# Large Temporal Models: Unlocking Temporal Understanding in LLMs for Temporal Relation Classification

**Omri Homburger and Kfir Bar**
Efi Arazi School of Computer Science, Reichman University, Herzliya, Israel
omri.homburger@post.runi.ac.il, kfir.bar@runi.ac.il

## Abstract

We present *Large Temporal Model*, a Large Language Model (LLM) that excels in Temporal Relation Classification (TRC). We show how a carefully designed fine-tuning strategy, using a novel two-step fine-tuning approach, can adapt LLMs for TRC. Our approach is focused on global TRC, enabling simultaneous classification of all temporal relations within a document. Unlike traditional pairwise methods, our approach performs global inference in a single step, improving both efficiency and consistency. Evaluations on the MATRES and OmniTemp benchmarks demonstrate that, for the first time, an LLM achieves state-of-the-art performance, outperforming previous pairwise and global TRC methods. Results show that our global approach produces more consistent and accurate temporal graphs. Ablation studies further validate the effectiveness of our two-step fine-tuning strategy, while analyses reveal why our approach succeeds in increasing performance and reducing inconsistencies.

## 1 Introduction

Temporal Relation Classification (TRC), also referred to as Temporal Relation Extraction (TRE), is the task of identifying and classifying the temporal ordering between events mentioned in a text. Given a document annotated with a set of events—typically verbs—the goal is to predict the temporal relations (e.g., before, after, simultaneous) between all pairs of events, thereby constructing a directed graph that captures the event chronology. A practical application of TRC is the automatic construction of event timelines, which are valuable for various downstream tasks across domains such as historical analysis, news summarization, and clinical narratives (e.g., (Bakker et al., 2024; Sezgin et al., 2023)).

We distinguish between two operational modes of TRC: (1) *pairwise* TRC, where each event pair is classified independently, and (2) *global* TRC, where the model classifies all event pairs jointly in a single inference step. Pairwise TRC has been the predominant approach for many years and can be viewed as a conventional classification task applied repeatedly across event pairs. However, this approach often suffers from consistency issues, as it does not enforce logical coherence across the predictions—such as transitivity or temporal logic constraints. For instance, predicting that event A occurs before event B, and event B occurs before event C, should imply that event A occurs before event C, a constraint that pairwise models frequently violate.

In contrast, global TRC requires models capable of generating an entire temporal relation graph in one step. Large Language Models (LLMs), with their strong generative capabilities, are well-suited for this mode. Recent work (Eirew et al., 2025) has demonstrated the potential of LLMs in global TRC, even in zero-shot settings without fine-tuning.

Furthermore, a notable trend in recent studies involves augmenting models with external knowledge sources, a strategy shown to enhance model capabilities and yield superior performance. For instance, the integration of the ATOMIC-2020 commonsense knowledge graph (Hwang et al., 2021) in a previous work (Tan et al., 2023) has proven effective in infusing models with broad temporal understanding across diverse events, demonstrably leading to performance gains over models trained without such external knowledge.

Despite their potential, LLMs have not yet shown superior capabilities compared to modern methods in TRC. Prior research (Roccabruna et al., 2024; Yuan et al., 2023; Eirew et al., 2025) indicates that LLMs perform worse than current approaches, which typically use smaller encoder-only models, when classifying temporal relations. This performance gap has been observed in both zero-shot and fine-tuning settings.

Our work aims to fill this gap by investigating



*Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2156–2171

the effectiveness of fine-tuning LLMs. In this work, we demonstrate for the first time that a carefully designed fine-tuning strategy for TRC substantially boosts LLMs performance, outperforming current approaches in both pairwise and global settings.

We introduce a two-step fine-tuning strategy. In the first step, we inject a mixture of relevant datasets containing temporal information to build a general temporal understanding within the model. In the second step, we fine-tune the model specifically on the global TRC task using a parameter-efficient adaptation technique, which enables better control over the fine-tuning impact and preserves the general temporal reasoning acquired in the first phase.

We evaluate our approach on two benchmark datasets. The first, MATRES, includes annotations only between event pairs that occur within a two-sentence window, thus reflecting local temporal reasoning. The second, OmniTemp, is a more recent and comprehensive dataset that provides annotations for all possible event pairs within a document, enabling evaluation of truly global temporal reasoning. Our experiments show that our fine-tuned models outperform existing state-of-the-art methods on both datasets, highlighting the effectiveness of our approach in enhancing LLMs' capacity for TRC.

In our work, we make the following three contributions:

- We show, for the first time, how to fine-tune LLMs to achieve superior capabilities on TRC, by applying the two-step fine-tuning approach. We demonstrate that this approach achieves new state-of-the-art performance on standard benchmarks for both pairwise and global TRC approaches.

- We perform ablation studies and qualitative analyses to understand why incorporating diverse external temporal knowledge, as opposed to using core TRC data only, can mitigate the inferior LLM performance on TRC, particularly for the complex global TRC task.

- We provide quantitative evidence that when performing global TRC, our model produce more coherent graph, compared to the pairwise approach. This highlights the value of a holistic, document-level prediction strategy.

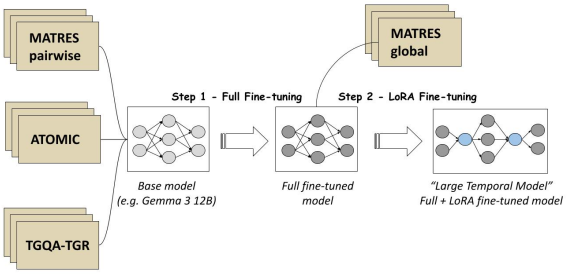


Figure 1: Our 2 step approach. In the first step, we perform a full fine-tuning with data from 3 sources - MATRES (pairwise), ATOMIC-2020 and TGQA-TGR. In the second step, we perform a LoRA fine-tuning with data from MATRES (global).

