Temporal reasoning for timeline summarisation in social media

Jiayu Song[1], **Mahmud Akhter**[1], **Dana Atzil-Slonim**[2], **Maria Liakata**[1,3]
[1]Queen Mary University of London, London, UK
[2]Bar-Ilan University, Israel
[3]The Alan Turing Institute, London, UK
{jiayu.song,m.akhter,m.liakata}@qmul.ac.uk
dana.slonim@gmail.com

Abstract

This paper explores whether enhancing temporal reasoning capabilities in Large Language Models (LLMs) can improve the quality of timeline summarization, the task of summarising long texts containing sequences of events, particularly social media threads . We introduce *NarrativeReason*, a novel dataset focused on temporal relationships among sequential events within narratives, distinguishing it from existing temporal reasoning datasets that primarily address pair-wise event relationships. Our approach then combines temporal reasoning with timeline summarization through a knowledge distillation framework, where we first fine-tune a teacher model on temporal reasoning tasks and then distill this knowledge into a student model while simultaneously training it for the task of timeline summarization. Experimental results demonstrate that our model achieves superior performance on mental health-related timeline summarization tasks, which involve long social media threads with repetitions of events and a mix of emotions, highlighting the importance of leveraging temporal reasoning to improve timeline summarisation.

1 Introduction

Timeline summarization organizes and presents a sequence of events in a coherent and concise manner (Steen and Markert, 2019; Li et al., 2021; Hu et al., 2024). It involves extracting event-related timelines and then summarising them (Hu et al., 2024; Rajaby Faghihi et al., 2022). Researchers generally create event graphs (Li et al., 2021) or cluster event related timelines (Hu et al., 2024) to identify relevant events. Recent work (Song et al., 2024) has introduced the challenging task of social media timeline summarisation, especially in the context of capturing fluctuations in individuals' state of mind as reflected in posts shared online over time. In these posts, numerous events may occur without explicit timestamps, requiring contextual inference to determine their chronological sequence. Moreover, mental health-related events are not easy to identify: they can be connected to an individual's emotions, interpersonal interactions, and the entire timeline is necessary to provide enough context(Song et al., 2024). It is particularly challenging to identify events pertaining to psychological states and to extract these from posts. When generating mental health related summaries from longitudinal posts, models need to understand related events and maintain temporal consistency to make inferences. This raises the question of whether temporal reasoning can be leveraged to enhance the quality of complex timeline summaries.

Temporal reasoning involves understanding and processing temporal information in text to deduce time-based relationships between events (Wenzel and Jatowt, 2023a). (Zhou et al., 2019) categorises temporal commonsense reasoning with respect to five aspects (duration, temporal ordering, typical time, frequency and stationarity). Subsequently, (Tan et al., 2023a; Jain et al., 2023) explore the temporal reasoning capabilities of Large Language Models (LLMs) with respect to the temporal commonsense aspects. LLMs with a strong understanding of temporal context can perform better on downstream tasks, including storytelling, natural language inference, timeline comprehension and tracking user status (Jain et al., 2023). Thus temporal commonsense reasoning is beneficial for timeline summarisation, as it helps maintain temporal consistency and the correct event order (Wenzel and Jatowt, 2023b; Vashishtha et al., 2020). Recent work (Chan et al., 2024; Feng et al., 2023; Chen et al., 2024; Zhang et al., 2024) has primarily focused on improving the temporal reasoning capabilities of LLMs, without exploring how enhancing these abilities impacts downstream tasks, such as timeline summarisation.

Here we propose combining temporal reasoningwith timeline summarisation using LLMs, to enhance the generation of timeline summaries. Specifically, we first fine-tune a teacher model using a novel temporal reasoning dataset (NarrativeReason) and then distill temporal reasoning knowledge into a smaller student model, which is simultaneously fine-tuned on the timeline summarisation task.

We make the following contributions:

- We are the first to explore how enhancing temporal reasoning in LLMs can improve timeline summarisation.
- Based on the timelines created by NarrativeTime on the TimeBankNT corpus (Rogers et al., 2024), we develop a new dataset *NarrativeReason* for temporal reasoning. Unlike existing temporal reasoning datasets (Tan et al., 2023a; Chu et al., 2024; Wang and Zhao, 2024), *NarrativeReason* focuses on the temporal relationships among a series of events within a story rather than event pairs. This can help LLMs process a series of events to generate a coherent and accurate timeline summary.
- We apply the fine-tuned LLM to a completely different domain from what it is trained on, specifically generation of mental health related timeline summaries. Experimental results show that our model achieves the best performance on the timeline summarisation dataset by (Song et al., 2024). Not only does it generate more accurate summaries but it also reduces hallucinations in LLMs.
- We show why knowledge distillation works well, and how it induces better learned representations, through activation analysis of the fine-tuned LLM.

