	61.9
LARGE-IN	**74.8**
Human	**97.8**

Table 5: Model and human performance on TIMEDIAL. BASE and LARGE denote the size of the pre-trained BERT and T5; ZERO, IN, and OUT denote that the model is zero-shot (with no finetuning), fintuned using the in-domain DailyDialog data, or finetuned using the out-of-domain Meena data, respectively. THe full dialog context is used for all models.

other words, if the model erroneously ranks an incorrect answer over a correct one, we consider it to be an error case. Note that we use the ranking-based metric as opposed to classification-based ones (for example, by asking the model to classify whether each individual candidate answer is correct or not (e.g., Zhou et al., 2019)) and because it presents a stricter measure that penalizes any incorrect answers being ranked over correct answers, and the ranking metric is not influenced by specific choices of the threshold hyperparameter that cuts off positive and negative predictions.

5.1 Model Performance

Table 5 shows model results and human performance. Human performance achieves a near-perfect level (97.80, with Cohen's kappa score of 0.86 showing almost perfect inter-rater agreement (Landis and Koch, 1977)).

Overall. The generation model based on T5-LARGE and finetuned on the in-domain DailyDialog data achieves the best performance. However, its 2-best accuracy (74.8) lagged far behind the human performance, demonstrating the difficulty of the TIMEDIAL challenge set.

Dialog Context	Options	GOLD	CLS	MF	GEN
A: What's the date **today**? B: **Today** is **September 28th, 2007**. A: I have a meeting this **afternoon**. B: When will it begin? A: It will begin at **three o'clock**. What's the time **now**? B: It is _____. A: I have to go **now**. I don't want to be late. B: Don't worry, time is enough.	half past one	✓	✓	✗	✓
	quarter to two	✓	✗	✗	✗
	half past three	✗	✓	✓	✓
	half past nine	✗	✗	✓	✗
A: Doctor, I feel much better **now**. Will I be able to go home **some time this week**? B: That's good to hear. You've had an ideal recovery from your operation. We're going to send you home tomorrow. A: Do you think I can get back to work **very soon**? B:Don't be in such a hurry. I'm confident that you'll be completely recovered in _____. A: Is there anything I should do? B: You'd better have a good rest for **a week**.	4 to 6 weeks	✓	✗	✗	✓
	5 to 7 weeks	✓	✓	✗	✗
	a week	✗	✓	✓	✓
	a day	✗	✗	✓	✗

Table 6: Example prediction errors made by different models for cases with challenging options, based on the phrase and numeral matching rules (§3). GOLD denotes the true labels. The model predictions show that the models get confused by learning shallow text matching in terms of pre-existing temporal concepts (marked by bold faced text) in the context.

Zero-shot vs. out-of-domain vs. in-domain. When comparing the different training data setup, we observe that models with in-domain training using the DailyDialog data (e.g., LARGE-IN) consistently outperforms those trained on the large out-of-domain Meena dataset (e.g., LARGE-OUT). Both setups outperform the zero-shot models (without any fine-tuning) (e.g., LARGE-ZERO). The results show that the large LMs still highly depend on in-domain or at least dialog data to grasp and enhance their temporal reasoning ability in dialog context. Further, we see increasing performance with increasing model size, which is not unexpected given the complexity of the task.

5.2 Error Analysis

Next, we analyze the different types of errors based on different rules for negative option creation in the annotation process. In particular, the *phrase matching* rule picks an exact time span from the dialog context, and *numeral matching* picks numerals from the dialog context. Thus, models picking those incorrect options imply reliance on spurious shallow text matching features.

Figure 2 shows the percentage of errors in terms of the different rules. For example, the BERT-based classification model CLS-IN erroneously picks 52% of negative options created by the phrase matching rule as correct answers (i.e., by ranking those negative options over the true correct options). We observe that the various models are all most vulnerable to the phrase matching options compared
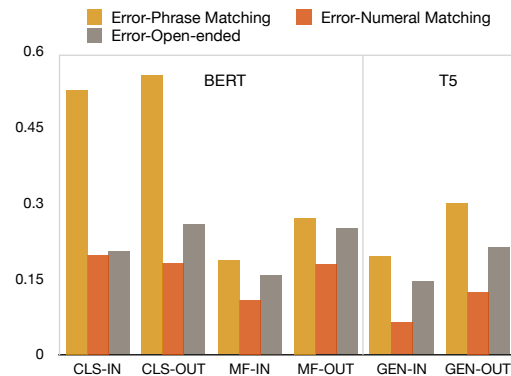

Figure 2: Percentage of errors on options created by different rules. CLS, MF, and GEN represent classification, mask-filling, and generation models, respectively; and IN and OUT denote in-domain and out-of-domain training. All models are of large size.

to other types of negative options, showing that they rely on spurious text matching to a significant extent. Between BERT and T5, we find T5 being more robust to shallow text matching.