## 2 Related work

TRC is a well-established and active research area within the global natural language processing (NLP) field. Over the past decade, many efforts have been dedicated to developing robust methods for this task.

Recent approaches (e.g. (Tan et al., 2023; Wu et al., 2025; Ning et al., 2024; Cohen and Bar, 2023; Zhang et al., 2022; Tan et al., 2021)) have leveraged neural architectures, mainly encoder-based models like BERT (Devlin et al., 2019) and its variants. These models are typically fine-tuned on TRC-specific datasets, learning to classify the relationship between pairs of marked events or time expressions within a given context.

Some studies (Tan et al., 2023; Zhuang et al., 2023) have explored enhancing these models by incorporating external knowledge sources during training. Notable examples include leveraging commonsense knowledge graphs like ATOMIC-2020 (Hwang et al., 2021), which contains relevant event-centric temporal relations, or utilizing datasets designed for related temporal tasks, such as TGQA-TGR (Xiong et al., 2024), originally created for temporal graph reasoning.

Much of the research has focused on a pairwise TRC approach, where each event pair is classified independently. The most commonly used benchmarks for this paradigm are MATRES (Ning et al., 2018) and TBDense (Cassidy et al., 2014). While both contain annotated news articles, they differ in scale and relation granularity. MATRES offers a larger corpus (275 documents) and uses a set of four relations (BEFORE, AFTER, EQUAL, VAGUE). TBDense, though smaller (36 documents), includes two additional relations (IN-

CLUDES, IS_INCLUDED). The standard evaluation metric is typically the micro-averaged F1 score, although protocols can vary. Most contemporary work follows the practice established by Ning et al. (2019), where VAGUE relation is excluded during evaluation but is used during training. However, alternative protocols exist, such as removing VAGUE instances entirely (Alsayyahi and Batista-Navarro, 2023).

The emergence of LLMs has introduced new opportunities for advancing TRC. Initial explorations have investigated zero-shot or few-shot capabilities. For instance, Yuan et al. (2023) proposed a chain-of-thought prompting strategy to elicit temporal relations from LLMs without task-specific training.

Fine-tuning LLMs specifically for TRC is still emerging. One study (Roccabruna et al., 2024) applied LoRA (Hu et al., 2022) to fine-tune LLMs for pairwise TRC, finding that this approach did not surpass fine-tuned encoder models under their experimental setup. We note that while direct comparison to this study is not possible due to a different protocol used, where VAGUE label is removed entirely, it demonstrated subpar LLMs capabilities compared to encoder-based models in TRC when using a standard training approach.

Concurrently, while traditional approaches to TRC often focus on pairwise classification of adjacent events, their limitations, particularly in generating globally inconsistent temporal graphs, have spurred interest in building more robust and consistent temporal structures. Several studies (Wang et al., 2020; Mathur et al., 2021; Yao et al., 2024) have addressed these issues and proposed mitigation strategies. For example, Niu et al. 2024 proposed a method to increase the model's understanding of temporal relations by focusing on their symmetric or antisymmetric properties.

A recent work (Eirew et al., 2025) highlighted the benefits of a global approach to TRC, which aims to predict the entire set of relations within a document simultaneously and helps in reducing inconsistencies within a temporal graph. It also introduced the OmniTemp dataset, which features comprehensive document-level annotations suitable for this paradigm. A similar effort is NarrativeTime (Rogers et al., 2024), a comprehensive annotation of the TB-Dense corpus, covering all possible event pairs.

## 3 Methodology

We develop a "Large Temporal Model" (LTM), which is an LLM capable of performing TRC. We propose a two-step fine-tuning approach designed to effectively adapt LLMs for the task of TRC, as demonstrated in Figure 1.

### 3.1 Step 1: Temporal Domain Adaptation Full Fine-tuning

The primary objective of this initial step is to instill broad temporal reasoning capabilities within the LLM. We achieve this through comprehensive full fine-tuning on a diverse corpus aggregated from multiple datasets relevant to temporal understanding. This initial full fine-tuning step equips the model with foundational concepts necessary for the downstream task-specific adaptation. The aggregated corpus for this stage comprises approximately 18,000 instances, each contains instructions of a specific task, the input text, and the expected output. Examples of the instructions given to the model are presented in Table 1. The instances are constructed from the following three sources:

**MATRES (Ning et al., 2018).** We utilize a portion of the widely used MATRES dataset to introduce the model to the core concepts and the format of the TRC task. This dataset contains 275 news articles (compiled from TimeBank, AQUAINT, and PLATINUM) annotated with events. Each pair of events has a temporal relation—BEFORE, AFTER, EQUAL, or VAGUE—indicating the temporal ordering between them. Following standard practice, the PLATINUM section, which contains 20 articles, serves as the test set. Of the 255 remaining articles compiled from TimeBank and AQUAINT, we allocate 100 articles for training and 20 articles for validation which we use for this full fine-tuning step, reserving the remaining 135 articles for the second step. From these 120 articles, we extract all annotated event pairs, treating each pair as a separate instance. This yields 5,946 training instances and 586 validation instances, each containing the original article text annotated with event markers on all events (e.g., *"He <sold(ei391)> the property to five buyers and <said(ei392)> he..."*) and the label of a single event pair. The prompt for the model includes the instructions for the task, followed by the marked article and the pair to be labeled. The model is required to output the correct temporal relation between the two events. An example is provided in Figure 3 in Appendix E.