2 Related Work

Temporal reasoning for LLMs Temporal reasoning in Natural Language Processing (NLP) is the ability to understand and process information related to time within natural language text. It includes reasoning about the chronology and duration of events, and understanding and capturing different temporal relations (Vashishtha et al., 2020). Despite the impressive performace of Large Language Models (LLMs), like GPT-4, across a wide range of tasks (e.g. translation, generation), they have been shown to perform suboptimally in temporal reasoning (Wang and Zhao, 2024; Chu et al., 2024; Qiu et al., 2023). However the ability to perform temporal reasoning is crucial for understanding narratives (Nakhimovsky, 1987; Jung et al., 2011a; Cheng et al., 2013), answering questions (Bruce, 1972; Khashabi, 2019; Ning et al., 2020), and summarising events (Jung et al., 2011b; Vashishtha et al., 2020). Consequently, efforts are being made to enhance the temporal reasoning capabilities of LLMs. (Xing and Tsang, 2023) (Huang et al., 2024). To increase understanding of temporal expressions, Tan et al. introduced the TEMPREASON dataset which addresses three types of relations (time-time, time-event, event-event). TEMPREASON was used to fine-tune a LLM to improve the model's temporal reasoning, and they analysed the performance of different LLMs on this dataset showing it is challenging for LLMs to capture the temporal relationships between different events. Xiong et al. use the aligned timeline to help the LLM to understand temporal reasoning by translating the context into a temporal graph,identifying valid time expressions and generating related temporal knowledge. The temporal relationship between events is inferred based on specific times (e.g., the year the events occurred). However in a narrative, events often occur without a clear indication of time.

Temporal reasoning for summarization Jung et al. developed a natural language understanding (NLU) system with a temporal reasoning component to create comprehensive timelines, applied tomedical records, presenting medical history in a more intuitive way. They found that temporal reasoning in NLU is tightly integrated into the NLP system's deep semantic analysis. Temporal reasoning can help the LM analyze temporal relationships between different events, which is very beneficial for event or news summarisation (Vashishtha et al., 2020). However, few studies explore how improvements in temporal reasoning in LLMs directly benefit downstream tasks such as text summarisation.

3 Methodology

Task Given an individual's timeline (a series of posts between two dates (Tsakalidis et al., 2022)), the goal is to generate an abstractive summary that reflects changes in the individual over time (Song et al., 2024).

3.1 Proposed architecture

To generate timeline summaries on social media we consider two sub-processes (see Fig. 1):

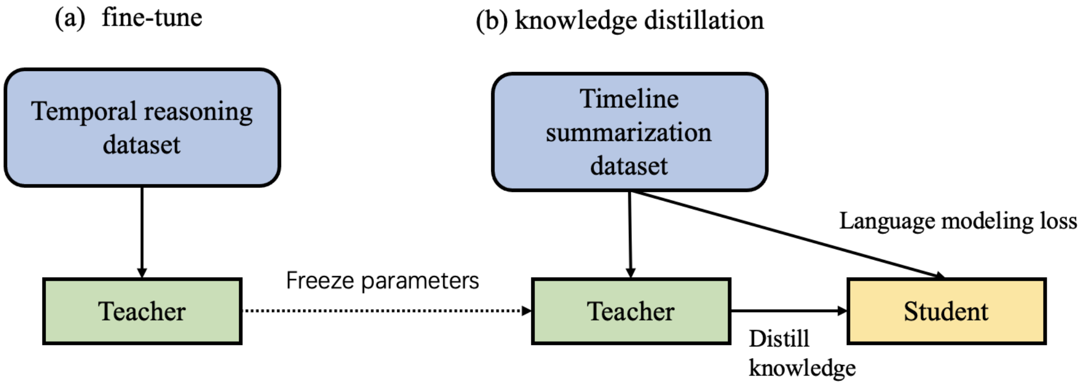


Figure 1: Overview of proposed method. (a) represents fine-tuning the teacher model on the temporal reasoning dataset.

(1) Improving temporal reasoning. We fine-tune a large LLM as a teacher model on the 'NarrativeReason' dataset §3.2.
(2) After fine-tuning the teacher model, we freeze its parameters. At this stage, we fine-tune a student model (a smaller LLM) on the the task of timeline summarisation (Chen et al., 2023a) from news. During this process, the teacher model transfers temporal reasoning knowledge to the student model, while the student simultaneously leverages the acquired temporal reasoning knowledge to perform timeline summarisation. For knowledge distillation (KD), we adopt three different strategies: Neuron Selectivity Transfer (NST), Contrastive Representation Distillation (CRD) and Probabilistic Knowledge Transfer (PRT). When generating the timeline summary, we conduct experiments on the TalkLife dataset, a very different domain to the one the student is trained on. We prompt the student model to generate mental health related summaries pertaining to aspects such as diagnostic states, inter- and intra- personal relationships and fluctuations in mood (Song et al., 2024).

3.2 Teacher Model

Here the goal is to improve an LLM's temporal reasoning. There is evidence showing that fine-tuning on datasets such as TEMPLAMA may enable an LLM to memorise the most frequent answer rather develop temporal reasoning (Tan et al., 2023b). In other words, the model does not truly understand temporal relationships, such as "before" and "after". Most temporal reasoning datasets involve pairs of events rather than multiple events. However, processing a sequence of events requires more intricate reasoning, including recognising patterns, dependencies, and causal chains among multiple events. This is useful for more sophisticated tasks such as narrative comprehension and timeline summarisation, where understanding the full sequence of events is crucial.

To prevent the LLM from learning shortcuts and memorising the most frequent answer, we created a temporal reasoning dataset *NarrativeReason*, which contains relationships between a series of events based on a given story.