Table 6 provides further examples of prediction errors, illustrating confusions due to shallow text matching. In the first dialog, both incorrect answers already partially occur in the context or are related to preexisting concepts (i.e., *"three"* to *"three o'clock"*, and *"nine"* to *"September"*). All the three models were confused and chose either of the two as the top prediction for the blank, even though the options clearly violate the context. Interestingly, the mask filling model was completely confused and ranked both incorrect answers over the correct ones. Similarly in the second example,

Size	BASE		LARGE	
Training	IN	OUT	IN	OUT
Classification (BERT)				
TARGET	50.5	40.0	50.5	47.5
LOCAL	+ 3.4	+ 3.3	+ 7.5	+ 2.0
FULL	− 0.6	− 0.1	+ 2.7	+ 1.2
Mask Filling (BERT)				
TARGET	57.8	44.3	60.3	46.8
LOCAL	+ 5.4	+ 3.0	+ 8.1	+ 4.9
FULL	+ 9.6	+ 3.1	+ 9.6	+ 8.0
Generation (T5)				
TARGET	55.5	45.9	66.7	56.1
LOCAL	+ 3.7	+ 2.7	+ 6.1	+ 3.7
FULL	+ 3.7	+ 4.7	+ 8.2	+ 5.8

Table 7: Impact of dialog context on reasoning accuracy. IN and OUT denote in-domain and out-of-domain training, respectively. We use *2-best accuracy* of *target* context as reference and report the absolute changes in performance of *local* and *full* context, respectively. Local dialog context results in better performance to full dialog context on 5 of the 12 cases, which are highlighted in the table.

the models fail to capture the contextual semantics.

5.3 Influence of Dialog Context

Table 7 shows how different scopes of dialog context (§4.2) affect model performance. First, the most restrictive target-only context is insufficient for accurate reasoning, by producing the weakest performance of most models. This highlights the importance of context information for temporal commonsense reasoning in dialog, which differs from previous temporal reasoning studies based on limited context (e.g., single-sentence question answering). Second, we note that the full dialog context does not always lead to the best performance. In 5 out of the 12 cases, using the local context yields equal or higher reasoning accuracy. The results show that the LMs still fall short of properly modeling the rich dialog contexts and making effective use of all information to do reasoning.

5.4 Errors of Reasoning Categories

Figure 3 shows the percentage of errors in each reasoning category. We observe that the models tend to make non-trivial portions of errors on commonsense/world knowledge questions. For example, the strongest model, T5 GEN-IN, failed on 18% of the instances that require commonsense or world knowledge, while BERT CLS-IN made errors on 48% of such instances. The performance
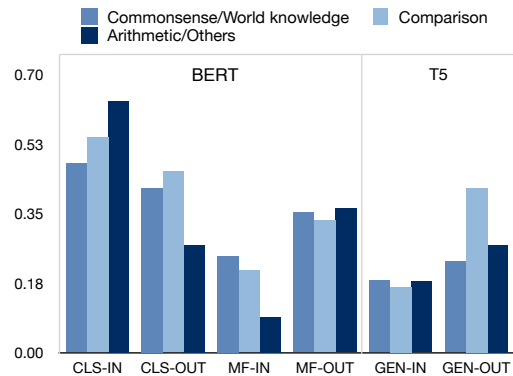

Figure 3: Percentage of errors on different reasoning types. CLS, MF, and GEN represent classification, mask-filling, and generation models, respectively. All models are of large size.

on comparison-based instances seems similar.

6 Related Work

Temporal commonsense reasoning. Early studies related to temporal analysis define time in the context of sets and relations (Bruce, 1972; Allen, 1983). More recent works often associate time with events and focus on identifying time expressions (Chang and Manning, 2012; Angeli et al., 2012; Lee et al., 2014), extracting temporal relations among events (Setzer and Gaizauskas, 2000; Pustejovsky et al., 2005; Lapata and Lascarides, 2006; Chambers et al., 2007; Ning et al., 2018b), and timeline construction (Do et al., 2012; Leeuwenberg and Moens, 2018).

Some recent work has focused on building challenging benchmarks for temporal commonsense reasoning. Story Cloze Test focuses on stereotypical causal temporal and causal relations between events (Mostafazadeh et al., 2016). Vashishtha et al. (2020) recast temporal reasoning datasets for event duration and event ordering into the natural language inference (NLI) format. Turque (Ning et al., 2020) is an reading comprehension dataset where the model needs to answer questions such as "what happens before/after [event]". Most related to our work is McTaco (Zhou et al., 2019), a dataset for evaluating temporal commonsense in the form of multiple-choice reading comprehension, where the context usually consists of a single sentence. Our work instead studies temporal commonsense reasoning in dialogs which often require significant commonsense and world knowledge to reason over rich context (Qin et al., 2019b; Dinan et al., 2018).

