Feroze *et al. Intell. Robot.* **2025**, *5*(1), 228-47
DOI: 10.20517/ir.2025.12

Intelligence & Robotics

Research Article

Open Access

Enhancing temporal commonsense understanding using disentangled attention-based method with a hybrid data framework

Wasif Feroze[1], Muhammad Shahid[2], Shaohuan Cheng[1], Elias Lemuye Jimale[1], Yi Yang[3], Hong Qu[1,4], Yulin Wang[1,4]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, Sichuan, China.
[2]Department of Computer Sciences, Bahria University, Islamabad, Punjab 54600, Pakistan.
[3]Key Laboratory of Maritime Intelligent Cyberspace Technology (Hohai University), Ministry of Education, Changzhou 213200, Jiangsu, China.
[4]Tianfu Jiangxi Laboratory, Chengdu 641419, Sichuan, China.

Correspondence to: Prof. Yulin Wang, School of Computer Science and Engineering, University of Electronic Science and Technology of China, No.2006, Xiyuan Ave, West Hi-Tech Zone, Chengdu 611731, Sichuan, China. E-mail: wyl@uestc.edu.cn

How to cite this article: Feroze, W.; Shahid, M.; Cheng, S.; Jimale, E. L.; Yang, Y.; Qu, H.; Wang, Y. Enhancing temporal commonsense understanding using disentangled attention-based method with a hybrid data framework. *Intell. Robot.* **2025**, *5*(1), 228-47. http://dx.doi.org/10.20517/ir.2025.12

Received: 30 Nov 2024 **First Decision:** 23 Jan 2025 **Revised:** 5 Feb 2025 **Accepted:** 11 Feb 2025 **Published:** 19 Mar 2025

Academic Editor: Jinhai Li **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Understanding and capturing temporal relationships between time-related events expressed in text is a crucial aspect of natural language understanding (NLU). Although transformer-based pre-trained language models such as bidirectional encoder representations from transformers (BERT) have achieved significant success in various natural language processing (NLP) tasks, they are still believed to underperform in temporal commonsense tasks due to the limitation of vanilla self-attention. This paper proposes a methodology for developing language models to understand temporal commonsense reasoning over several tasks better. The proposed framework integrates a multi-data hybrid curation approach for dataset preparation, a collaborative synthetic dataset generation process involving chat agents and human domain experts, and a multi-stage fine-tuning strategy that leverages curated, intermediate, and target datasets to enhance temporal commonsense reasoning capabilities. The models we use in our proposed methodology are superior due to the use of an advanced attention mechanism and effective utilization of our framework. These models utilize disentangled attention, which is relative encoding position, which proved crucial for temporal commonsense by understanding temporal cues and indicators efficiently. Our extensive experiments show that models built with our proposed methodology enhance results on several temporal commonsense categories. Our results

show that we achieved better performance than the previous published work by utilizing a disentangled attention mechanism and hybrid data framework. Most impressively, our approach has demonstrated state-of-the-art (SOTA) results, surpassing all previous studies on temporal commonsense for the MC-TACO dataset.

Keywords: Temporal commonsense, commonsense reasoning, disentangled attention mechanism, large language models, data augmentation

1. INTRODUCTION

Commonsense is the basic unit of common knowledge among most people, which humans develop over time without explicit learning, and it is essential for understanding human language text. In natural language processing (NLP), commonsense reasoning is critical to discuss and develop methods that improve the commonsense ability of machine learning models[1]. Effectively understanding natural language text requires commonsense knowledge; recently, research focusing on NLP has increased in this direction, with Zhang *et al.*[2] developing a framework for extracting and analyzing ordinal commonsense knowledge, Yang *et al.*[3] introducing a model that enhances commonsense question answering by leveraging implicit relations in knowledge graphs, Wang *et al.*[4] proposing a lexical taxonomy-based method for inferring isA commonsense knowledge with high accuracy and Nguyen *et al.*[5] constructing a large-scale refined commonsense knowledge base with improved precision and recall. There are several commonsense reasoning types in NLP and applications of reasoning in machine learning[6], and our work investigates temporal commonsense, a special case of commonsense reasoning used to study time and its reasoning aspects. Temporal commonsense involves reasoning about time-related concepts such as event duration ("It takes 30 min to brew coffee"), ordering ("Brushing teeth happens after waking up"), and typical times ("Schools open at 8 AM"). Understanding time, whether stated explicitly or implicitly in any form of data, is very easy for humans; however, it is difficult for computers to comprehend temporal aspects in text and do reasoning in temporal commonsense. A few examples are presented in Figure 1 to show temporal commonsense in text data.

Over the years, NLP research has made significant progress in understanding time[7]; it is crucial for event extraction and relationships, which have many applications for deep learning[8,9]. Initially, the focus was on temporal information retrieval and temporal reasoning[10], which aimed to extract and organize time-related data from text, but now researchers have shifted their attention towards temporal commonsense reasoning[11], the latest form that emphasizes understanding how temporal events relate to everyday knowledge. While most of the research on temporal reasoning focuses on the relationship between temporal events and event extraction[12], some researchers also explore specific aspects individually, such as event duration[13], frequency[14], ordering[15], and infilling[16]. Various masking techniques have recently been used for intermediate training, followed by fine-tuning on temporal commonsense data, resulting in significant progress in current state-of-the-art (SOTA) results[17].

Temporal commonsense reasoning has been the subject of many research studies, but most have only focused on specific aspects of events individually, such as duration, ordering, or infilling. On the other hand, our work is unique in that it examines all categories of temporal commonsense, similar to the work by Zhou *et al.*[11]. Finding time-related events stated in the text and deriving temporal commonsense relations is crucial to comprehending sentences in natural language understanding (NLU). Earlier studies have utilized bidirectional encoder representations from transformers (BERT)[18] and RoBERTa[19] models but have not shown effectiveness in generalizing temporal commonsense tasks. While BERT and RoBERTa have achieved SOTA performance on numerous NLP benchmarks, their pre-training objectives do not explicitly capture dynamic temporal relations, which leads to diminished effectiveness on temporal commonsense tasks[20] as models were not able to identify simple temporal distinctions such "before" and "after" on straightforward sentiment

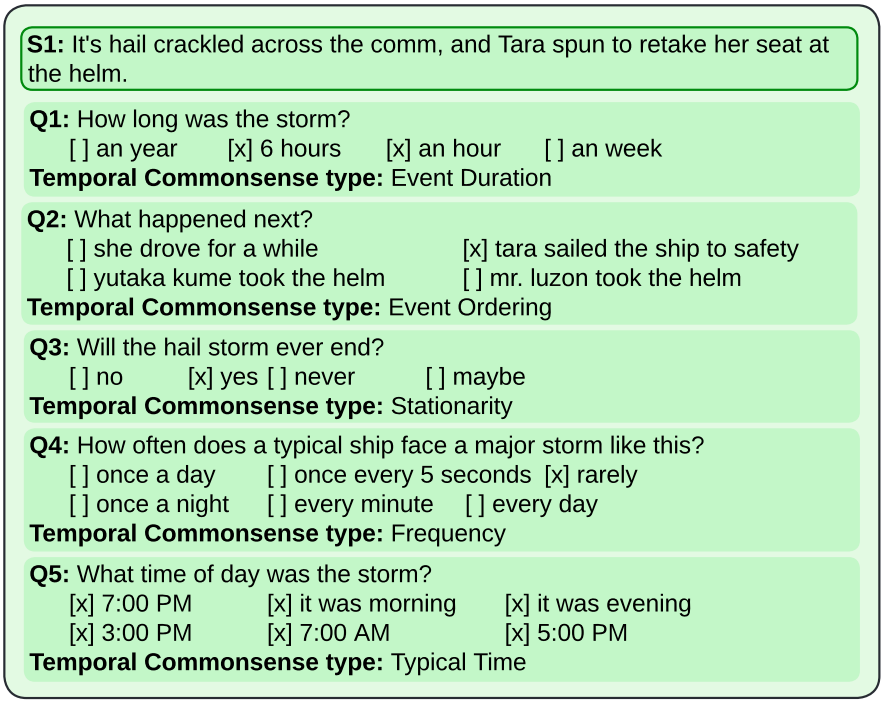

Figure 1. Sample examples for temporal commonsense.

analysis data. Kadari *et al.*[21] have tried to address this issue using more recent language models such as T5[22], a text-to-text pre-trained model; however, their approach only involved different preprocessing techniques such as time unit and duration unit normalization, which has limited effectiveness for large pre-trained models. For adversarial methods, ALICE[23] is improving the generalization of pre-trained language models on commonsense inference tasks by enhancing their robustness in the embedding space using the true label and model prediction for perturbations. ALICE++[24] is an improved method that employs a dynamic search algorithm for the best layers to use perturbations compared to the embedding layer in the former approach. The ML-ALICE[25] method introduces perturbations to the attention representation of layers. However, adversarial training methods have been ineffective due to their increased training time compared to standard fine-tuning approaches and the need for extensive human labeling for perturbation data.

