ChronoFact: Timeline-based Temporal Fact Verification

Anab Maulana Barik[1] , **Wynne Hsu**[1,2] and **Mong Li Lee**[1,3]
[1]School of Computing, National University of Singapore, Singapore
[2]Institute of Data Science, National University of Singapore, Singapore
[3]Centre for Trusted Internet and Community, National University of Singapore, Singapore
anabmaulana@u.nus.edu, {whsu,leeml}@comp.nus.edu.sg

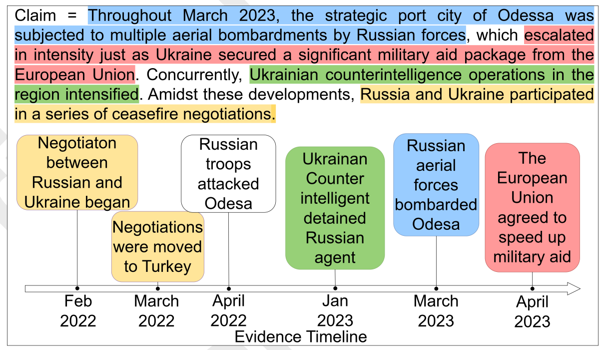
Abstract

Temporal claims, often riddled with inaccuracies, are a significant challenge in the digital misinformation landscape. Fact-checking systems that can accurately verify such claims are crucial for combating misinformation. Current systems struggle with the complexities of evaluating the accuracy of these claims, especially when they include multiple, overlapping, or recurring events. We introduce a novel timeline-based fact verification framework that identify events from both claim and evidence and organize them into their respective chronological timelines. The framework systematically examines the relationships between the events in both claim and evidence to predict the veracity of each claim event and their chronological accuracy. This allows us to accurately determine the overall veracity of the claim. We also introduce a new dataset of complex temporal claims involving timeline-based reasoning for the training and evaluation of our proposed framework. Experimental results demonstrate the effectiveness of our approach in handling the intricacies of temporal claim verification.

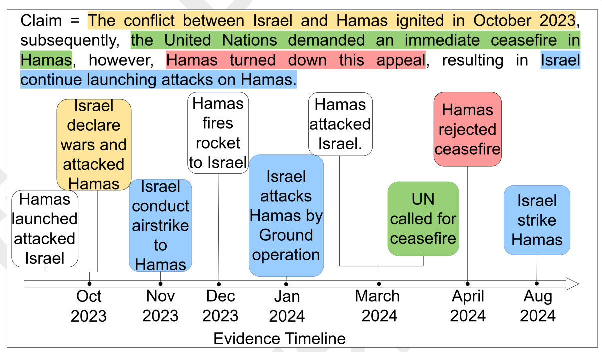
1 Introduction

The spread of false information has reached alarming levels, undermining social trust significantly. One prominent category of false information involves temporal claims, which are statements that include time-specific elements, either explicitly (e.g., "in 1953") or implicitly (e.g., "before [another event]"). The complexity of verifying these claims escalates with the number of events mentioned. Effective verification must assess not only the veracity of each event within its temporal context but also understand the relationships between these events, especially their chronological order.

Existing research has largely focused on verifying the veracity of individual events within claims, and overlook the chronological order of these events [Barik *et al.*, 2024; Qudus *et al.*, 2023]. This oversight can undermine the effectiveness of fact-checking systems, particularly when dealing with complex narratives where the sequence of events is crucial for determining the truth. Understanding the chronological order of events is hindered by the absence of the explicit



(a) Claim with multiple events



(b) Claim with recurring event

Figure 1: Illustration of complex temporal claim verification.

temporal cues. The complexity increases when events overlap or recur, complicating the task of establishing a clear and accurate timeline for the verification process.