**ATOMIC-2020 (Hwang et al., 2021).** To improve the model's understanding of basic temporal ordering, we incorporate data from the ATOMIC-2020 common sense knowledge dataset. We specifically select instances corresponding to the event-centered *isBefore* and *isAfter* relations. These instances consist of two sentences, each provided with a single label indicating their temporal relation: *isBefore*, meaning the event described in the first sentence occurs before the second, and *isAfter*, meaning the event in the first sentence occurs after the one in the second. For example, *"PersonX calls the cable company isAfter PersonX can't watch TV."* For consistency with the MATRES dataset, we convert the *isBefore* and *isAfter* labels to *BEFORE* and *AFTER*, respectively. Of the 33,579 available instances for these relations, we use 30,000 for training and 3,575 for validation. To encourage the model to handle multiple temporal questions concurrently, we structure the training data by consolidating every five original ATOMIC instances into a single instance. This results in 6,000 composite training instances and 715 validation instances, where the model must predict five independent relations per instance. The prompt for the model includes the instructions for the task, followed by the five pairs of sentences. The model is required to output the correct five temporal relations. An example is provided in Figure 4 in Appendix E.

**TGQA-TGR (Xiong et al., 2024).** To expose the model to more complex temporal reasoning scenarios, we include the TGQA-TGR dataset, a synthetic corpus originally developed for temporal graph reasoning. Each sample presents a narrative alongside temporal queries about events in it (e.g., *"When did the event X start?"*, *"Given the following five events, which event is the second one in chronological order?"*) and their answers in natural language. We use the original training and test splits combined as our training data (6,110 instances) and the original validation split (698 instances) for validation during this stage. The prompt for the model contains the instructions for the task, followed by the narrative and a single question. The model is required to output the answer for the question asked. An example is provided in Figure 5 in Appendix E.

### 3.2 Step 2: Downstream-task Fine-tuning (TRC)

The second step focuses on adapting the model from Step 1 specifically to the target task of global TRC, where the goal is to predict the temporal relationships between all relevant event pairs within a given document simultaneously.

| Dataset | Prompt |
|---|---|
| MATRES | Given the text below where events are marked with <eventName(identifier)>, for the specified pair of events below, determine the temporal relationships (BEFORE, AFTER, EQUAL, VAGUE) between them. |
| ATOMIC | Given the pairs of sentences below, for each pair determine if the first sentence happened BEFORE or AFTER the second sentence. |
| TGQA-TGR | You are given the following text. Answer the question below. |

Table 1: Examples of instructions given to the model on each dataset used in training.

For this adaptation, we employ the parameter-efficient Low-Rank Adaptation (LoRA) fine-tuning technique (Hu et al., 2022). LoRA substantially reduces the number of trainable parameters and minimizes computational overhead by introducing low-rank decomposition matrices into the model layers.

The training set for this step consists of the remaining 135 training articles from the MATRES dataset that were not used in Step 1 (validation set is the same). In contrast to Step 1 and aligning with the global TRC task formulation, each training instance in this phase consists of the full article text with marked events, accompanied by the complete set of originally annotated event pairs. The model is expected to generate the full set of event pairs along with their correct labels in a single inference step. An example is provided in Figure 6 in Appendix E.

## 4 Experimental Setup

### 4.1 Models

Our experiments utilize two instruction-tuned, open-source LLMs: Meta's Llama 3.1 8B (Grattafiori et al., 2024) and Google's Gemma 3 12B (Team et al., 2025). We specifically use these open-source models because they are widely used and offer strong performance while remaining compact enough to run on a single GPU, substantially reducing training and serving costs. These models were trained using the Together AI[1] framework.

Additional details on the hyperparameters used during training and evaluation are provided in Ap-

[1] https://www.together.ai/

pendix A. Information on the total cost of training and evaluation is provided in Appendix B.

### 4.2 Datasets

We evaluate our proposed method on two publicly available benchmark test sets used for TRC. We provide additional information in Appendix C.

**MATRES (Ning et al., 2018).** While MATRES is used for training, we also include the MATRES dataset in our evaluation by utilizing articles from its original test set, known as the PLATINUM corpus. Notably, MATRES annotations are restricted to event pairs occurring in consecutive or nearby sentences, reflecting local temporal reasoning and leaving many other event pairs in the text unannotated.

**OmniTemp (Eirew et al., 2025).** This relatively new dataset builds upon the foundational structure of the MATRES dataset and employing the same core relations (i.e. BEFORE, AFTER, EQUAL, and VAGUE). Unlike MATRES, OmniTemp aims to annotate all valid event pairs within a document, irrespective of sentence distance. This denser, document-level annotation potentially introduces greater complexity, especially for global TRC approaches that consider all pairs simultaneously. Furthermore, given the model was not trained on such long distance relations poses additional complexity.

We note our initial attempt to utilize the official OmniTemp training set for the Step 2 task-specific fine-tuning of our global TRC model. However, this attempt proved unsuccessful, as the limited dataset size (20 documents only) was insufficient for the model to effectively learn the task-specific nuances.

### 4.3 Evaluation Metrics

**Micro-Averaged F1 Score.** Following standard practice in TRC evaluation (Ning et al., 2019), we use the micro-averaged F1 score as the primary measure of label classification performance. This metric is calculated only over the instances where the ground truth label is BEFORE, AFTER, or EQUAL, thereby excluding the ambiguous VAGUE category. We compare our results directly against previous works that adopt the same evaluation protocol.