Limitations of prior work highlight critical gaps in temporal commonsense research. Therefore, our focus to address these gaps is on developing computationally efficient methods and presenting the latest findings in this domain. Our objective is to explore novel approaches that can effectively generalize temporal commonsense tasks without requiring extensive human labeling or computational resources and advance this important area of research, which has significant importance for various fields. Our approach introduces a hybrid data framework and uses an advanced attention mechanism to achieve these goals.

This paper presents our approach to understanding temporal commonsense reasoning using DeBERTa models[26]. We employ various training and fine-tuning methods, including multi-stage fine-tuning, with augmented data generated by different methods. Our methodology integrates advanced techniques for data curation, synthetic dataset generation, and multi-stage fine-tuning, which elevate the model's performance on temporal commonsense tasks. After conducting extensive experiments and evaluating our technique against standard question-answering metrics, our experimental results show SOTA performance, exceeding all previous studies on temporal commonsense with the MC-TACO dataset. Following are contributions of our research to understanding temporal commonsense using disentangled attention mechanism-based methods:

- We propose a multi-data hybrid curation framework, which involves the preparation of several sets of diverse datasets for temporal commonsense reasoning, significantly improving model training and performance.
- We introduce an innovative process for generating a synthetic dataset through the hybrid collaboration of chat agents and human experts, which enriches the model's ability to reason on time-based events by providing high-quality, diverse data.
- We perform multi-stage fine-tuning of DeBERTa models, where the model is fine-tuned first on curated datasets, then on intermediate datasets, and finally on our target dataset, which enables the model to better generalize across temporal commonsense reasoning tasks.
- We fine-tune four variants of the DeBERTa pre-trained model using both the target dataset and synthetic datasets, with a detailed comparison of performance between the different fine-tuned models, demonstrating the impact of the data augmentation and fine-tuning strategies on model accuracy.

2. RELATED WORK

Commonsense research has recently become increasingly popular in the NLP research community. While earlier research was primarily focused on physical commonsense, researchers have expanded their investigations to include a broader range of topics. Forbes looked into commonsense reasoning related to size, weight, and strength in the physical world[27]. Cocos explored concepts such as the intensity, roundness, and deliciousness of objects[28]. Other researchers have looked at commonsense knowledge related to events, including the intentions and reactions of participants[29] and subsequent event selection[30]. Despite the importance of temporal commonsense for understanding event reasoning, very few studies have focused on it. Ning emphasized the importance of understanding event duration for constructing a story timeline[31].

Understanding time and temporal inference in natural language has captured the attention of researchers for an extensive period. Previous studies in this field have included the extraction of temporal expressions[32], extraction of temporal relations[33], and timeline construction[34]. Other works have focused on implicit temporal commonsense, such as event ordering and understanding what happens after an event[15,35], as well as event duration[32,36,37]. One study has explored five types of temporal commonsense together and presented them as a unified framework rather than as individual aspects of temporal commonsense[11]. This work has tailored temporal commonsense as reading comprehension and presents a dataset as question-answering tasks. The field of question-answering has seen steady research progress in the NLP community, with a focus on general comprehension of text[38–40].

Several challenging benchmarks have been recently developed for temporal commonsense inference. The Story Cloze Test[41] dataset discusses typical temporal ordering and event causal relationships. TORQUE[42] dataset presents temporal order of events as a reading comprehension task in NLP. MC-TACO[11] is the only benchmark that focuses on five complex temporal commonsense categories altogether. TIMEDIAL[43] is a dialogues dataset comprising multi-turn dialogues with complex temporal information. The construction of language models for temporal reasoning tasks is proposed by Zhou *et al.*, which presents and produces the representation of time and other events relevant to tasks[14].

Data augmentation is a process to expand available data without collecting new samples, initially proposed and used in the computer vision field. Later, it also proved significant in NLP despite the challenges posed by the discrete nature of languages to maintain invariance by continuous noising[44]. There are several approaches for text data augmentation; we will discuss the background of two methods, synonym replacement (SR), and back-translation (BT), for the relevant scope of this research. Kobayashi uses predictive language models to replace relevant words with the most similar words, which does not change the original labels[45]. Other data augmentation methods for substituting synonyms utilize well-known ontologies[46] or word similarity

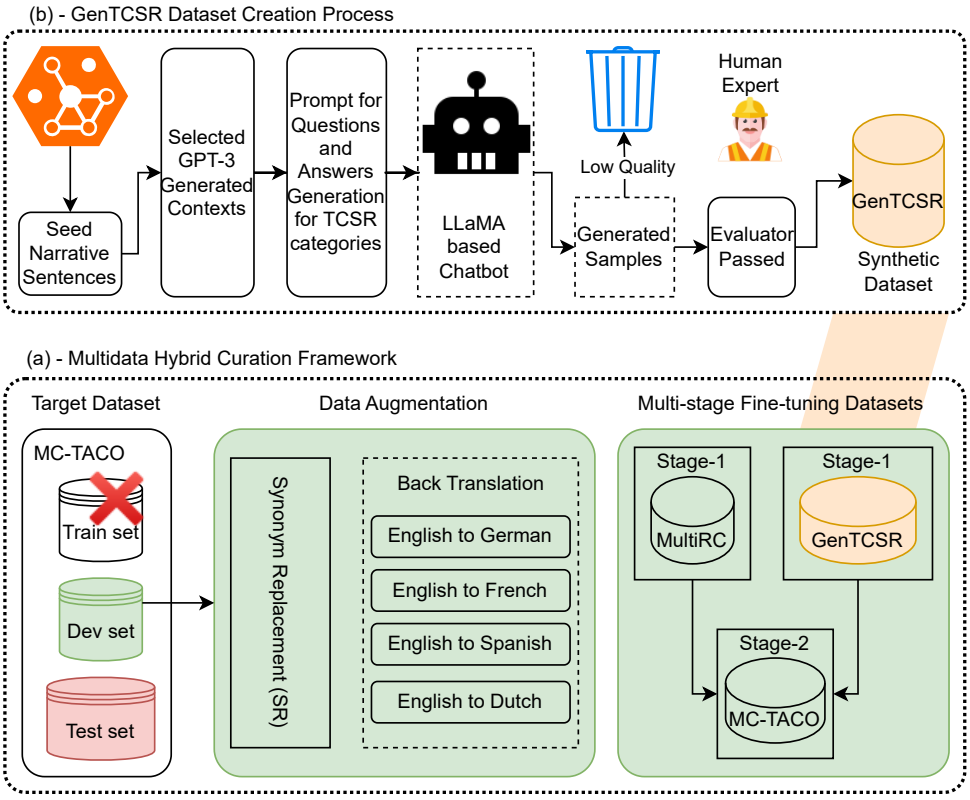

Figure 2. Overall architecture of proposed framework for temporal commonsense reasoning.

calculations[47]. The technique of using machine translation to generate new data from monolingual source data was first introduced by Sennrich to train neural machine translation models[48]. Yu *et al.* expanded the target data by generating new data by translating it into French and then back to English for BT data augmentation[49]. Fadaee uses BT to expand data and improve model performance without changing the training process of machine translation models[50]. Sugiyama proposes a method to prove the advantages of BT-based data augmentation for neural machine translation, which is context-aware translation and uses different language directions for BT[51]. Despite recent advancements in temporal reasoning, several technical gaps remain: simple time unit and duration unit normalization methods fail to efficiently leverage pre-trained models, adversarial techniques, while boosting performance, are computationally inefficient, vanilla attention mechanisms struggle to effectively identify temporal cues in textual data, and the issue of data scarcity persists, although it could be alleviated through data augmentation approaches.

3. METHODS

3.1. Proposed framework overview

Our proposed methodology framework consists of three main components: (a) a multi-data hybrid curation framework to prepare several sets of datasets for temporal commonsense reasoning; (b) a process to generate a synthetic dataset with a hybrid collaboration of chat agents and human experts; (c) fine-tuning pre-trained models using curated datasets, our target data and multi-stage fine-tuning using intermediate datasets with our target dataset. Figure 2 illustrates the overall framework for our proposed approach, and the following subsections provide additional details about these components.