Example 1. Figure 1(a) shows a claim involving four events, with evidence relevant to each event indicated by matching color boxes. While each event appears to be supported by some evidence when analyzed independently, closer scrutiny of the timeline reveals discrepancies. For instance, evidence indicates that negotiations between Ukraine and Russia took place in Feb 2022, before the bombardments in March 2023. Therefore, the claim event *"Russia and Ukraine participated in a series of ceasefire negotiations"*, suggested by the temporal cues to have occurred in March 2023, is not actually supported by the evidence. Existing works that analyze individ-

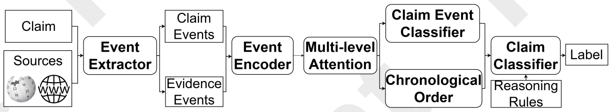


Figure 2: Overview of `ChronoFact` Framework.

ual events without considering the timeline would incorrectly conclude that all the events in the claim are supported, and deem the claim true. This shows the importance of **chronological order of events** in claim verification.

Example 2. Figure 1(b) shows a claim event *"Israel attacks Hamas"*, with evidence indicating it occurs at three different times: Nov 2023, Jan 2024, and Aug 2024. Analysis of the timeline shows that the occurrence in Aug 2024 aligns with the order of events in the claim, thereby concluding that the claim is supported. Previous studies that have overlooked the timeline of events might match the claim event to earlier occurrence in Nov 2023, or Jan 2024, leading to incorrect conclusion that this claim is refuted. This highlights the need to **align claim and evidence events with their chronological order** for accurate claim verification.

To overcome the limitations in existing temporal fact verification methods, we introduce `ChronoFact`, a framework that systematically identify events from both claim and evidence to construct a coherent timeline for temporal fact verification. `ChronoFact` examines the relationships between claim and evidence events at three levels: event-level, token-level, and time-level and learns a model to predict the veracity of each claim event and their chronological accuracy.

We develop a new dataset called `ChronoClaims` for timeline-based fact verification. This dataset encompasses complex claims involving multiple, related events that unfold over time with both implicit and explicit temporal expressions. These features are lacking in current datasets like T-FEVER, which mainly consist of single-event claims, and T-FEVEROUS, which, despite its more complex claims, often has events that are not chronologically related.

2 Related Work

There has been a stream of research on evidence-based claim verification to assess whether the evidence sentences support or refute the claim [Stammbach and Neumann, 2019; Soleimani *et al.*, 2020]. Techniques like GEAR [Zhou *et al.*, 2019], KGAT [Liu *et al.*, 2020], and DREAM [Zhong *et al.*, 2020] model claim and evidence into a graph using Graph Attention Network to facilitate information propagation of information between them. CGAT [Barik *et al.*, 2022] enhances this by incorporating commonsense knowledge from ConceptNet to enrich the contextual representation. These works do not consider temporal information.

Only a few studies have incorporated temporal information in the claim verification process. NMT [Mori *et al.*, 2022] verifies economic claims against time series data in tabular format by translating the claims into Datalog rules to query the tabular evidence. TemporalFC [Qudus *et al.*, 2023] focuses on verifying claims represented as tuples against a knowledge graph. It utilizes temporal graph embeddings to determine the validity timing of underlying triples. However, these works are limited to structured data, hindering its applicability to natural language claims and evidence.

[Allein *et al.*, 2021] capture the temporal relevance of evidence through a re-ranking process based on the proximity of publication dates between the evidence and the claim. ITR [Allein *et al.*, 2023] extends this by assigning publication dates of claims and evidence to fixed-size time buckets for temporal reasoning. However, it ignores implicit temporal cues or chronological order in the claim and evidence.

TACV [Barik *et al.*, 2024] decomposes claim and evidence sentences into events and employ temporal-aware representation encoder to retrieve evidence that are both semantically and temporally related to the claim. It utilizes GPT for temporal reasoning to verify individual claim events. Unlike TACV, our approach also evaluates the chronological timeline accuracy between the claim and evidence, providing a more thorough verification of complex claims.

A related area of research, question answering (QA), has also begun to incorporate temporal information. FAITH [Jia *et al.*, 2024] focuses on implicit temporal reasoning for the temporal question answering task by generating intermediate questions with explicit temporal information. This method uses heterogeneous sources to improve the completeness of evidence retrieval and employs a general-purpose QA system to respond to the question. While it is common to utilize QA in the claim verification process [Pan *et al.*, 2023], they do not consider the chronological order of events. In contrast, our work focuses on verifying complex claims that require timeline-based reasoning.