For evaluating the model's performance on global TRC, we employ an inference strategy where the model processes each document comprehensively in a single call. The input to the model for each inference includes the full document text with annotated event mentions and a complete list of potential event pairs within that document. The model is expected to output a corresponding list enumerating each input pair along with its predicted temporal relation. This approach requires one inference call per document in the test set, resulting in a total of $N$ inferences, where $N$ is the number of documents.

In contrast, the evaluation of pairwise TRC involves assessing the model's ability to classify the temporal relation for individual event pairs independently. For each potential pair of events within a document, the model is invoked separately. The input for each inference consists of the document text containing the marked event mentions and the specific event pair under consideration. The model is required to output the temporal relation solely for the given pair. This fine-grained evaluation strategy leads to a significantly higher number of inference calls, roughly $N \times P$ inferences, where $N$ is the number of documents and $P$ is the average number of event pairs per document.

**Transitive Inconsistencies.** To assess the structural coherence and logical consistency of the temporal predictions, we also measure the number of transitive inconsistencies in the temporal graphs implicitly generated by our models predictions on the test sets. A transitive inconsistency occurs when predicted relations violate logical constraints (e.g., predicting A BEFORE B, B BEFORE C, and C BEFORE A). We compare this metric on both global and pairwise TRC.

To quantify the level of inconsistency within the predicted temporal relations, we adopt the methodology for calculating transitive inconsistencies as described in (Eirew et al., 2025). Specifically, for each document, we consider the complete set of temporal relations output by the model, either from the single inference pass in the global TRC approach or by aggregating the results from all pairwise inference calls in the pairwise TRC approach. We then iterate through all possible triplets of events $(e_i, e_j, e_k)$ within that document. For each triplet, we check for transitive inconsistencies among the predicted relations for the pairs $(e_i, e_j)$, $(e_j, e_k)$ and $(e_i, e_k)$, if these relations exist in the model's output for that document.

The number of such inconsistencies is counted

| Model | MATRES | OmniTemp |
|---|---|---|
| ZSL-GlobalConsistency | 63.0 | 74.5 |
| Gemma-3-12B - standard training | 66.0 | 52.9 |
| LTM - LLama-3.1-8B (ours) | 83.6 | 72.4 |
| LTM - Gemma-3-12B (ours) | **83.7** | **78.5** |

Table 2: Comparing micro-average F1 scores on MATRES and OmniTemp for global TRC.

| Model | MATRES | OmniTemp |
|---|---|---|
| Bayesian-Trans (Tan et al., 2023) | 82.7 | - |
| Unified-Framework (Huang et al., 2023) | 82.6 | - |
| OntoEnhance (Zhuang et al., 2023) | 82.6 | - |
| TCT (Ning et al., 2024) | 82.9 | - |
| GenTRE (Wu et al., 2025) | 83.5 | - |
| Bayesian + constraints (Tan et al., 2023) * | 79.2 | 80.7 |
| CoT (Yuan et al., 2023) * | 56.6 | 67.2 |
| LTM - Llama-3.1-8B (ours) | 84.8 | 81.3 |
| LTM - Gemma-3-12B (ours) | **85.0** | **82.9** |

Table 3: Comparing micro-average F1 scores on MATRES and OmniTemp for pairwise TRC. Models marked with (*) have their scores adapted from (Eirew et al., 2025).

for each document. The total number of transitive inconsistencies across the entire test set is then calculated by summing the inconsistency counts from all individual documents.

## 5 Results

### 5.1 Classification Performance

**Global Methods.** Table 2 compares our global TRC approach against a prior study that utilized larger LLMs for global TRC without fine-tuning (Eirew et al., 2025). To the best of our knowledge, this is the only modern study that uses global TRC. We directly compare our approach against the best-performing reported method, and our fine-tuned models substantially outperform this approach. For an additional comparison, we performed a standard training baseline with removing Step 1 and only applying LoRA while using the full set of available MATRES training examples. Our approach yields substantial gains of +17.7 F1 points on MATRES and +4 F1 points on OmniTemp, highlighting the critical role of our proposed domain adaptation and task-specific fine-tuning stages for achieving high performance in global TRC with LLMs.

**Pairwise Methods.** Table 3 presents the performance of our fine-tuned models compared to previous state-of-the-art (SOTA) methods that use a pairwise classification approach. Both our models—the fine-tuned Llama 3.1 8B and Gemma 3 12B—establish new SOTA results on both datasets. Our best-performing model achieves an improvement of +1.5 F1 points over the previous pairwise SOTA on MATRES and +2.2 F1 points on OmniTemp, demonstrating the effectiveness of leveraging LLMs with our fine-tuning strategy even when compared against expert pairwise models. We note that comparison with several recent studies (Roccabruna et al., 2024; Xu et al., 2025; Hu et al., 2025) is challenging due to different or unreported evaluation protocols used, which often differ in the way the VAGUE relation is treated during training and evaluation. These variations can substantially influence the final scores.

Interestingly, results on MATRES for the global method outperforms all pairwise methods, further emphasizing the performance of our solution.

### 5.2 Temporal Graph Consistency

We analyze the logical coherence of the temporal graphs produced by our models. We quantify the total number of transitive inconsistencies across all predictions generated for each test set. Table 4 summarizes these inconsistency counts. Consistent with its denser annotation structure, OmniTemp exhibits a higher absolute number of inconsistencies compared to MATRES across different methods.