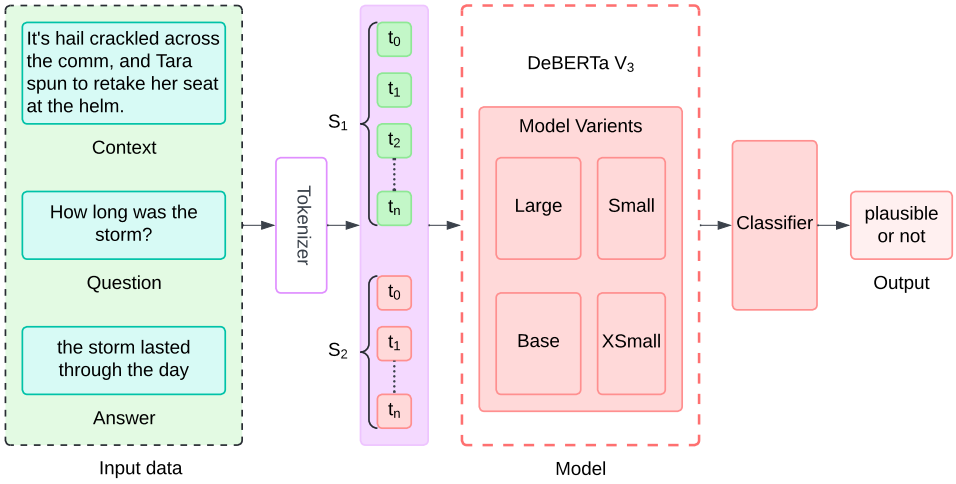

Figure 3. Overall process to fine-tune models using our proposed method for temporal commonsense reasoning.

3.2. Problem definition

Following previous work, we also utilize pre-trained language models to fine-tune our approach to answering questions that need an understanding of commonsense knowledge about time. We choose the DeBERTa as a pre-trained model. We formulate our temporal commonsense reasoning task using sequence-pair classification as a binary classification method. For sequence classification, the model is provided two sequences as input; for sequence one, the context and question of our target dataset are concatenated, and the candidate answer is sequence two. Special tokens of the model separate these sequences. The model's dense output layer utilizes the final hidden state of the classification special token in the sequence to make predictions about the plausibility of the answer of each instance in our dataset.

3.3. Pre-trained models

DeBERTa [52] is an advanced version of encoder-based models similar to the BERT [18] and RoBERTa [19] models, incorporating two innovative techniques. One of these techniques involves disentangled attention instead of a vanilla self-attention mechanism that uses content embedding as two vectors where a content vector and a position vector represent every word in sequence. This mechanism calculates attention weights for words with the relative position of other words using disentangled matrices based on these vectors, resulting in better contextual understanding. The second technique involves an improved mask decoder to predict masked tokens during pre-training instead of the output softmax layer. Together, all these design strategies contribute to significant performance enhancement of DeBERTa compared to its predecessors, BERT and RoBERTa.

We use DeBERTaV3 [26], particularly for the pre-trained language model, an upgraded iteration of the original DeBERTa model, implementing several advancements to enhance its performance. One notable improvement is changing the pre-training objective from traditional BERT-based masked language modeling (MLM) [18] to the Electra style replaced token detection (RTD) [53] technique, which is superior to MLM and a more sample-efficient technique. Additionally, DeBERTaV3 introduces a novel gradient-disentangled embedding sharing method. This innovative approach effectively addresses the tug-of-war dynamics commonly encountered during training [54], leading to enhanced training efficiency and superior quality in the pre-trained model. Furthermore, DeBERTaV3 offers a range of model sizes, including xsmall, small, base, and large. Figure 3 illustrates the overall process to fine-tune pre-trained models using our proposed approach.

3.4. Data augmentation techniques

We implemented two data augmentation techniques to expand our dataset and compare the performance of multiple fine-tuned models with different perspectives. We only applied data augmentation to the context and question sentences of our dataset, keeping the original sentences for the answers in both techniques.

3.4.1 SR

SR technique is a simple data augmentation technique where several words from a sentence are chosen and replaced with synonyms; it improved the performance of models for text classification where label data is not available in huge quantities[45]. For SR, we use the Easy Data Augmentation library (https://github.com/jasonwei20/eda_nlp). The noise absorption ability of sentences for data augmentation depends on the length of the sentence, so a long sentence can absorb more words without changing the original context and does not affect ground truth class labels. In order to balance this trade-off for short and long sentences, we use $n = \alpha l$ to select varying words to be changed[55]. Here, l represents the sentence length, and α is a parameter for the percentage of words to change in a sentence. This formula ensures that the number of words changed is proportional to the sentence length, maintaining a consistent level of augmentation across different sentence lengths. Following this setting, we generate four more augmented sentences for each original sample in our training data split. So, in this way, our original dataset expanded to four times.

3.4.2 BT

Given the limited size of our training dataset, we use the BT technique[48] to expand it[50]. We selected four languages that are closely related to our source language (English), and the available open-source machine translation models for these languages have demonstrated exceptional performance. We utilize the EasyNMT (https://github.com/UKPLab/EasyNMT) translation service on our server, a leading machine translation open software library offering various models for nearly 100 languages and different translation directions. Specifically, we utilize the Opus-MT model[56], a neural machine translation model based on encoder-decoder architecture, to translate the original English text to the selected languages. We use four languages - German (de), French (fr), Spanish (es), and Dutch (nl) - for forward translation. The translated text is then back-translated into English. Our back-translated dataset is also four times the size of the original dataset, ensuring consistency for comparison purposes.

3.5. Synthetic dataset creation process

We have utilized AI chat agents to create a synthetic dataset focused on the characteristics of our target temporal commonsense reasoning dataset, which we named generative temporal commonsense reasoning (GenTCSR). The process for creating this dataset, as illustrated in Figure 2B, involved a careful selection of 1,000 seed contexts to provide to the chat agent. From these contexts, we generated questions and answers for temporal commonsense reasoning using a customized prompt tailored to the specific characteristics of the data. Our proposed process ensures the reliability of the GenTCSR dataset. These seed contexts were sourced from the Soda dataset, which includes seed narrative sentences derived from temporal knowledge graphs. The seed sentences were then inputted into GPT-3 to generate complex contextual stories. After obtaining samples of questions and answers of these selected contexts using an LLaMA-based chatbot in our process, a human expert evaluated their quality and decided which ones should be included in the dataset. Ultimately, GenTCSR serves as one of the intermediate datasets for our proposed methodology.

3.6. Fine-tuned models

We have fine-tuned DeBERTa models on the MC-TACO dataset using the standard fine-tuning process, which we refer to as the standard fine-tuned model. We have also adapted two other fine-tuning models to train on our augmented data to compare the model performance for standard fine-tuning processes and fine-tuning after data augmentation. Figure 4 illustrates the overall working of our fine-tuned models. There are two important concepts in our DeBERTa.

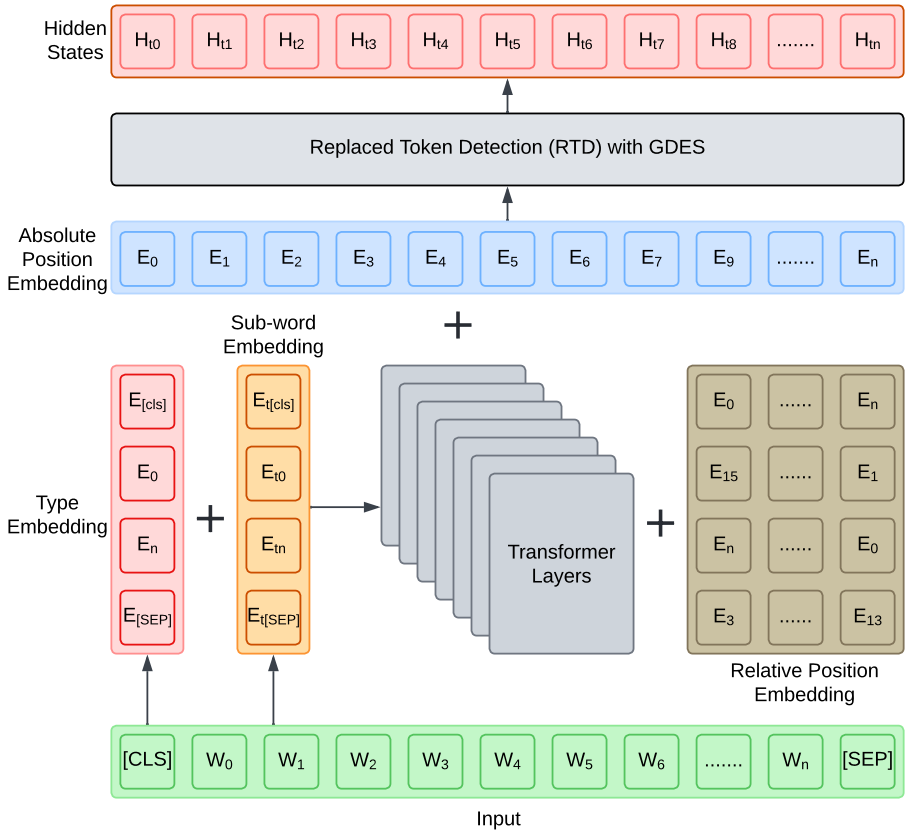

Figure 4. Overview of our model's processing pipeline for temporal commonsense reasoning. The input is first converted into embeddings and positional embeddings, which are then processed through the transformer layer. The output is passed to the RTD module with GDES, and the final hidden states are computed for reasoning tasks. RTD: Replaced token detection; GDES: gradient disentangled embedding sharing.