Event extraction To create NarrativeReason we restructured the NarrativeTime dataset (Rogers et al., 2024), which re-annotated TimeBankDense (Cassidy et al., 2014) with a timeline-based annotation framework, NarrativeTime. They annotated all possible temporal links (TLINKS) and the temporal relationships of all events occurring within astory. They thus provide a complete temporal relation for the sequence of events in the text, rather than just the temporal relation between event pairs. That means they provide a clear timeline for these events. However, they only used one word (e.g. "fallen") to represent an event, which does not fully capture the complete picture of the event (e.g. "the Indonesian stock market has fallen by twelve percent"). In general, an event is considered an action involving its corresponding participants, and is marked by annotating a representative expression of the event. In the context of temporal reasoning, events are usually the head of the verb phrase. (Ning et al., 2018; Pustejovsky et al., 2003).

Therefore, we filtered the dataset TimeBankNT, keeping only annotated verbs to represent events. In order To represent a complete event, we use these verbs as triggers to extract relational triplets e.g. *<Indonesian stock market value, fallen, by twelve percent>* to represent the event *fallen*. Then, we use these triplets to construct the temporal relationship of a series event. (e.g. *Event <Indonesian*

stock market value, fallen, by twelve percent> is BEFORE Event <financial week, turning, bad for Asia>).

Dataset construction In a story, we consider the temporal relations for all events, and construct event-event relation question/answer pairs, which addrees the chronological relation between events, such as 'before', 'after', 'during', and 'simultaneous' (Tan et al., 2023a). Specifically, we obtain the temporal relations of all events and then use **questioanswering** prompts to reconstruct the dataset. **Question**: your task is to identify the temporal relation between *EVENT A* and *EVENT B*: based on the Story: *STORY*. **Answer**: *EVENT A* temporal relation (BEFORE/ AFTER/ INCLUDES/ IS_INCLUDED/ SIMULTANEOUS) *EVENT B*. Although a single question-answer pair is used to determine the temporal relationship between a pair of events, for a complete story, we construct multiple question-answer pairs to cover the temporal relationsamong all events. This ensures that the model can clearly outline the temporal relations between all events in the story.

Fine-tuning task We apply supervised fine-tuning (SFT) on a large LLM (teacher model) utilising Low-Rank Adaptation (LoRA) (Hu et al., 2022). The input and output of the model are the temporal questions and corresponding answers. Our experiments reveal that fine-tuning on this dataset brings improvements to other temporal reasoning tasks (Appendix A.2) .

3.3 Student Model

After we fine-tune a teacher model on the reconstructed dataset, we transfer the temporal reasoning knowledge to a student model, while, also fine-tuning the student on a news timeline summarisation dataset (Chen et al., 2023b). Thus we aim for the student to learn temporal reasoning and use this ability to generate timeline summaries. We fine-tune Phi-3-mini-4k-instruct as a student model. We use three knowledge distillation (KD) objectives to transfer knowledge from the teacher to the student: Neuron Selectivity Transfer (NST) (Huang and Wang, 2017), transfers heatmap like spatial activation patterns of teacher neurons to student neurons; Contrastive Representation Distillation (CRD) (Tian et al., 2019), maximises the mutual information between the teacher and student representations with contrastive learning; Probabilistic Knowledge Transfer (PRT) (Passalis and Tefas, 2018), matches the probability distribution of the data in the feature space.

PRT: The core idea is to match the probability distribution of the data in the feature space between teacher and student models. However, learning a significantly smaller model that accurately recreates the whole geometry of a complex teacher model is often impossible. (Passalis and Tefas, 2018) uses the conditional probability distribution to describe the samples. Here, $\mathbf{Y_t} = \{\mathbf{y_t}_1, \mathbf{y_t}_2, ..., \mathbf{y_t}_l\} \in \mathbb{R}^{vocab_t}$ denote the output logits of the teacher model, and $\mathbf{Y_s} = \{\mathbf{y_s}_1, \mathbf{y_s}_2, ..., \mathbf{y_s}_l\} \in \mathbb{R}^{vocab_s}$ denote the output logits of the student model, where $\mathbf{y_t}/\mathbf{y_s}$ is vector and l is the length of sentences, $vocab_t$ and $vocab_t$ are the vocabulary sizes of teacher and student models respectively. We can define the conditional probability distribution for the teacher model as Eq. 1, and student model as Eq. 2.

$$p_{i|j} = \frac{K(\mathbf{y_t}_i, \mathbf{y_t}_j; 2\sigma_t^2)}{\sum_{k=1,k\neq j}^{l} K(\mathbf{y_t}_k, \mathbf{y_t}_j; 2\sigma_t^2)} \quad (1)$$

$$q_{i|j} = \frac{K(\mathbf{x_s}_i, \mathbf{x_s}_j; 2\sigma_s^2)}{\sum_{k=1,k\neq j}^{l} K(\mathbf{x_s}_k, \mathbf{x_s}_j; 2\sigma_s^2)} \quad (2)$$

In these two equations (Eq. 1 and Eq. 2), we use the cosine similarity as a kernel metric allows for more robust affinity estimations.

$$K_{consine}(\mathbf{y_t}_i, \mathbf{y_t}_j) = \frac{1}{2}(\frac{\mathbf{y_t}_i^{\mathrm{T}} \mathbf{y_t}_j}{\|\mathbf{y_t}_i\|_2 \|\mathbf{y_t}_j\|_2} + 1) \in [0, 1].$$

We use Kullback-Leibler (KL) divergence to calculate the distance between the conditional probability distributions of the teacher model and the student model, as follows:

$$L_{PKT} = \sum_{i=1}^{l} \sum_{j=1,i\neq j}^{l} p_{i|j} \log(\frac{p_{i|j}}{q_{i|j}}).$$