However, a key finding is that our global TRC approach on our best performing model, trained on Gemma 3 12B, substantially reduces the occurrence of such inconsistencies, lowering the total count by more than 50% on both datasets compared

| Model | Method | MATRES | OmniTemp |
|---|---|---|---|
| LTM - Gemma-3-12B | pairwise | 3 | 61 |
| | global | 1 | 22 |
| LTM - Llama-3.1-8B | pairwise | 13 | 54 |
| | global | 2 | 54 |

Table 4: The total number of inconsistencies of our fine-tuned models across MATRES and OmniTemp test sets, evaluating performance in both pairwise and global approaches.

to results from the pairwise approach. Further analysis reveals that our global approach yields at least one inconsistency in only 3 documents, whereas the pairwise approach exhibits inconsistencies in 7 documents.

While the Llama 3.1 8B model also demonstrated the anticipated reduction in inconsistencies with the global approach on the MATRES test set, an initial anomaly was observed on the OmniTemp test set. Here, the total inconsistency counts were similar for both the pairwise and global methods. Further analysis revealed that a single outlier document accounted for an unusually high 35 inconsistencies in the global method. These were primarily due to the model struggling to classify two specific events within that document, leading to numerous incorrect relations. Excluding this document, the global method achieved 19 inconsistencies compared to 42 for the pairwise approach, aligning with the trend of improved graph consistency.

This shows that adopting a global perspective allows the model to generate more reliable and internally consistent temporal graphs.

## 6 Ablation Studies

### 6.1 Global Method

To assess the individual contributions of the components within our proposed two-step training methodology, particularly the temporal domain adaptation full fine-tuning (Step 1), we conducted a systematic ablation study on our best performing model (Gemma 3 12B). We assess how omitting or simplifying the Step 1 datasets before applying the standard Step 2 LoRA fine-tuning affects the final performance on the global TRC task.

First, we remove Step 1 entirely. The base instruction-tuned LLM undergoes only the downstream task LoRA fine-tuning in Step 2, completely omitting the Step 1 domain adaptation pre-training. We note that the omission of Step 1 allows us to use all 235 news articles from the MATRES training set during LoRA fine-tuning in Step 2.

| Configuration | F1-score |
|---|---|
| No Step 1 | 66.0 |
| MATRES only | 75.9 |
| MATRES + ATOMIC | 78.8 |
| MATRES + TGQA-TGR | 82.7 |
| MATRES + ATOMIC + TGQA-TGR | **83.7** |

Table 5: Ablation study for global approach on MATRES. Configurations are described in Section 6.1.

For ablation, we tested various combinations of the Step 1 datasets. We included the MATRES training set in all configurations for Step 1 to ensure the model was initially trained on core TRC concepts before incorporating additional datasets. Therefore, we perform Step 1 with the following configurations: (1) MATRES only, (2) MATRES + ATOMIC, and (3) MATRES + TGQA-TGR.

Table 5 presents the performance of all possible configurations on the MATRES test set for the global TRC task. The results yield several key insights, which are also supported in the analysis presented in Section 7.

**The Need of Domain Knowledge.** The *No Step 1* configuration exhibits substantially lower performance compared to all configurations incorporating Step 1 full fine-tuning. This finding strongly suggests that the initial domain adaptation phase is crucial for instilling foundational temporal concepts and reasoning abilities in the LLM, which are necessary precursors for effective task-specific fine-tuning.

**Benefit of Diverse Full Fine-Tuning Data.** We observe a clear trend of performance improvement as the diversity and complexity of the dataset used in Step 1 increase. This confirms the value of incorporating varied temporal datasets during full fine-tuning; the commonsense precedence knowledge from ATOMIC and the complex reasoning scenarios from TGQA-TGR demonstrably contribute to enhancing the model's capabilities beyond what is learned from the TRC examples in MATRES alone.

### 6.2 Pairwise Method

We further experiment with Step 1 using MATRES pairwise dataset only to verify the impact to the model's performance of adding additional datasets.

The results demonstrate comparable performance, with 85 F1 score in both configurations. This finding suggests that, for the task of pairwise TRC, the inclusion of the additional datasets did not yield an adverse impact on model performance.

## 7 Qualitative Analysis

**Dataset Configuration.** We conduct an analysis of the different dataset configurations presented in Section 6.1. Specifically, we examine their performance on individual labels — BEFORE, AFTER, and EQUAL — on the MATRES datasets using the global TRC approach.

A significant observation is the model's inability to output the EQUAL label. This is likely attributable to the sparse representation of EQUAL relations in the MATRES training set, where only 3% of pairs are labeled as such, so learning such classification is challenging. This is further explained by the frequent annotation errors within the EQUAL labels of the MATRES dataset, as noted by Niu et al. 2024.

Figure 2 illustrates the percentage of correct classifications for the BEFORE and AFTER labels. It is clear that the model not trained with Step 1 exhibits a substantial number of errors on both labels. While incorporating the MATRES dataset alone helps mitigate some of these errors, the improvement is not sufficient.

A key insight from this analysis is that adding TGQA-TGR, a complex temporal reasoning dataset, substantially helps the model to learn temporal concepts and boosts performance on the BEFORE label. Specifically, training with the TGQA-TGR dataset significantly improved performance by reducing the number of errors where BEFORE was incorrectly classified as AFTER by over 50% compared to the model trained solely on MATRES.