3.6.1 Disentangled attention

The traditional attention mechanism[57] used in BERT uses a single vector consisting of word embedding and position embedding for input layer representation; in contrast, disentangled attention[52] uses two different vectors to represent each embedding. Attention weight scores for each word are calculated using content vectors and position vectors for disentangled matrices. Two tokens, i and j, can be represented as $\{\mathbf{H_i}\}$, $\left\{\mathbf{P_{i|j}}\right\}$ and $\left\{\mathbf{H_j}\right\}$, $\left\{\mathbf{P_{j|i}}\right\}$ for the two vector approach of disentangled attention mechanism. To calculate attention score $\mathbf{A_{i,j}}$,

$$\begin{aligned}\mathbf{A_{i,j}} &= \left\{\mathbf{H_i}, \mathbf{P_{i|j}}\right\} \times \left\{\mathbf{H_j}, \mathbf{P_{j|i}}\right\}^{\mathsf{T}} \\ &= \mathbf{H_i}\mathbf{H_j}^{\mathsf{T}} + \mathbf{H_i}\mathbf{P_{j|i}}^{\mathsf{T}} + \mathbf{P_{i|j}}\mathbf{H_j}^{\mathsf{T}} + \mathbf{P_{i|j}}\mathbf{P_{j|i}}^{\mathsf{T}}\end{aligned} \tag{1}$$

where $\{\mathbf{H_i}\}$, $\left\{\mathbf{P_{i|j}}\right\}$, $\left\{\mathbf{H_j}\right\}$, and $\left\{\mathbf{P_{j|i}}\right\}$ represent the content of i, the position of i relative to j, the content of j, and the position of j relative to i, respectively. Equation (1) represents the disentangled matrices of content-to-content, content-to-position, position-to-content, and position-to-position, respectively. The position-to-position matrix has been removed from the equation since it does not add additional information.

The relative distance from token i to token j represented as $\delta(i, j) \in [0, 2k)$ can be defined as:

$$\delta(i,j) = \begin{cases} 0 & \text{for} \quad i-j \le -k \\ 2k-1 & \text{for} \quad i-j \ge k \\ i-j+k & \text{others.} \end{cases} \tag{2}$$

Here, k represents the maximum value for relative distance.

The standard self-attention[57] for single-head can be reformulated for disentangled attention using similar terms in Equation (3). The query ($\mathbf{Q_c}$), key ($\mathbf{K_c}$), and value ($\mathbf{V_c}$) vectors represent projected content derived from projection matrices $\mathbf{W_{q,c}}$, $\mathbf{W_{k,c}}$, $\mathbf{W_{v,c}} \in \mathbf{R^{d\times d}}$, respectively. A fixed vector for relative position embedding is represented as $\mathbf{P} \in \mathbf{R^{2k\times d}}$. Two other vectors ($\mathbf{Q_r}, \mathbf{K_r}$) are generated from projection matrices ($\mathbf{W_{q,r}}, \mathbf{W_{k,r}} \in \mathbf{R^{d\times d}}$) to represent projected relative position vectors.

$$\begin{gathered}\mathbf{Q_c} = \mathbf{HW_{q,c}}, \mathbf{K_c} = \mathbf{HW_{k,c}}, \mathbf{V_c} = \mathbf{HW_{v,c}}, \mathbf{Q_r} = \mathbf{PW_{q,r}}, \mathbf{K_r} = \mathbf{PW_{k,r}} \\ \tilde{\mathbf{A}}_{\mathbf{i,j}} = \mathbf{Q_i^c K_j^{c\intercal}} + \mathbf{Q_i^c K^r_{\delta(i,j)}}^{\intercal} + \mathbf{K_j^c Q^r_{\delta(j,i)}}^{\intercal} \\ \mathbf{H}_\circ = \text{softmax}\left(\frac{\tilde{\mathbf{A}}}{\sqrt{3d}}\right)\mathbf{V_c}\end{gathered} \tag{3}$$

Here, $\tilde{\mathbf{A}}_{\mathbf{i,j}}$ is the attention score for a token i relative to token j, $\delta(j,i)$ and $\delta(i,j)$ is the relative position matrix respective to token positions. The scaling factor applied here is $\frac{1}{\sqrt{3d}}$ instead of $\frac{1}{\sqrt{d}}$ in[57] for training stability. The attention score for each token in sequence can be computed using the algorithm[52].

3.6.2 Gradient disentangled embedding sharing

The DeBERTaV3 model training objective is inspired by the ELECTRA model, where the traditional MLM is replaced by a generative adversarial network (GAN)-based generator and discriminator approach. The two separate networks for the generator and discriminator share an embedding layer to learn from each other; this causes a drawback of tug-of-war dynamics. To overcome this problem, gradient disentangled embedding sharing (GDES)[26] is a novel approach that offers superior embedding sharing for each generator and discriminator. In GDES, token embeddings are shared by both the generator and discriminator. However, the gradients of the generator are not affected by RTD loss, which helps to avoid deficiencies due to conflicting objectives. GDES allows only the change of embeddings of the generator with the MLM loss, which helps to achieve the same converging speed of traditional embedding sharing with better results of generator output.

GDES implementation is presented as in Figure 5; the discriminator embedding is represented as $\mathbf{E}_D = sg\left(\mathbf{E}_G\right) + \mathbf{E}_\triangle$, where sg is an operator to stop gradients being updated for generator embedding $\mathbf{E}_G$ and only allow residual embedding $\mathbf{E}_\triangle$ to be updated. $\mathbf{E}_\triangle$ is initialed as a zero matrix and updated during training using no embedding sharing method. During training, inputs are generated through a generator for the discriminator for each iteration; MLM loss updates both $\mathbf{E}_G$ and $\mathbf{E}_\triangle$. Next $\mathbf{E}_\triangle$ updated through discriminator on inputs generated previously with RTD loss. Finally, $\mathbf{E}_G$ and $\mathbf{E}_\triangle$ are added to produce matrix for discriminator embedding $\mathbf{E}_D$.

4. EXPERIMENTS AND RESULTS

4.1. Datasets

We use four types of datasets to fine-tune our models with different experimental settings and evaluate our approach concerning temporal commonsense reasoning: a target dataset, an augmented dataset, and two intermediate datasets. We selected MC-TACO as our primary benchmark due to its comprehensive coverage of temporal commonsense categories (e.g., duration, event ordering) and its status as the standard for temporal commonsense reasoning evaluation benchmark to compare with prior work. MultiRC is included to overcome the data scarcity problem for temporal commonsense reasoning and is used as an intermediate dataset for a multi-stage fine-tuning approach. GenTCSR is used to increase the diversity of the dataset domain for temporal commonsense reasoning and also acts as a comparing approach to a data augmentation approach.

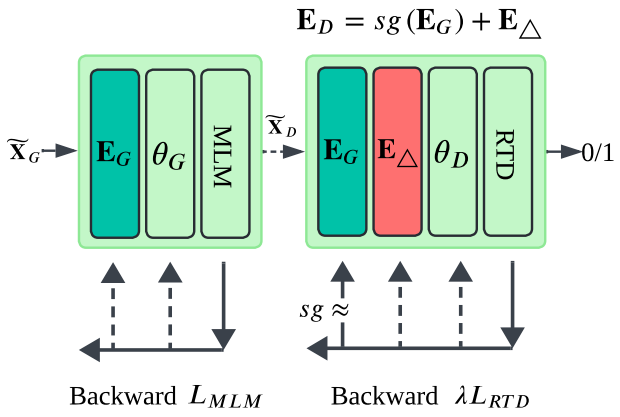

Figure 5. Overview of the GDES method. The discriminator embedding ($\mathbf{E}_D$) is formed by stopping gradients for the generator embedding ($\mathbf{E}_G$) while allowing updates to the residual embedding ($\mathbf{E}_\triangle$). GDES: Gradient-disentangled embedding sharing.