NST matches the distributions of neuron selectivity patterns between teacher and student networks. We transfer the last hidden layer $\mathbf{T} = \mathbf{t}(\mathbf{x})$ of the teacher model to the last hidden layer $\mathbf{S} = \mathbf{s}(\mathbf{x})$ of the student model given input text $\mathbf{x}$. Specifically, we transfer neuron selectivity knowledge from $\{\mathbf{t}(\mathbf{x})_{*,i}\}_{i=1}^{N}$ to $\{\mathbf{s}(\mathbf{x})_{*,i}\}_{i=1}^{M}$, where N and M are the hidden state dimensions. Then we follow the idea in (Huang and Wang, 2017) using Maximum Mean Discrepancy (MMD) to calculate the distance between these two distributions. The basic idea of MMD is that if the moments of two

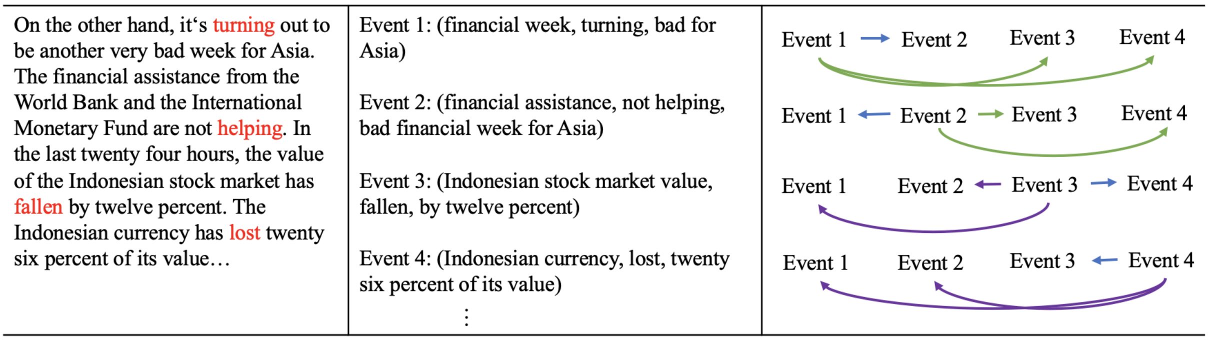


Figure 2: The temporal relationship between events. We represent events triggered by verbs using relational triplets, as shown in the middle column. In the right column, we list a series of events and use arrows of different colors to represent the temporal relations between these events. In this example, we use blue, green and purple arrows to indicate 'Simultaneous', 'After' and 'Before' respectively.

random variables are the same for all orders, then the two distributions are identical $L_{MMD^2}(\mathbf{t},\mathbf{s})$.

$$L_{MMD^2}(\mathbf{t},\mathbf{s}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N} k\left[\mathbf{t}(\mathbf{x})_{*,i};\mathbf{t}(\mathbf{x})_{*,i'}\right] + \frac{1}{M^2}\sum_{j=1}^{M}\sum_{j'=1}^{M} k\left[\mathbf{s}(\mathbf{x})_{*,j};\mathbf{s}(\mathbf{x})_{*,j'}\right] - \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} k\left[\mathbf{s}(\mathbf{x})_{*,i};\mathbf{t}(\mathbf{x})_{*,j}\right].$$

In this formula we use a Gaussian Kernel $k(x,y) = exp(-\frac{||x-y||_2^2}{2\sigma^2})$ with $\sigma = 1$. We transfer the teacher activation patterns to the student by minimizing L_{MMD^2}.

CRD maximises the lower-bound to the mutual information between the teacher and student representations. Here, we follow (Tang et al., 2021) to sample one positive pair from the joint distribution $p(\mathbf{S},\mathbf{T}) = q(\mathbf{S},\mathbf{T}|positive)$ for N negative pairs sampled from the product of marginal $p(\mathbf{S})p(\mathbf{T}) = q(\mathbf{S},\mathbf{T}|negtive)$, where N is the batch size. We can maximize the lower bound of mutual information by minimizing the following loss function:

$$\begin{aligned} L_{CRD}(\mathbf{x}) = -\mathbb{E}_{q(\mathbf{s},\mathbf{t}|positive)}\left[\log h(\mathbf{s},\mathbf{t})\right] \\ -N.\mathbb{E}_{q(\mathbf{s},\mathbf{t}|negtive)}\left[\log(1-h(\mathbf{s},\mathbf{t}))\right] \end{aligned} \quad (3)$$

In Eq 3, function h should satisfy $h : \{\mathbf{s},\mathbf{t}\} \rightarrow [0,1]$,

$$h(\mathbf{s},\mathbf{t}) = \frac{exp(\mathbf{s}^{\mathrm{T}}\mathbf{t})}{exp(\mathbf{s}^{\mathrm{T}}\mathbf{t}) + \frac{N}{M}},$$

where M is the cardinality of the dataset, and we need to normalize $\mathbf{s}$ and $\mathbf{t}$ by L-2 norm before the inner product.

The knowledge distillation method transfers temporal reasoning knowledge from the teacher model to the student model. At the same time, we aim for this knowledge to benefit the timeline summarization task. Therefore, we fine-tune the student model on the timeline summarization dataset §4.1, enabling it to simultaneously learn from the teacher model and use the language modeling loss (for next token prediction) $L_{language}$ to integrate temporal reasoning knowledge with timeline summarization information.