At the same time, this boost comes at the cost of a slight decrease in performance on the AFTER label. This behavior might stem from the construction of the TGQA-TGR training set, which often requires the model to place events within a timeline or asks the model to focus on the start time of the events. Such tasks could potentially emphasize prior events, leading to improved classification of the BEFORE relation.

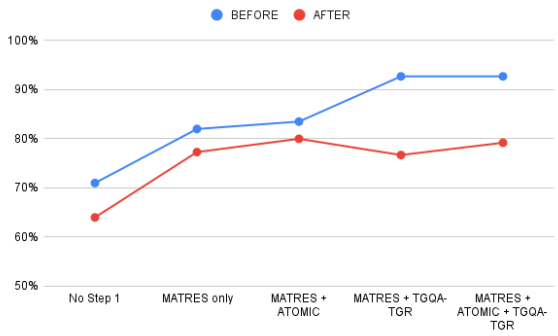


Figure 2: Comparing performance on BEFORE and AFTER labels on different models. Values are the percentage of correct classification.

Interestingly, the ATOMIC dataset only marginally improves performance, primarily on the AFTER label. This suggests that ATOMIC may not introduce many novel temporal concepts to the model, and its primary contribution might be to slightly balance the classification performance between the BEFORE and AFTER labels.

Combined, these datasets achieve what MATRES alone cannot: a notable improvement in temporal understanding and superior performance on the TRC task. This suggests that LLMs may inherently lack robust temporal understanding, and the MATRES dataset by itself is insufficient to instill the necessary temporal concepts for successful temporal relation classification.

**Transitive Inconsistencies.** We conducted a deeper analysis of the results presented in Section 5.2, specifically examining the performance of our Gemma-3-12B model on the OmniTemp dataset. This investigation yielded two key insights.

First, we found that over 90% of inconsistencies are attributable to errors on long-distance event pairs, defined as pairs with more than 50 tokens separating them. This observation strongly supports our assertion that global TRC is the preferred method for constructing temporal graphs involving long-distance events.

Second, our analysis revealed that inconsistencies often stem from a small number of specific events for which the model struggles to accurately classify related pairs. This is a significant finding, as it implies that the temporal graphs generated by our model, particularly in global settings, are largely consistent, with inconsistencies localized to a limited subset of nodes and their associated relations.

## 8 Conclusion

In this work, we proposed and evaluated a two-step fine-tuning methodology specifically designed to utilize the capabilities of LLMs for TRC.

Our models achieved new state-of-the-art results, substantially surpassing both previous best pairwise and global methods. Evidently, it offers a versatile approach applicable to both approaches. Furthermore, our analysis revealed that the global TRC models trained with our methodology produce substantially more logically consistent temporal graphs, mitigating a common issue in pairwise prediction schemes. Ablation studies and qualitative analyses confirmed the critical role of the initial domain adaptation stage and the benefit of diverse external knowledge data. The results strongly indicate that appropriately structured fine-tuning strategies can effectively unlock the latent temporal reasoning capabilities within large pre-trained models.

It is also important to note that the quality of the training data limits achievable performance, as MATRES is known to contain many annotation errors, as reported in (Niu et al., 2024). Future work should continue exploring the utility of LLMs in the temporal domain and leverage alternative, higher-quality datasets, such as OmniTemp.

This work not only advances the state-of-the-art in TRC but also provides a validated methodology for adapting LLMs to complex, knowledge-intensive NLP tasks. While the multi-step strategy presented in this paper was designed to address the unique data challenges inherent to the TRC task, future work can investigate how to apply a similar approach to address challenges in other domains.

## Limitations

There are three main limitations to this work. First, fine-tuning techniques often require large datasets for training. In our case, global TRC, we need many annotated documents since each is translated into a single training instance. Unfortunately, most of the datasets used for TRC are relatively small in terms of number of documents, leaving only MATRES as a potential training set for the two steps with its 275 documents. Other datasets, such as TBDense or OmniTemp, contain only 20-25 documents that can be used for training, making it challenging to train with them only. Second, we were limited in this study to medium size open source LLMs due to limited resources and the lack of fine tuning options offered by closed source models. Third, the diversity in evaluation protocols across different studies poses a significant challenge for comprehensive comparisons. Consequently, our comparative analysis is limited to studies employing the most common evaluation protocol.

Information on the the usage of AI assistants in this work is provided in Appendix D.

## References


Sarah Alsayyahi and Riza Batista-Navarro. 2023. TIMELINE: Exhaustive annotation of temporal relations supporting the automatic ordering of events in news articles. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16336–16348, Singapore. Association for Computational Linguistics.

Femke Bakker, Ruben Van Heusden, and Maarten Marx. 2024. Timeline extraction from decision letters using chatgpt. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 24–31.

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 (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Omer Cohen and Kfir Bar. 2023. Temporal relation classification using Boolean question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1843–1852, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Eirew, Kfir Bar, and Ido Dagan. 2025. Beyond pairwise: Global zero-shot temporal graph generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31428–31446, Suzhou, China. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational*

*Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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 *International Conference on Learning Representations*.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Jiafeng Guo, and Xueqi Cheng. 2025. Large language model-based event relation extraction with rationales. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7484–7496, Abu Dhabi, UAE. Association for Computational Linguistics.

Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, and Dongyan Zhao. 2023. More than classification: A unified framework for event temporal relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9631–9646, Toronto, Canada. Association for Computational Linguistics.