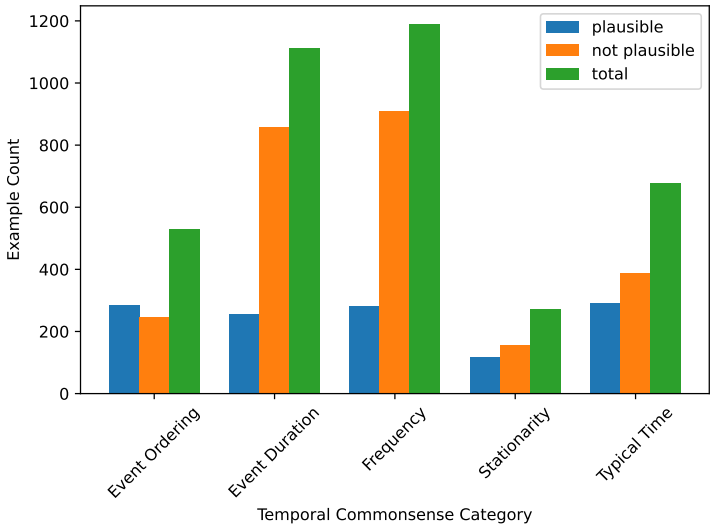

Figure 6. Statistical overview of dataset.

Below, we briefly describe these datasets and present statistics in the tables.

4.1.1 MC-TACO

MC-TACO [11] dataset introduces the temporal commonsense reasoning task to evaluate a pre-trained model's knowledge to understand temporal associations between events and entities in natural language text. MC-TACO is our target dataset; the development set is used for fine-tuning and dataset augmentation, and the test set is used to evaluate the performance of models trained with different experiments. The dataset comprises 13,000 samples, each presenting a distinct temporal reasoning challenge [11]. MC-TACO dataset categorizes temporal commonsense into five classes; no previous work has studied all these classes, so it proved a challenging benchmark for models to test temporal commonsense. The temporal categories are duration, temporal ordering, typical time, frequency, and stationarity. A tuple in each sample consists of a sentence for context, a question with some temporal cue, a candidate answer, and a temporal category, which we are not using as input for features for our models. The statistical distribution for the dataset is presented in Figure 6.

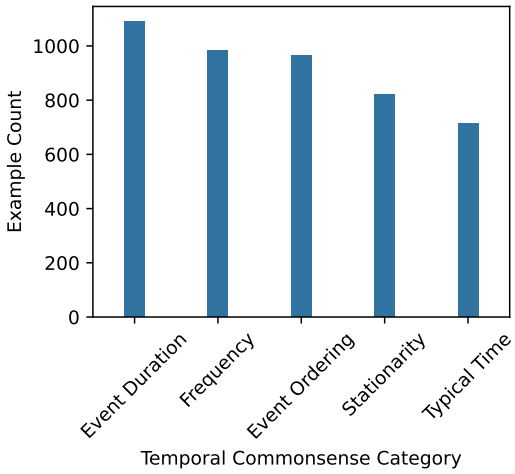

Figure 7. Statistical overview of GenTCSR dataset. GenTCSR: Generative temporal commonsense reasoning.

Table 1. Parameters for our baseline models

Model name	Vocabulary (K)	Backbone parameters (M)	Hidden size	Layers
$DeBERTaV3_{xsmall}$	128	22	384	12
$DeBERTaV3_{small}$	128	44	768	6
$DeBERTaV3_{base}$	128	86	768	12
$DeBERTaV3_{large}$	128	304	1024	24

4.1.2 MultiRC

MultiRC is a reading comprehension dataset with short paragraphs and multi-sentence questions from diverse domains (e.g., science books from elementary school, travel guides, *etc.*). It challenges the model to independently identify the correct answers from multiple-choice questions without being a span in the text from a given passage. It has around 10 K questions, and 60% of the data is released as training and development data. For our purpose, we used only the train set of this dataset as an intermediate dataset. It is a general machine reading dataset and does not have explicit temporal commonsense aspects.

4.1.3 GenTCSR

We used large language models to generate a temporal commonsense reasoning dataset named GenTCSR. This synthetic data is also used as an intermediate dataset for our experiments; the creation of this dataset is presented in our methods section. This generated dataset has 13,221 samples. We adopted MC-TACO temporal commonsense categories for this dataset, and the statistics of these categories are shown in Figure 7.

4.2. Baseline models

The DeBERTa model, known for its disentangled attention mechanism, improved encoding capabilities, and enhanced training, has established itself as a reliable and effective model for baseline in various NLP tasks (i.e., text classification, named entity recognition, and question answering). In our study, we have meticulously chosen four baseline models of different sizes: $DeBERTaV3_{large}$, $DeBERTaV3_{base}$, $DeBERTaV3_{small}$, and $DeBERTaV3_{xsmall}$, all of which are pre-trained models. We are following the DeBERTa codebase setting for our study, ensuring a rigorous and thorough approach. Detailed parameter information about these baselines is presented in Table 1.

4.3. Experiment settings

We utilized the Transformers library with Pytorch for our experiments to fine-tune all models. We used a learning rate ranging from 9×10^6 to 1×10^5 for the ADAM optimizer and training batch size ranging from

Table 2. Experiment settings for fine-tuning different DeBERTaV3 models

Hyperparameter	XSmall	Small	Base	Large
Dropout of task layer	0.1	0.1	0.1	0.1
Warmup steps	0.01	0.01	0.01	0.01
Learning rates	5e-6	5e-6	5e-6	5e-6
Batch size	64	64	64	64
Weight decay	0.01	0.01	0.01	0.01
Max training epochs	5,10	5,10	5,10	5,10
Learning rate decay	Linear	Linear	Linear	Linear
Adam ϵ	1e-6	1e-6	1e-6	1e-6
Adam β_1	0.9	0.9	0.9	0.9
Adam β_2	0.999	0.999	0.999	0.999
Gradient clipping	1.0	1.0	1.0	1.0

Table 3. Results for DeBERTaV3 baseline models

Model name	Exact match (%)	F1 (%)
XSmall	10.66	17.42
Small	17.42	17.42
Base	17.42	17.42
Large	12.16	49.84

16 to 128. Depending on the dataset type, we used five or ten training epochs. Gradient clipping of 1 is used to evade gradient explosion. We used WordPiece [58] to tokenize all texts and did not chop text for maximum sequence length. The more detailed hyperparameters setting is presented in Table 2 for all models following the default settings of DeBERTaV3.

4.4. Evaluation metrics

Following other models, we also used two question-level metrics [Exact match (EM) and F1 score] for evaluation. EM is a binary metric that checks if the prediction exactly matches the original answer. If there is even a single character difference, it is considered wrong. If the predicted string matches the reference string exactly, its EM score is 1; otherwise, it is 0. The EM score for a dataset is the sum of all individual EM scores divided by the total number of predictions. F1 score is another popular metric for classification tasks. It compares words in the predicted answer to those in the ground truth answer.

4.5. Results

This section summarizes our experiments' results on an MC-TACO test set-based evaluator and presents results for each fine-tuned model in detail. The key findings of our experiments include the effectiveness of augmenting text data with different methods, such as SR, BT, and multi-stage fine-tuning, in enhancing the understanding of temporal commonsense tasks.

4.5.1 Results for baseline and standard fine-tuning

Table 3 summarizes the results of baseline models. The baseline models were evaluated without fine-tuning on the target dataset, serving as a reference point for comparison. Table 4 presents the results of the standard fine-tuning method using our target dataset without any data augmentation. To compare the results of the baseline model and our standard fine-tuned model, we can see an apparent performance increase; fine-tuning significantly improves both EM and F1 scores across all model sizes. Notably, the DeBERTaV3-Large model achieves the highest performance, with an EM score of 63.06% and an F1 score of 82.17%, demonstrating the effectiveness of fine-tuning on our target dataset. The Base model improves from 17.42% EM and 17.42% F1 in the baseline setting to 48.72% EM and 75.87% F1 after fine-tuning. Similarly, the Small model exhibits a substantial boost, increasing its EM from 17.42% to 32.36% and F1 from 17.42% to 65.36%. These improvements highlight the effectiveness of task-specific fine-tuning, with larger models benefiting more significantly.

Table 4. Results for standard fine-tuned models

Model name	Exact match (%)	F1 (%)
XSmall	17.42	25.81
Small	32.36	65.36
Base	48.72	75.87
Large	63.06	82.17

Table 5. Results for fine-tuned models on SR augmented data

Model name	Exact match (%)	F1 (%)
XSmall	35.06	66.01
Small	46.10	73.17
Base	55.26	78.89
Large	62.76	83.11

SR: Synonym replacement.