Commonsense reasoning with LMs. With the recent success of large pre-trained language models

(LMs) (Devlin et al., 2019; Brown et al., 2020), it is an open question whether these models, pretrained on large amounts of data, capture commonsense knowledge. Several works have been proposed to assess the ability of LMs for commonsense or numerical reasoning (Zhang et al., 2020; Bouraoui et al., 2020), or to mine commonsense knowledge from LMs (Davison et al., 2019). Lin et al. (2020) showed that state-of-the-art LMs such as BERT and RoBERTa performs poorly on numerical reasoning tasks without any finetuning. Works have also been proposed to improve language model's commonsense reasoning (Qin et al., 2020, 2019a; Zhou et al., 2020) and numerical reasoning abilities (Geva et al., 2020). In our work, we study several modeling approaches and finetuning settings of large LMs, and establish strong baselines for temporal commonsense reasoning in dialogs.

7 Conclusions

We introduced TimeDial, a challenge set consisting of 1.1K multiple-choice cloze questions for temporal commonsense reasoning in dialog. The dataset is carefully curated to evaluate a models' ability to do temporal commonsense/numerical reasoning over dialog context. In order to establish strong baselines and provide information on future model development, we conducted extensive experiments with state-of-the-art language models with different settings: the scope of context, weak supervision strategies, and learning objectives. While humans can easily answer these questions (97.8% accuracy), even our best model variant (T5-large with in-domain training) struggles on this challenge set (73%). Moreover, our qualitative error analyses show that these large language models often rely on shallow, spurious features (particularly text matching) when answering these questions, instead of truly doing reasoning over the context.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

James F Allen. 1984. Towards a general theory of action and time. *Artificial intelligence*, 23(2):123–154.

Gabor Angeli, Christopher D Manning, and Dan Jurafsky. 2012. Parsing time: Learning to interpret time expressions. In *Proc. of NAACL*.

Zied Bouraoui, José Camacho-Collados, and S. Schockaert. 2020. Inducing relational knowledge from bert. In *Proc. of AAAI*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam Mc-Candlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Bertram C Bruce. 1972. A model for temporal references and its application in a question answering program. *Artificial intelligence*, 3:1–25.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. Classifying temporal relations between events. In *Proc. of ACL*.

Angel X Chang and Christopher D Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In *Proc. of LREC*.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proc. of EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proc. of ICLR*.

Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proc. of EMNLP*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proc. of ACL*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proc. of NAACL*, pages 107–112.

Kenneth Kahn and G.Anthony Gorry. 1977. Mechanizing temporal knowledge. *Artificial Intelligence*, 9(1):87 – 108.

Z. Kozareva and E. Hovy. 2011. Learning temporal information for states and events. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 424–429.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Mirella Lapata and Alex Lascarides. 2006. Learning sentence-internal temporal relations. *Journal of Artificial Intelligence Research*, 27:85–117.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *Proc. of ICML*, pages 1078–1088. PMLR.

Kenton Lee, Yoav Artzi, Jesse Dodge, and Luke Zettlemoyer. 2014. Context-dependent semantic parsing for time expressions. In *Proc. of ACL*.

Artuur Leeuwenberg and Marie Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In *Proc. of EMNLP*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proc. of EMNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proc. of ACL*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proc. of NAACL*.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proc. of ACL*.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proc. of EMNLP*.

Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018b. Improving temporal relation extraction with a globally acquired statistical resource. In *Proc. of NAACL*.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proc. of EMNLP*.

James Pustejovsky. 2017. Iso-timeml and the annotation of temporal information. In *Handbook of Linguistic Annotation*, pages 941–968. Springer.

James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Saurí. 2005. Temporal and event information in natural language text. *Language resources and evaluation*, 39(2):123–164.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019a. Counterfactual story reasoning and generation. In *Proc. of EMNLP*.

Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, William B Dolan, Yejin Choi, and Jianfeng Gao. 2019b. Conversing by reading: Contentful neural conversation with on-demand machine reading. In *Proc. of ACL*.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Backpropagation-based decoding for unsupervised counterfactual and abductive reasoning. In *Proc. of EMNLP*.

Gabriel A Radvansky and Jeffrey M Zacks. 2014. *Event cognition*. Oxford University Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Andrea Setzer and Robert J Gaizauskas. 2000. Annotating events and temporal information in newswire texts. In *Proc. of LREC*.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Proc. of SemEval*.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Proc. of Findings of EMNLP*.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Proc. of Findings of EMNLP*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proc. of EMNLP*.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proc. of ACL*.