Jena Hwang, Chandra Bhagavatula, Ronan Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Choi Yejin. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:6384–6392.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. TIMERS: Document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–533, Online. Association for Computational Linguistics.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6203–6209, Hong Kong, China. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. 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)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Wanting Ning, Lishuang Li, Xueyang Qin, Yubo Feng, and Jingyao Tang. 2024. Temporal cognitive tree: A hierarchical modeling approach for event temporal relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 855–864, Miami, Florida, USA. Association for Computational Linguistics.

Jingcheng Niu, Saifei Liao, Victoria Ng, Simon De Montigny, and Gerald Penn. 2024. ConTempo: A unified temporally contrastive framework for temporal relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1521–1533, Bangkok, Thailand. Association for Computational Linguistics.

Gabriel Roccabruna, Massimo Rizzoli, and Giuseppe Riccardi. 2024. Will LLMs replace the encoder-only models in temporal relation classification? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20402–20415, Miami, Florida, USA. 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)*, pages 12053–12073, Torino, Italia. ELRA and ICCL.

Emre Sezgin, Syed-Amad Hussain, Steve Rust, and Yungui Huang. 2023. Extracting medical information from free-text and unstructured patient-generated health data using natural language processing methods: feasibility study with real-world data. *JMIR Formative Research*, 7:e43014.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2021. Extracting event temporal relations via hyperbolic geometry. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8065–8077, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2023. Event temporal relation extraction with Bayesian translational model. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1125–1138, Dubrovnik, Croatia. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev,

and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706, Online. Association for Computational Linguistics.

Zhonghua Wu, Wenzhong Yang, Meng Zhang, Fuyuan Wei, and Xinfang Liu. 2025. A reinforcement learning-based generative approach for event temporal relation extraction. *Entropy*, 27(3).

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)*, pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

Jun Xu, Mengshu Sun, Zhiqiang Zhang, and Jun Zhou. 2025. Maqinstruct: Instruction-based unified event relation extraction. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1441–1445, New York, NY, USA. Association for Computing Machinery.

Hao-Ren Yao, Luke Breitfeller, Aakanksha Naik, Chunxiao Zhou, and Carolyn Rosé. 2024. Distilling multi-scale knowledge for event temporal relation extraction.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Shuaicheng Zhang, Qiang Ning, and Lifu Huang. 2022. Extracting temporal event relation with syntax-guided graph transformer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 379–390, Seattle, United States. Association for Computational Linguistics.

Ling Zhuang, Hao Fei, and Po Hu. 2023. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100(C).

## A Hyperparameters Configuration

**Training Configuration.** Hyperparameter values were chosen based on the performance of the validation set. For Step 1, we used the full fine-tuning configuration option in Together AI, with a learning rate of 1e-5 and a batch size of 16 instances, and ran over all training instances during a single epoch. For Step 2, we used the LoRA configuration, with rank = 64 and alpha = 64, learning rate of 1e-4 and a batch size of 8 instances, and ran over all training instances during 3 epochs.

**Evaluation Configuration.** We ran inference using the trained models on the Together AI framework. We experimented with different temperature values, where we conducted 3 runs for each value and calculated the average score. We found out the temperature = 0 results in the best overall score and yields the most consistent and accurate results.

## B Training and Model Usage Cost

As discussed in Section 4.1, we used two open source models for applying our two-step approach and run our experiments - Gemma 3 12B and Llama 3.1 8B. We used the API of the Together AI framework for training and running the models. We now estimate the total cost of our experiments.

The fine-tuning price depends on the type (full fine-tuning or LoRA fine-tuning), the size of the model, and the total number of tokens used. In our experiments, a single run of Step 1 in our approach costs approximately $15-20, and Step 2 costs around $5.

The price of running the model depends on the duration in which the model endpoint is online and costs approximately $0.2 per minute. We only used endpoints when running experiments on the test sets.

Following the above analysis, we estimate that the total cost of our experiments is around $200.

## C Artifacts Information and Licenses

In this paper, we use the following common public artifacts - MATRES (Ning et al., 2018), provided without a license, OmniTemp (Eirew et al., 2025) uses summaries from the Multi-News corpus (Fabbri et al., 2019), which is distributed under a custom license that permits free academic use, ATOMIC-2020 (Hwang et al., 2021) provided under the CC-BY license, TGQA-TGR (Xiong et al., 2024) provided under the MIT license, Gemma 3 (Team et al., 2025) provided under the CC-BY 4.0 license and Llama 3.1 (Grattafiori et al., 2024) provided under Llama 3.1 Community License.

All datasets presented in this paper were used according to their original design and intended use. Their data is written in English only. We did not find any content in these datasets that required any further steps to protect or anonymize it. Specifically, data in MATRES or OmniTemp is public news articles, data in ATOMIC-2020 is common sentences that do not contain any offensive or problematic content, and data in TGQA-TGR was gen-

erated by LLM and does not contain any harmful content.

## D Use of AI Assistants

During the preparation of this paper, AI-powered language tools were utilized to assist with curating the text. This assistance was limited to improving phrasing, style, clarity, and grammatical correctness. The core research, conceptualization of ideas, experimental design, data analysis, and interpretation of results were conducted entirely by the human authors.

## E Training Instances

Below are examples of the training instances used during the two-step training presented in this paper.

Given the text below where events are marked with <eventName(identifier)>, for the specified pair of events below, determine the temporal relationships (BEFORE, AFTER, EQUAL, VAGUE) between them.

Text -

Major job cuts at AT and T. The long distance giant <slashing(ei255)> up to eighteen thousand jobs, freezing executive salaries and <shaking(ei257)> up management. The <changes(ei258)> are part of a one point six billion dollar cost cutting initiative to <revitalize(ei261)> its position in the telecommunications business. Earlier AT and T also <announced(ei262)> an eighteen percent <drop(ei263)> in profits for the fourth quarter. The company's sales force <applauded(ei265)> the shake up.