4.5.2 Results for augmented data fine-tuning

Table 5 shows the results of models fine-tuned on SR augmented data. For comparison purposes, it shows a visible gain in the performance results of all model variants over standard fine-tuned models. Notably, even the XSmall model exhibits significant improvements compared to standard fine-tuning, with EM increasing from 17.42% to 35.06% and F1 from 25.81% to 66.01%. This suggests that SR augmentation enhances model robustness, particularly for smaller architectures. Table 6 shows a complete detail of the results accomplished through fine-tuning models on BT-based augmented data, covering all languages and respective languages. The thoroughness of our technique is evident in the detailed results presented, which cover four languages and four models. The best results and models are in bold to emphasize the best-performing augmented data for each language. Our results show that French is the best-performing language, giving two better models than other languages: XSmall and Large. The second best-performing language is Dutch, which gives us the best results for the Small model. Using all language data, we also achieved the best results for the Base model. This table proves our hypothesis that augmenting text data with different techniques helps enhance the understanding of temporal commonsense tasks. Overall, these results suggest that data augmentation, particularly BT, substantially improves model generalization, with the best performance observed for French and Spanish augmentations on larger models. Additionally, SR offers a lightweight yet effective enhancement, particularly for smaller models.

4.5.3 Results for multi-stage fine-tuning

We also evaluated the performance of multi-stage fine-tuning models trained on two intermediate datasets, MultiRC and GenTCSR. The results demonstrate notable differences in the effectiveness of the intermediate datasets and reveal trends in model scalability. Table 7 summarizes the results of multi-stage fine-tuned models across the two intermediate datasets. Fine-tuning on MultiRC before training on our target dataset leads to consistent performance gains across all model sizes. Notably, the Large model achieves 64.11% EM and 84.21% F1, outperforming both standard fine-tuning and SR augmentation. Similarly, the Base model benefits significantly, reaching 53.68% EM and 77.53% F1, marking a strong improvement over its baseline and standard fine-tuned counterparts. Smaller models (XSmall and Small) also exhibit notable performance increases, reinforcing the effectiveness of knowledge transfer from MultiRC. In contrast, using GenTCSR as an intermediate dataset yields mixed results. While the Large model still achieves strong performance (59.01% EM, 80.89% F1), the Base model's performance drops (34.76% EM, 62.06% F1) compared to MultiRC-based fine-tuning. Interestingly, the Small model performs better than expected (38.06% EM, 66.66% F1), suggesting that GenTCSR may provide more benefits to smaller architectures. However, the XSmall model sees a significant drop in performance compared to MultiRC, indicating that its domain or task formulation may not transfer as effectively. Another reason for the performance drop is that MultiRC has more examples than GenTCSR. To

Table 6. Results for fine-tuned models on back translation augmented data

Language	Model name	Exact match (%)	F1 (%)
German (de)	XSmall	27.55	47.84
	Small	33.33	66.47
	Base	50.10	77.00
	Large	62.99	83.11
Spanish (es)	XSmall	19.44	22.61
	Small	34.16	67.12
	Base	50.23	76.98
	Large	64.49	83.29
French (fr)	**XSmall**	**30.78**	**62.72**
	Small	38.44	68.87
	Base	50.90	77.49
	Large	**64.19**	**84.35**
Dutch (nl)	XSmall	24.62	36.15
	Small	**42.34**	**71.07**
	Base	48.05	75.22
	Large	61.56	82.50
all	XSmall	31.46	65.64
	Small	39.79	69.97
	Base	**56.46**	**79.44**
	Large	63.96	83.62

Bold denotes the best performance.

Table 7. Results for multi-stage fine-tuned models on two intermediate datasets

Intermediate dataset	Model name	Exact match (%)	F1 (%)
MultiRC	**XSmall**	35.36	68.02
	Small	37.31	68.72
	Base	53.68	77.53
	Large	64.11	84.21
GenTCSR	XSmall	27.03	55.30
	Small	38.06	66.66
	Base	34.76	62.06
	Large	59.01	80.89

Bold denotes the best performance. GenTCSR: Generative temporal commonsense reasoning.

conclude, models fine-tuned on MultiRC consistently achieved higher performance across all sizes compared to those fine-tuned on GenTCSR, underscoring the potential benefits of human-created datasets over synthetic datasets generated by large language models for temporal commonsense reasoning.

4.5.4 Results comparison with prior work

We chose the previous SOTA studies to compare our results with other works and reported their best results in Table 8. The upper part of the table presents results for other transformer-based models; these models are encoder-based architecture except the T5, which is encoder-decoder architecture. The results for TacoLM are reported for individual categories of temporal commonsense, so we calculated the average of these results and presented them here to compare overall results with fairness. The middle part of the table presents results for adversarial-based methods. These two models are current SOTA results for temporal commonsense, and as for comparison, we can see these models are better performing than the other transformer-based methods. The bottom part of the table presents our results; as we can see, our large model surpasses the current SOTA ALICE++ on both EM and F1, scoring approximately 4% more. To compare our result with another transformer-based method, our large model also outperforms the T5 model by approximately 5% on both metrics and also, our model has fewer backbone parameters than the T5 model. Our Base model also achieves better results than the RoBERTa large and BERT large models, and it has substantially fewer parameters than the RoBERTa and BERT models.

Table 8. Results comparison of our best-fine-tuned models with other relevant studies

Methods	Model name	Exact match (%)	F1 (%)
Other transformer-based	TacoLM [14]	42.72	-
	BERT [59]	45.2	72.5
	RoBERTa [59]	51.2	76.2
	ALBERT [59]	59.2	79.7
	T5-3B [21]	59.08	79.46
Adversarial based	ALICE [23]	56.45	79.50
	ALICE++ [24]	59.90	80.88
Ours (best results)	**XSmall**	**35.36**	**68.02**
	Small	**37.31**	**68.72**
	Base	**55.41**	**78.99**
	Large	**64.19**	**84.35**

Bold represents our best results.

5. ANALYSIS AND DISCUSSION

5.1. Quality and biases in GenTCSR

Our statistical analysis indicates that the quality of GenTCSR is comparable to that of MC-TACO across multiple dimensions. The average narrative length in GenTCSR is 183.14 words, and the average question length is 50.95 words, closely aligning with MC-TACO's distribution. Moreover, GenTCSR maintains a similar coverage of temporal commonsense categories, ensuring a balanced representation of event ordering, duration, frequency, stationarity, and typical time. Additionally, lexical analysis of the most frequent words in GenTCSR questions reveals strong alignment with MC-TACO, with key temporal indicators such as often, time, when, and before appearing at comparable frequencies. These similarities suggest that GenTCSR effectively captures the structural and semantic characteristics of real-world temporal reasoning datasets, reinforcing its utility for benchmarking temporal commonsense understanding. While GenTCSR enhances temporal reasoning, its synthetic nature limits exposure to real-world ambiguity. It may also introduce common biases of pre-trained models of AI chatbots used for data creation. Future work will integrate crowdsourced ambiguous examples and deep analysis of biases introduced.

5.2. Performance analysis on temporal categories

Our approach significantly improved all our models' performance across all evaluation metrics in the MC-TACO dataset by integrating SR and BT augmentations during training, as demonstrated in Tables 5 and 6. Our results evidently depict the significant impact of our approach with and without data augmentation compared to the baseline results for all model variants.

In Table 9, we have presented the results for each individual category of Temporal Commonsense to show detailed findings. The table implies the percentage of correct answers for each category. These results are grouped according to model size and represent a structured result for each dataset. The XSmall model fine-tuned on an augmented dataset of German and French back-translated data achieved the best score for event ordering, event duration, frequency, and typical time categories. The model trained on BT(de) performed best for the stationarity category. The Small model fine-tuned on the BT (NL) dataset gained the best scores for event ordering and event duration; also, trained on the DA(sr) dataset, it performed best for frequency, stationarity, and typical time categories. For the Base model, the best scores for three categories, event ordering, event duration, and frequency, were achieved by the BT(all) trained model, and for two categories, stationarity and typical time were achieved for the DA(sr) dataset. The Large model displayed diverse results for each category. The BT(fr) trained model achieved the best scores for event ordering, frequency, and typical time categories. The model trained on BT(de) achieved the best score for event duration, while the model trained on DA(sr) achieved the best score for the stationarity category.