3.4 Mental Health Timeline Summary

We apply the student model to other domains, specifically to generate mental health-related summaries for timelines §4.1 from social media. For the mental health summary, we use the format proposed by Song et al. (2024), which includes three key clinical concepts (diagnosis, inter- and intra-personal relations, moments of change). We follow their method to prompt the student model to generate a summary for each timeline.

4 Experiments

4.1 Datasets

We conduct experiments on three different datasets. We fine-tune the teacher model on the 'NarrativeReason' dataset. When distilling the temporal reason knowledge to the student model, we use a news timeline summarisation dataset(Chen et al., 2023a). Finally, we apply the model to a different domain, specifically generating mental health-related timeline summaries from social media.

NarrativeReason We extracted 668 events from 30 articles, containing a total of 19,614 temporal relations between events Rogers et al. (2024), leading to 19,614 question/answer pairs for event-event relations. We use these question/answer pairs to fine-tune the teacher model to enhance its temporal reasoning capability.

Timeline summarisation Dataset This timeline summarisation dataset is sourced from (Chen et al., 2023a). They collected timeline summaries from Wikipedia websites, with a total of 5,000 timelines and summaries.

TalkLife When generating the summary, we use the dataset collected by Tsakalidis et al. (2022) comprising 500 anonymised user timelines from Talklife. Song et al. (2024) sample 30 timelines from it and annotate them with corresponding mental health-related summaries. These human-written summaries include diagnosis, intra- and interpersonal patterns and mental state changes over time, and they also highlight information related to individuals' mental states, which can be used for automated evaluation.

4.2 Models & Baselines

We compare our method against existing LLMs for mental health related summarisation. We introduce the implementation details in the appendix A.1

L-phi This model uses LLaMA as the teacher model and a smaller model, Phi, as the student. We apply different knowledge distillation (KD) methods to transfer temporal reasoning knowledge to the student model. Subsequently, we directly prompt this model to generate mental health-related timeline summaries.

P-phi In this model, we use Phi as the teacher model and another Phi of the same size as the student model.

Phi$_{joint}$ To compare with the phi-phi, we use joint learning to fine-tune Phi on both the NarrativeReason and Timeline summarisation datasets. This allows us to observe whether knowledge distillation (KD) outperforms directly fine-tuning on the two datasets. In addition, we fine-tune Phi on the NarrativeReason and timeline summarisation datasets and obtain models **Phi**$_{temp}$ and **Phi**$_{timeline}$ separately. We use these two models to generate timeline summaries for comparison with the KD derived model.

Phi$_{ICL}$ We use in context learning (ICL) to guide Phi to generate summaries. We provide the model with a pair consisting of a timeline and its corresponding summary as an example, and then let it generate summaries for other timelines.

LLaMA We prompt LLaMA to generate mental health related summaries on timelines directly as a baseline.

4.3 Evaluation

We use the timeline summaries from (Song et al., 2024)(§4.1) forautomated evaluation. We employ*Factual Consistency (FC)*, to measure whether timeline summaries are consistent with the original timelines, and *Evidence Appropriateness (EA)*, to measure the consistency of human written summaries with their corresponding timeline summaries (Song et al., 2024).

For human evaluation, we worked with two clinical psychology graduate students fluent in English to evaluate 30 summaries generated from 30 timelines (TalkLfe). We follow the metrics used in (Song et al., 2024) to evaluate the summaries from the perspectives of *Factual Consistency* and *Usefulness* (general/diagnosis/inter-&Interpersonal/MOC). A factually consistent summary should accurately represent the content of the timeline and excludes any information not present in the timeline. We also evaluate the usefulness of a mental health summary on the basis of four aspects: *General* (contains the most clinically important information from the timeline and aids the clinician in understanding a patient's condition); *Diagnosis* (provides useful information about an individual's diagnosis); *inter-&Interpersonal* (provides helpful information about an individuals' main needs and patterns of self and other relationships); *MOC* (provides useful information about an individual's changes over time in terms of emotion/cognition and behaviour).

5 Results and Discussion

5.1 Automatic evaluation

We conducted experiments with different combinations of KD methods. Table 1 shows experiment results. Since we did not change the output dimensions when fine-tuning LLaMA, the CRD method was not used during the KD process. Among the individual methods, PKT performed the best; however, combining PKT with NST achieved the best overall results.

Table 2 shows the results of applying alternative fine-tuning strategies on LLMs. The results for **Phi**$_{temp}$ and **Phi**$_{timeline}$ on FC indicate that fine-tuning on a single dataset does not improve model performance; instead, it exacerbates hallucination

Metric	P-phi$_{NST}$	P-phi$_{CRD}$	P-phi$_{PRT}$	P-phi$_{NST\&CRD}$	P-phi$_{NST\&PRT}$	P-phi$_{PRT\&CRD}$
FC	.344	.369	.378	.424	**.438**	.345
EA	.968	.954	.965	.969	**.973**	.961
Metric	L-phi$_{NST}$	L-phi$_{PRT}$	L-phi$_{NST\&PRT}$	LLaMA	–	–
FC	.367	.385	**.397**	.372	–	–
EA	.968	.966	**.971**	.956	–	–

Table 1: Automatic evaluation for factual consistency (FC), evidence appropriateness (EA) on student model. Higher is better, best in **bold**

issues. Notably, **Phi**$_{temp}$ performed the worst in both FC and EA metrics, suggesting that the incorporation of temporal reasoning information appears to interfere with the LLM's ability to effectively handle the timeline summarization task. Additionally, in strategy **Phi**$_{jont}$, combining the two types of data directly during training failed to integrate them effectively. As a result, the performance of **Phi**$_{jont}$ was even worse than using in-context learning to guide the LLM (**Phi**$_{ICL}$).