It's not something we're not <used(ei288)> to in the industry. But I <think(ei268)> right now, for AT and T and the people here, it's fairly <intense(ei269)> as you might w- well imagine.

I mean look at the stock. You know, since he's been here the stock <skyrocketed(ei274)> so, Yeah I <think(ei275)> he's <doing(ei276)> the right thing.

I <think(ei277)> it's a good thing that they're finally going to be <downsizing(ei279)> you know some management because there is a whole lot of waste.

Good news for AT and T today. I <think(ei282)> it's <excellent(ei283)> for the company.

But investors are <approaching(ei284)> the <changes(ei285)> with caution shares of AT and T down nearly four at sixty-one and a half.

Pair -

announced(ei262) – applauded(ei265)

Answer -

---

announced(ei262) BEFORE applauded(ei265)

Figure 3: An example of Step 1 MATRES training instance's input and output. Text after the dotted line is the expected model response.

Given the pairs of sentences below, for each pair determine if the first sentence happened BEFORE or AFTER the second sentence.

Sentences -
PersonX forms PersonY army – PersonX puts up recruitment posters
PersonX takes pictures of it – PersonX posts the photo online to help find the owner
PersonX takes it out of the oven – PersonX mixes a cake mix
PersonX asks PersonY's help – PersonX slips on the ice
PersonX watches sports – PersonX turned on the tv

Answer -

---

PersonX forms PersonY army AFTER PersonX puts up recruitment posters
PersonX takes pictures of it BEFORE PersonX posts the photo online to help find the owner
PersonX takes it out of the oven AFTER PersonX mixes a cake mix
PersonX asks PersonY's help AFTER PersonX slips on the ice
PersonX watches sports AFTER PersonX turned on the tv

Figure 4: An example of Step 1 ATOMIC training instance's input and output. Text after the dotted line is the expected model response.

You are given the following text. Answer the question below.

Text -

Once upon a time in Austin, Texas, a young boy named Liam Davis was born in 1972. From a very early age, Liam showed a natural talent for soccer. He joined the local youth team, the Phoenix Eagles, in 1988, where his skills quickly caught the attention of coaches and scouts. In 1989, Liam had the opportunity to play for the Seattle Warriors under-17 football team. His performance at this level was outstanding, and he was also selected to represent the Cheshire national under-16 football team. Liam's passion for the game grew stronger with each passing day. However, Liam's time with the Phoenix Eagles came to an end in 1990 as he wanted to explore new horizons. In 1991, he joined R.W. Eastbridge, a team that provided him with more challenging competition. Liam later briefly reunited with his former team, R.T. Wolverhampton, before finally bidding farewell to R.W. Eastbridge in 1992. In 1993, Liam embarked on a new journey by joining Oceanside United FC. This move proved to be pivotal in his career as he honed his skills and became a formidable player. A year later, in 1994, Liam's exceptional abilities drew the attention of Chelsea United FC, and he joined their ranks. He spent a year with the team, showcasing his talent and contributing to their success. However, Liam couldn't seem to settle down, and in 1995 he left Chelsea United FC. He took a break from competitive soccer for a few years to reassess his goals and ambitions. But Liam's love for the game never faded, and in 1998 he joined Dallas City FC. His return to the field was celebrated, and he played with unmatched passion until 1999 marked the end of his time with Dallas City FC. Although Liam Davis's soccer journey had its ups and downs, his perseverance and talent were undeniable. He left an indelible mark on every team he played for, inspiring his teammates and leaving a lasting impression on coaches and fans alike. Liam's story serves as a reminder that passion, determination, and love for the sport drive us to overcome obstacles and achieve greatness.

Question -

When did the event (Liam Davis was born in Austin) start?

Answer -

---

1972

Figure 5: An example of Step 1 TGQA-TGR training instance's input and output. Text after the dotted line is the expected model response.

Given the text below where events are marked with <eventName(identifier)>, for each pair of events below, determine the temporal relationships (BEFORE, AFTER, EQUAL, VAGUE) between them.

Text -

Philip Morris Cos., New York, <adopted(ei55)> a defense measure <designed(ei56)> to <make(ei57)> a hostile <takeover(ei58)> prohibitively expensive.

The giant foods, tobacco and brewing company <said(ei59)> it will <issue(ei60)> common-share purchase rights to shareholders of record Nov. 8. Under certain circumstances, the rights would <entitle(ei62)> Philip Morris holders to <buy(ei63)> shares of either the company or its acquirer for half price.board isn't <aware(ei64)> of any <attempts(ei65)> to <take(ei66)> over Philip Morris, the company <said(ei67)>. As of Sept. 30, Philip Morris <had(ei68)> 926 million shares outstanding. In composite trading on the New York Stock Exchange, Philip Morris shares <closed(ei69)> yesterday at $43.50 each, down $1.

Pairs -

adopted(ei55) – designed(ei56)
adopted(ei55) – said(ei59)
designed(ei56) – said(ei59)
said(ei67) – had(ei68)
had(ei68) – closed(ei69)

Answer -

---

adopted(ei55) AFTER designed(ei56)
adopted(ei55) BEFORE said(ei59)
designed(ei56) BEFORE said(ei59)
said(ei67) AFTER had(ei68)
had(ei68) BEFORE closed(ei69)

Figure 6: An example of Step 2 training instance's input and output. Text after the dotted line is the expected model response.