Our main results and secondary results for each temporal commonsense category show the impact of data aug-

Table 9. Results comparison according to temporal commonsense categories

Configration		Temporal CSR				
Model	**Setting**	**Event ordering**	**Event duration**	**Frequency**	**Stationarity**	**Typical time**
XSmall	Baseline	49.46	54.78	49.16	42.38	48.83
	STD	52.45	74.04	73.65	56.28	56.90
	DA(sr)	51.43	71.77	72.81	66.83	66.39
	BT(de)	53.20	76.65	75.28	**72.36**	68.96
	BT(es)	52.45	74.04	73.65	56.28	56.90
	BT(fr)	**59.13**	**82.42**	**81.33**	56.66	**73.00**
	BT(nl)	55.18	74.01	73.01	66.16	65.03
	BT(all)	53.75	75.76	77.55	67.50	67.76
Small	Baseline	52.45	74.04	73.65	56.28	56.90
	STD	62.33	83.38	81.45	66.16	76.87
	DA(sr)	70.78	86.74	**86.03**	**75.71**	**85.38**
	BT(de)	63.69	82.88	81.73	67.34	75.72
	BT(es)	63.90	82.52	81.33	70.18	76.70
	BT(fr)	63.28	85.55	83.88	71.36	80.96
	BT(nl)	**71.32**	**86.77**	85.15	74.87	82.87
	BT(all)	68.53	83.87	83.64	73.03	80.74
Base	Baseline	52.45	74.04	73.65	56.28	56.90
	STD	74.32	88.92	87.38	78.39	87.78
	DA(sr)	81.06	90.20	88.30	**84.42**	**89.80**
	BT(de)	75.95	89.78	87.22	77.72	87.40
	BT(es)	76.57	89.35	87.14	79.90	88.82
	BT(fr)	75.75	88.56	87.66	77.22	89.58
	BT(nl)	71.87	87.86	85.35	75.21	88.11
	BT(all)	**81.13**	**90.11**	**89.17**	82.58	89.25
Large	Baseline	47.55	25.96	26.35	43.72	43.10
	STD	86.92	91.75	91.12	84.76	92.69
	DA(sr)	85.22	91.66	91.84	**85.93**	91.82
	BT(de)	85.97	**92.71**	92.24	83.92	92.03
	BT(es)	87.06	91.26	91.72	84.59	92.74
	BT(fr)	**88.49**	92.25	**92.36**	84.09	**92.85**
	BT(nl)	85.35	91.79	91.28	83.25	91.76
	BT(all)	87.74	92.05	91.68	84.59	91.98

Bold denotes the best performance. CSR: Commonsense reasoning; STD: standard fine-tuned.

mentation on our models' performance. Both data augmentation SR and BT are very effective in improving the performance of our fine-tuned models compared to baseline models. Data augmentation using BT consistently enhances performance results for all model sizes on the temporal commonsense reasoning dataset. These enhancements indicate that BT augmentation facilitates a more thorough comprehension of the text, allowing the model to recognize various categories and contexts of temporal commonsense better.

6. CONCLUSIONS

This paper has focused on understanding temporal commonsense reasoning using DeBERTa models. Temporal commonsense is significant for commonsense reasoning in the NLP field. Understanding and reasoning about time pose challenges for computers, making it an important area for further study and development in NLP. It is evident that NLP research has evolved significantly in understanding temporal aspects of text data. The shift towards temporal commonsense reasoning reflects the dynamic nature of this field. Previous research only focuses on one aspect of temporal commonsense, such as duration, ordering, or infilling, but our paper presents a unique contribution by introducing five categories of temporal commonsense. Our research introduces a novel framework that unifies multiple datasets and generates a synthetic dataset for these five distinct categories of temporal commonsense, enabling a more holistic approach to modeling temporal phenomena in text data. We utilized various training and fine-tuning methods, including multi-stage fine-tuning, and augmented our data using different methods. After conducting extensive experiments and evaluating our technique against standard question-answering metrics, our approach has demonstrated SOTA results, surpassing all previous studies on temporal commonsense within the MC-TACO dataset. Overall, our focus has

been on developing computationally efficient methods and presenting the latest findings in the domain of temporal commonsense reasoning. We aimed to explore and propose novel approaches that effectively generalize the temporal commonsense tasks without requiring extensive human labeling or computational resources and advance this crucial area of research, which has significant implications for various fields. While our work has limitations, such as the need for human supervision in synthetic dataset validation, it is still a faster alternative to manual human labeling of datasets. Future directions include extending this framework to multilingual contexts, integrating external knowledge sources for richer temporal grounding, and developing fully automated data generation pipelines to eliminate dependency on human validation.

DECLARATIONS

Authors' contributions

Made substantial contributions to the conception and design of the study; performed data analysis and experimentation: Feroze, W.

Provided technical support and conducted reviews: Shahid, M.; Cheng, S.; Jimale, E. L.

Supervised the project; provided administrative, technical, and material support; performed proofreading: Qu, H.; Wang, Y.

Secured funding and conducted the final proofreading: Yang, Y.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Financial support and sponsorship

This work was supported by the National Defense Basic Scientific Research Program of China under Grant WDZC20235250407.

Conflicts of interest

Qu, H. is an Editorial Board Member of the journal *Intelligence & Robotics*. Qu, H. was not involved in any steps of editorial processing, notably including reviewer selection, manuscript handling, or decision-making. The other authors declare that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Davis, E.; Marcus, G. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM.* **2015**, *58*, 92–103. DOI
2. Zhang, S.; Rudinger, R.; Duh, K.; Van Durme, B. Ordinal common-sense Inference. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 379–95. DOI
3. Yang, P.; Liu, Z.; Li, B.; Zhang, P. Implicit relation inference with deep path extraction for commonsense question answering. *Neural. Process. Lett.* **2022**, *54*, 4751–68. DOI
4. Wang, C.; Liu, J.; Liu, J.; Wang, W. Inference of isA commonsense knowledge with lexical taxonomy. *Appl. Intell.* **2022**, *53*, 5290–303. DOI
5. Nguyen, T. P.; Razniewski, S.; Romero, J.; Weikum, G. Refined commonsense knowledge from large-scale web contents. *IEEE. Trans. Knowl. Data. Eng.* **2023**, *35*, 8431–47. DOI

6. Su, C.; Yu, G.; Wang, J.; Yan, Z.; Cui, L. A review of causality-based fairness machine learning. *Intell. Robot.* **2022**, *2*, 244-74. DOI
7. Zhong, X.; Cambria, E. Time expression recognition and normalization: a survey. *Artif. Intell. Rev.* **2023**, *56*, 9115–40. DOI
8. Ji, A.; Woo, W. L.; Wong, E. W. L.; Quek, Y. T. Rail track condition monitoring: a review on deep learning approaches. *Intell. Robot.* **2021**, *1*, 151-75. DOI
9. Lin, Y.; Xie, Z.; Chen, T.; Cheng, X.; Wen, H. Image privacy protection scheme based on high-quality reconstruction DCT compression and nonlinear dynamics. *Expert. Syst. Appl.* **2024**, *257*, 124891. DOI
10. Campos, R.; Dias, G.; Jorge, A. M.; Jatowt, A. Survey of temporal information retrieval and related applications. *ACM. Comput. Surv.* **2014**, *47*, 1-41. DOI
11. Zhou, B.; Khashabi, D.; Ning, Q.; Roth, D. "Going on a vacation" takes longer than "Going for a walk": a study of temporal commonsense understanding. 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)*, Hong Kong, China. Association for Computational Linguistics, 2019; pp. 3363–9. DOI
12. Huang, G.; Min, Z.; Ge, Q.; Yang, Z. Towards document-level event extraction via Binary Contrastive Generation. *Knowl. Based. Syst.* **2024**, *296*, 111896. DOI
13. Yang, Z.; Du, X.; Rush, A.; Cardie, C. Improving event duration prediction via time-aware pre-training. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020. pp. 3370–8. DOI
14. Zhou, B.; Ning, Q.; Khashabi, D.; Roth, D. Temporal common sense acquisition with minimal supervision. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020; pp. 7579–89. DOI
15. Ning, Q.; Wu, H.; Peng, H.; Roth, D. Improving temporal relation extraction with a globally acquired statistical resource. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana. Association for Computational Linguistics, 2018; pp. 841–51. DOI
16. Lin, S. T.; Chambers, N.; Durrett, G. Conditional generation of temporally-ordered event sequences. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021; pp. 7142–57. DOI
17. Cole, J. R.; Chaudhary, A.; Dhingra, B.; Talukdar, P. Salient span masking for temporal understanding. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia. Association for Computational Linguistics, 2023; pp. 3052–60. DOI
18. Devlin, J.; Chang, M. W.; Lee, K.; Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics, 2019; pp. 4171–86. DOI
19. Liu, Y.; Ott, M.; Goyal, N.; et al. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692. Available online: https://doi.org/10.48550/arXiv.1907.11692. (accessed 7 Mar 2025)
20. Ribeiro, M. T.; Wu, T.; Guestrin, C.; Singh, S. Beyond accuracy: behavioral testing of NLP models with CheckList. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020; pp. 4902–12. DOI
21. Kaddari, Z.; Mellah, Y.; Berrich, J.; Bouchentouf, T.; Belkasmi, M. G. Applying the T5 language model and duration units normalization to address temporal common sense understanding on the MCTACO dataset. In: *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, Fez, Morocco, 09-11 Jun, 2020. IEEE, 2020; pp. 1–4. DOI
22. Raffel, C.; Shazeer, N.; Roberts, A.; et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv* **2019**, arXiv:1910.10683. Available online: https://doi.org/10.48550/arXiv.1910.10683. (accessed 7 Mar 2025)
23. Pereira, L.; Liu, X.; Cheng, F.; Asahara, M.; Kobayashi, I. Adversarial training for commonsense inference. In: *Proceedings of the 5th Workshop on Representation Learning for NLP. Online: Association for Computational Linguistics*. Association for Computational Linguistics, 2020; pp. 55–60. DOI
24. Pereira, L.; Cheng, F.; Asahara, M.; Kobayashi, I. ALICE++: adversarial training for robust and effective temporal reasoning. In: *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China. Association for Computational Lingustics, 2021; pp. 373–82. https://aclanthology.org/2021.paclic-1.40/. (accessed 2025-03-07)
25. Kanashiro Pereira, L. Attention-focused adversarial training for robust temporal reasoning. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France*. European Language Resources Association, 2022; pp. 7352–59. https://aclanthology.org/2022.lrec-1.800/. (accessed 2025-03-07)
26. He, P.; Gao, J.; Chen, W. DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv* **2021**, arXiv:2111.09543. Available online: https://doi.org/10.48550/arXiv.2111.09543. (accessed 7 Mar 2025)
27. Forbes, M.; Choi, Y. Verb physics: relative physical knowledge of actions and objects. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics, 2017; pp. 266–76. DOI
28. Cocos, A.; Wharton, S.; Pavlick, E.; Apidianaki, M.; Callison-Burch, C. Learning scalar adjective intensity from paraphrases. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, 2018; pp. 1752–62. DOI
29. Rashkin, H.; Sap, M.; Allaway, E.; Smith, N. A.; Choi, Y. Event2Mind: commonsense inference on events, intents, and reactions. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia.