Metric	Phi$_{ICL}$	Phi$_{temp}$	Phi$_{timeline}$	Phi$_{joint}$
FC	**.412**	.141	.184	.238
EA	.965	.895	**.966**	.941

Table 2: Automatic evaluation for factual consistency (FC), evidence appropriateness (EA) across the compared models. Higher is better, best in **bold**

5.2 Human evaluation

Based on the results of the automatic evaluation, we selected the best-performing LLaMA-Phi and Phi-Phi models. Additionally, we included the non-fine-tuned versions of LLaMA and Phi. This can help us understand in which specific aspects the model has improved with the inclusion of temporal reasoning information. The fine-tuned model shows the greatest improvement in terms of factual consistency and usefulness (general). This aligns with our findings when analyzing the summaries, where the fine-tuned model significantly reduces hallucination, as shown in Table 3. In addition, we found that the fine-tuned model did not show significant improvement in terms of Moments of Change (MOC). This could be because, in the timelines, users do not experience many emotional switches, resulting in the model-generated summaries being relatively similar.

5.3 Why knowledge distillation works

In this section, we analyze why the Phi$_{timeline}$ model performed better from a representation learning perspective. For this purpose, we perform two experiments, a task understanding probing experiment and a Joint Task Representation Learning (JTRL) experiment. We analyze the internal representations of Phi$_{timeline}$ against Phi$_{joint}$. We construct our probing dataset by using Phi$_{joint}$'s test dataset. We pose the two tasks as a binary classification problem and then extract the activations for the last layer. Then we use UMAP (McInnes et al., 2018) to project these activations to lower dimensions (Sainburg et al., 2021; Tseriotou et al., 2023). From Figure 3, we can see that the activations of Phi$_{joint}$ are well separated for each task, whereas the activations for Phi$_{timeline}$ overlap. Given the performance of Phi$_{timeline}$, this leads us to hypothesize that the model learned better representations for the task due to more task specific polysemantic (Olah et al., 2020) neurons.

Aspect	phi	phi-phi	LLaMA	LLaMA-phi
Factual Consistency	2.90	3.32	3.58	**3.83**
Usefulness (General)	2.60	3.13	3.17	**3.48**
(Diagnosis)	2.90	3.37	3.45	**3.62**
(Inter-& Intrapersonal)	2.95	3.00	3.40	**3.51**
(MoC)	2.97	2.97	3.42	**3.47**

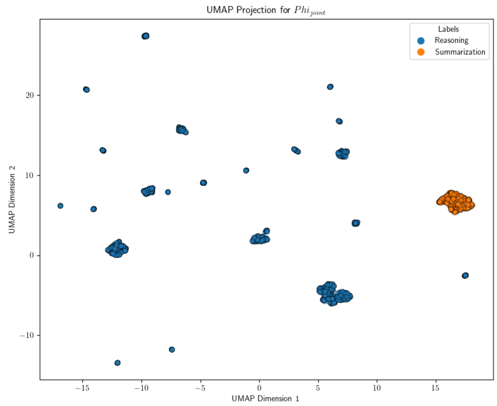
Table 3: Human evaluation results based on 5-point Likert scales (1 is worst, 5 is best). Best in **bold**.

To validate our hypothesis, we ran another set of experiments to analyze the internal representation difference between the two models, which we termed as JTRL. Our main idea here was to validate that knowledge distillation resulted in better representations due to more polysemantic neurons, and these representations varied highly compared to Phi$_{joint}$'s representation. To measure JTRL, we use the Centered Kernel Alignment (CKA) (Kornblith et al., 2019) similarity score. CKA can be used to measure the similarity between internal representations of modelsConneau et al. (2020); Muller et al. (2021); Del and Fishel (2021); Moosa et al. (2023).

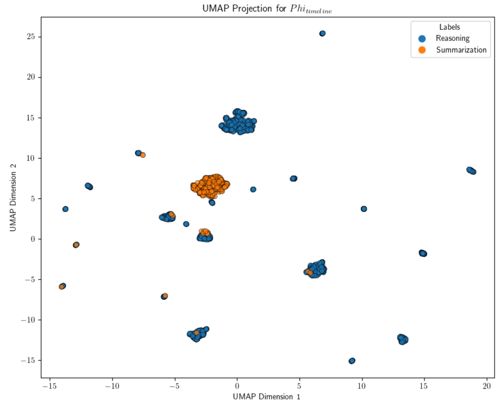
To calculate CKA, we used the same probing dataset. First, we calculate the sentence embeddings of each input by averaging the hidden state representation of the tokens. Then, we calculate the CKA similarity score between the mean sentence embeddings and each layer representation of the model. We did this both for the individual models and between the layers of the models. From Figure 4, we can see that Phi$_{timeline}$ shows a gradual increase in CKA across layers, peaking in the mid-to-late layers. This indicates that as the layers progress, Phi$_{timeline}$ preserves and refines the information from the initial embeddings in a task-relevant way. The is likely due to the fact that the distillation process encourages this alignment by transferring task-relevant knowledge from the teacher. On the other hand, Phi$_{joint}$ shows an initial increase but saturates and flattens early. This indicates that it fails to refine representations effectively in deeper layers, which could be due to conflicting objectives between the tasks. The lower CKA in later layers suggests that the model moves away from the initial embeddings in a manner that is less effective for task-specific learning. Lastly, the CKA values between the models show that they learn vastly different representations. In short, Phi$_{timeline}$'s ability to align with the initial embeddings correlates with its better task performance, as the CKA value reflects how well the model retains and transforms meaningful input information throughout its layers.

6 Conclusions

We created a dataset named *NarrativeReason*, which is utilised to enhance the temporal reasoning abilities of LLMs. This dataset enables models to uncover temporal relationships between events within timelines. We fine-tune a larger LLM with *NarrativeReason*, then distill its enhanced temporal reasoning capability to a smaller LLM while leveraging the distilled knowledge to enhance performance in timeline summarisation tasks. We apply the model to a different domain from the one it was trained on, namely generating mental health related timeline summaries. Our results demonstrate that our approach using knowledge distillation produces more accurate summaries while significantly reducing hallucinations. We further analyzed the internal representations of the model and found that knowledge distillation leads to better feature representations within the model that are more aligned with the task of timeline summarisation.



(a) UMAP projection Phi$_{joint}$



(b) UMAP projection Phi$_{timeline}$

Figure 3: The UMAP projection forPhi$_{joint}$ and Phi$_{timeline}$ show the last layer activations for both models. We can see that Phi$_{timeline}$ has more polysemantic activations compared to Phi$_{joint}$.

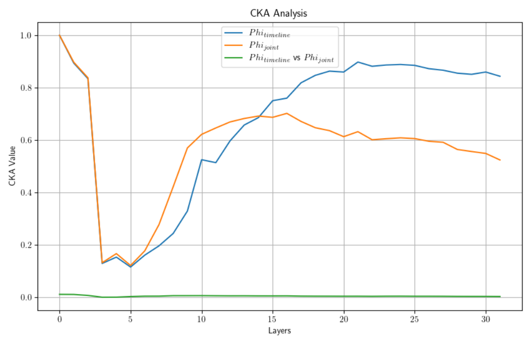


Figure 4: CKA similarity score of both within and between Phi$_{timeline}$ and Phi$_{joint}$ model representations.

Limitations

In our work we aim to leverage temporal reasoning to enhance the performance of LLMs in timeline summarisation tasks. We apply this to social media timelines from the mental health domain to enable LLMs to recognize and analyze events chronologically, in order to capture the dynamics of user behavior and evolving mental states more effectively. This faces the following limiting factors: (a) the existing knowledge of mental health embedded in the LLM and (b) the fixed formats of mental health-related texts that the model is trained on. By analyzing the generated summaries, we found that there seems to be a general tendency to make clear statements about specific DSM diagnoses (such as PTSD, bipolar disorder, etc.). In the vast majority of cases it is possible to write that there is evidence that can indicate such a potential, instead of providing a definite assessment. Moreover, many parts of the summaries seem very generic. This lack of personalisation can sometimes lower the quality of the summary. While this may not always have a significant impact, it generally reduces the depth and individuality of the analysis, sometimes even affecting the factual consistency.

These findings can help the exploration of LLMs in the future, particularly in the mental health domain. They can help refine future models that better understand social media posts, leading to more accurate mental health summaries and improved diagnostic insights for clinicians.

Ethics Statement

Ethics institutional review board (IRB) approval was obtained from the corresponding ethics board of the lead University prior to engaging in this research study. Our work involves ethical considerations around the analysis of user generated content shared on a peer support network (TalkLife). A license was obtained to work with the user data from TalkLife and a project proposal was submitted to them in order to embark on the project.

The final summaries in all cases are obtained by feeding the timeline summaries into an LLM. Given that LLMs are susceptible to factual inaccuracies, often referred to as 'hallucinations,' and tend to exhibit biases, the clinical summaries they generate may contain errors that could have serious consequences in the realm of mental health decision-making. These inaccuracies can encompass anything from flawed interpretations of the timeline data to incorrect diagnoses and even recommendations for potentially harmful treatments. Mental health professionals must exercise caution when relying on such generated clinical summaries. These summaries should not serve as substitutes for therapists in making clinical judgments. Instead, well-trained therapists must skillfully incorporate these summaries into their clinical thought processes and practices. Significant efforts are required to establish the scientific validity of the clinical benefits offered by these summaries before they can be integrated into routine clinical practice.

References

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

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 501–506. The Association for Computer Linguistics.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721.

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2024. Improving large language models in event relation logical prediction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9451–9478. Association for Computational Linguistics.

Xiuying Chen, Mingzhe Li, Shen Gao, Zhangming Chan, Dongyan Zhao, Xin Gao, Xiangliang Zhang, and Rui Yan. 2023a. Follow the timeline! generating an abstractive and extractive timeline summary in chronological order. *ACM Trans. Inf. Syst.*, 41(1):9:1–9:30.

Xiuying Chen, Mingzhe Li, Shen Gao, Zhangming Chan, Dongyan Zhao, Xin Gao, Xiangliang Zhang, and Rui Yan. 2023b. Follow the timeline! generating an abstractive and extractive timeline summary in chronological order. *ACM Transactions on Information Systems*, 41(1):1–30.

Yao Cheng, Peter Anick, Pengyu Hong, and Nianwen Xue. 2013. Temporal relation discovery between events and temporal expressions identified in clinical narrative. *Journal of biomedical informatics*, 46:S48–S53.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1204–1228. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Maksym Del and Mark Fishel. 2021. Establishing interlingua in multilingual language models. *CoRR*, abs/2109.01207.