Association for Computational Linguistics, 2018; pp. 463–73. DOI

30. Zellers, R.; Bisk, Y.; Schwartz, R.; Choi, Y. SWAG: a large-scale adversarial dataset for grounded commonsense inference. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, 2018; pp. 93–104. DOI
31. Ning, Q.; Wu, H.; Roth, D. A multi-axis annotation scheme for event temporal relations. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics, 2018; pp. 1318–28. DOI
32. Vashishtha, S.; Van Durme, B.; White, A. S. Fine-grained temporal relation extraction. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, 2019; pp. 2906–19. DOI
33. Ning, Q.; Zhou, B.; Feng, Z.; Peng, H.; Roth, D. CogCompTime: a tool for understanding time in natural language. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium. Association for Computational Linguistics, 2018; pp. 72–7. DOI
34. Leeuwenberg, A.; Moens, M. F. Temporal information extraction by predicting relative time-lines. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, 2018; pp. 1237–46. DOI
35. Li, Z.; Ding, X.; Liu, T. Constructing narrative event evolutionary graph for script event prediction. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2018; pp. 4201–7. DOI
36. Williams, J. Extracting fine-grained durations for verbs from Twitter. In: *Proceedings of ACL 2012 Student Research Workshop*, Jeju Island, Korea. Association for Computational Linguistics, 2012; pp. 49–54. https://aclanthology.org/W12-3309/. (accessed 2025-03-07)
37. Vempala, A.; Blanco, E.; Palmer, A. Determining event durations: models and error analysis. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, 2018; pp. 164–8. DOI
38. Clark, P.; Cowhey, I.; Etzioni, O.; et al. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv* **2018**, arXiv:1803.05457. Available online: https://doi.org/10.48550/arXiv.1803.05457. (accessed 7 Mar 2025)
39. Ostermann, S.; Roth, M.; Modi, A.; Thater, S.; Pinkal, M. SemEval-2018 Task 11: machine comprehension using commonsense knowledge. In: *Proceedings of the 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana. Association for Computational Linguistics, 2018; pp. 747–57. DOI
40. Merkhofer, E.; Henderson, J.; Bloom, D.; Strickhart, L.; Zarrella, G. MITRE at SemEval-2018 Task 11: commonsense reasoning without commonsense knowledge. In: *Proceedings of the 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana. Association for Computational Linguistics, 2018; pp. 1078–82. DOI
41. Mostafazadeh, N.; Chambers, N.; He, X.; et al. A corpus and cloze evaluation for deeper understanding of commonsense stories. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California. Association for Computational Linguistics, 2016; pp. 839–49. DOI
42. Ning, Q.; Wu, H.; Han, R.; Peng, N.; Gardner, M.; Roth, D. TORQUE: a reading comprehension dataset of temporal ordering questions. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020; pp. 1158–72. DOI
43. Qin, L.; Gupta, A.; Upadhyay, S.; He, L.; Choi, Y.; Faruqui, M. TIMEDIAL: temporal commonsense reasoning in dialog. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021; pp. 7066–76. DOI
44. Yin, S.; Xiang, Z. A hyper-heuristic algorithm via proximal policy optimization for multi-objective truss problems. *Expert. Syst. Appl.* **2024**, *256*, 124929. DOI
45. Kobayashi, S. Contextual augmentation: data augmentation by words with paradigmatic relations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana. Association for Computational Linguistics, 2018; pp. 452–7. DOI
46. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. *arXiv* **2015**, arXiv:1509.01626. Available online: https://doi.org/10.48550/arXiv.1509.01626. (accessed 7 Mar 2025)
47. Wang, W. Y.; Yang, D. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal. Association for Computational Linguistics, 2015; pp. 2557–63. DOI
48. Sennrich, R.; Haddow, B.; Birch, A. Improving neural machine translation models with monolingual data. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. Association for Computational Linguistics, 2016; pp. 86–96. DOI
49. Yu, A. W.; Dohan, D.; Luong, M.; et al. QANet: combining local convolution with global self-attention for reading comprehension. *arXiv* **2018**, arXiv:1804.09541. Available online: https://doi.org/10.48550/arXiv.1804.09541. (accessed 7 Mar 2025)
50. Fadaee, M.; Monz, C. Back-translation sampling by targeting difficult words in neural machine translation. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, 2018; pp. 436–46. DOI
51. Sugiyama, A.; Yoshinaga, N. Data augmentation using back-translation for context-aware neural machine translation. In: *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, Hong Kong, China. Association for Computational Linguis-

tics, 2019; pp. 35–44. DOI

52. He, P.; Liu, X.; Gao, J.; Chen, W. DeBERTa: decoding-enhanced BERT with disentangled attention. *arXiv* **2020**, arXiv:2006.03654. Available online: https://doi.org/10.48550/arXiv.2006.03654. (accessed 7 Mar 2025)
53. Clark, K.; Luong, M. T.; Le, Q. V.; Manning, C. D. ELECTRA: pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555. Available online: https://doi.org/10.48550/arXiv.2003.10555. (accessed 7 Mar 2025)
54. Hadsell, R.; Rao, D.; Rusu, A.; Pascanu, R. Embracing change: continual learning in deep neural networks. *Trends. Cogn. Sci.* **2020**, *24*, 1028–40. DOI
55. Wei, J.; Zou, K. EDA: easy data augmentation techniques for boosting performance on text classification tasks. 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)*, Hong Kong, China. Association for Computational Linguistics, 2019; pp. 6383–9. DOI
56. Tiedemann, J.; Thottingal, S. OPUS-MT - building open translation services for the world. In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal. European Association for Machine Translation, 2020; pp. 479–80. https://aclanthology.org/2020.eamt-1.61/. (accessed 2025-03-07)
57. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762. Available online: https://doi.org/10.48550/arXiv.1706.03762. (accessed 7 Mar 2025)
58. Song, X.; Salcianu, A.; Song, Y.; Dopson, D.; Zhou, D. Fast WordPiece tokenization. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021. pp. 2089–103. DOI
59. Kimura, M.; Kanashiro Pereira, L.; Kobayashi, I. Toward building a language model for understanding temporal commonsense. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*. Association for Computational Linguistics, 2022; pp. 17–24. DOI