Yu Feng, Ben Zhou, Haoyu Wang, Helen Jin, and Dan Roth. 2023. Generic temporal reasoning with differential analysis and explanation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12013–12029. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Qisheng Hu, Geonsik Moon, and Hwee Tou Ng. 2024. From moments to milestones: Incremental timeline summarization leveraging large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7232–7246, Bangkok, Thailand. Association for Computational Linguistics.

Rikui Huang, Wei Wei, Xiaoye Qu, Shengzhe Zhang, Dangyang Chen, and Yu Cheng. 2024. Confidence is not timeless: Modeling temporal validity for rule-based temporal knowledge graph forecasting. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10783–10794, Bangkok, Thailand. Association for Computational Linguistics.

Zehao Huang and Naiyan Wang. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6750–6774, Singapore. Association for Computational Linguistics.

Hyuckchul Jung, James Allen, Nate Blaylock, William de Beaumont, Lucian Galescu, and Mary Swift. 2011a. Building timelines from narrative clinical records: Initial results based-on deep natural language understanding. In *Proceedings of BioNLP 2011 Workshop*, pages 146–154, Portland, Oregon, USA. Association for Computational Linguistics.

Hyuckchul Jung, James Allen, Nate Blaylock, William de Beaumont, Lucian Galescu, and Mary Swift. 2011b. Building timelines from narrative clinical records: initial results based-on deep natural language understanding. In *Proceedings of BioNLP 2011 workshop*, pages 146–154.

Daniel Khashabi. 2019. *Reasoning-Driven Question-Answering for Natural Language Understanding*. University of Pennsylvania.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Manling Li, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen R. McKeown. 2021. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6443–6456. Association for Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2023. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 670–685, Dubrovnik, Croatia. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Alexander Nakhimovsky. 1987. Temporal reasoning in natural language understanding: The temporal structure of the narrative. In *Third Conference of the European Chapter of the Association for Computational*

Linguistics, Copenhagen, Denmark. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, Brussels, Belgium. Association for Computational Linguistics.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Nikolaos Passalis and Anastasios Tefas. 2018. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–284.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2023. Are large language models temporally grounded? *arXiv preprint arXiv:2311.08398*.

Hossein Rajaby Faghihi, Bashar Alhafni, Ke Zhang, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2022. CrisisLTLSum: A benchmark for local crisis event timeline extraction and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5455–5477, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anna Rogers, Marzena Karpinska, Ankita Gupta, Vladislav Lialin, Gregory Smelkov, and Anna Rumshisky. 2024. Narrativetime: Dense temporal annotation on a timeline. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 12053–12073. ELRA and ICCL.

Tim Sainburg, Leland McInnes, and Timothy Q. Gentner. 2021. Parametric UMAP embeddings for representation and semisupervised learning. *Neural Comput.*, 33(11):2881–2907.

Jiayu Song, Jenny Chim, Adam Tsakalidis, Julia Ive, Dana Atzil-Slonim, and Maria Liakata. 2024. Combining hierachical vaes with llms for clinically meaningful timeline summarisation in social media. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14651–14672. Association for Computational Linguistics.

Julius Steen and Katja Markert. 2019. Abstractive timeline summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 21–31, Hong Kong, China. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023a. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14820–14835. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023b. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 24468–24481.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660, Dublin, Ireland. Association for Computational Linguistics.

Talia Tseriotou, Adam Tsakalidis, Peter Foster, Terence Lyons, and Maria Liakata. 2023. Sequential path signature networks for personalised longitudinal language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5016–5031, Toronto, Canada. Association for Computational Linguistics.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078, Online. Association for Computational Linguistics.

Yuqing Wang and Yun Zhao. 2024. TRAM: benchmarking temporal reasoning for large language models.

In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6389–6415. Association for Computational Linguistics.

Georg Wenzel and Adam Jatowt. 2023a. An overview of temporal commonsense reasoning and acquisition. *CoRR*, abs/2308.00002.

Georg Wenzel and Adam Jatowt. 2023b. An overview of temporal commonsense reasoning and acquisition. *CoRR*, abs/2308.00002.

Bowen Xing and Ivor W. Tsang. 2023. Relational temporal graph reasoning for dual-task dialogue language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13170–13184.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10452–10470. Association for Computational Linguistics.

Xinliang Frederick Zhang, Nick Beauchamp, and Lu Wang. 2024. Narrative-of-thought: Improving temporal reasoning of large language models via recounted narratives. *arXiv preprint arXiv:2410.05558*.

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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3361–3367. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

We use Meta-Llama-3-8B as teacher model and Phi-3-mini-4k-instruct as student model. We use Low-Rank Adaptation (LoRA) (Hu et al., 2022) for both of these two models while fine-tuning. We use an AdamW (Kingma and Ba, 2015) optimizer with learning rate 5e-5. While fine-tuning, we set the batch-size as 1 for each task, but set gradient accumulation steps as 16.

A.2 Experiment on teacher model

After fine-tuning the teacher model, we used the *TEMPREASON* dataset (Tan et al., 2023b) to evaluate its performance pre- and post-fine-tuning. Specifically, we focused on the *L3* part of the dataset, which deals with event-event relations, to determine whether the model could accurately infer the temporal sequence between two events. We use **F1** as the evaluation metrics, the experimental results show that the fine-tuned model achieved a 0.07 improvement in this metric compared to the pre-fine-tuning model, highlighting its enhanced ability to infer event-event temporal relations.