3 Proposed Framework

`ChronoFact` follows the typical automated claim verification which involves collecting relevant evidence from credible sources and assessing the claim's veracity based on the evidence. Figure 2 shows the key modules in the framework.

Event Extractor. Given a claim, we first employ the GENRE sequence-to-sequence entity linking model [De Cao *et al.*, 2020] to retrieve relevant documents from Wikipedia and extract all evidence sentences. Then we utilize Semantic Role Labelling from AllenNLP [Allen, 1983] to extract events from both the claim and evidence. Each event has its core information and temporal argument. Finally, we score the evidence with the events extracted from the claim using an event representation encoder similar to [Barik *et al.*, 2024].

Event Encoder. We tokenize each event and pass each token i to the flan-T5 model to obtain the corresponding token representations H_i. For tokens that represent date, we apply mean pooling, followed by positional encoding [Vaswani *et al.*, 2017] where the position corresponds to the distance between the event date and the earliest date found in the claim and evidence events. The final representation of the event is given by $<H_{CLS}, H_1 \cdots H_d>$ where d is the number of tokens in the event, H_{CLS} is the event-level representation obtained by the average pooling of $H_j, 1 \leq j \leq d$, and $H_1 ... H_d$ are the token-level representations.

Multi-level Attention Encoder. We process each pair of claim event representation c_i and evidence event representation e_j through a multi-level attention module to determine the relevance of evidence events for each claim event. This involves calculating attention scores at three levels:

- **token-level attention score** α_{ij}: is the average of cosine similarities between all pairs of tokens in c_i and e_j.
- **event-level attention score** β_{ij}: is the cosine similarity between the H_{CLS} representations of c_i and e_j.
- **time-level attention score** γ_{ij}: is the cosine similarity between the mean pooled date representations in c_i and e_j.

The final multi-level attention score between c_i and e_j, denoted as ω_{ij}, is the average of the event-level, token-level, and time-level attention scores. These computed attention scores are then employed to predict the label of each event in the claim, assess the accuracy of the claim's chronological order based on the evidence timeline, and evaluate the claim's overall veracity. Note that we considered dynamically learn the weights of each attention level. However, our experiment indicated that this approach did not improve performance, leading us to adopt the average weighting method instead.

Claim Event Classifier. This module predicts the label of claim event c_i using the top-k evidence events $E' \subset E$ with the highest final attention scores. Let $H^{c_i}_{CLS}$ and $H^{e_j}_{CLS}$ be the representation of the CLS token for c_i and evidence event $e_j \in E'$ respectively. We concatenate the representations weighted by the attention scores to obtain u^{c_i} as follows:

$$u^{c_i} = H^{c_i}_{CLS} \oplus \omega_{i1} H^{e_1}_{CLS} \oplus \cdots \oplus \omega_{ik} H^{e_k}_{CLS}$$

This is then fed to two fully connected layers followed by softmax to obtain the probability distribution z^{c_i}:

$$z^{c_i} = \mathrm{softmax}(FC_2(\mathrm{ReLU}(FC_1(u^{c_i}))))$$

where $z^{c_i}[0]$, $z^{c_i}[1]$, and $z^{c_i}[2]$ are the probabilities of the labels "SUP", "REF", and "NEI" respectively. The label y^{c_i} for the claim event c_i is assigned based on the highest probability among these.

Chronological Order Classifier. This module predicts the accuracy of the chronological order of a claim C using the timeline of evidence events. Given n events in C, the relevance of each evidence event to C is given by:

$$r^{e_j} = \tanh\left(\sum_{i=1}^{n} \omega_{i,j}\right) \quad (1)$$

We use $\tanh$ for its bounded output range of $[-1, 1]$, which aligns with our interpretation of relevance as a continuous spectrum ranging from negative to positive. We sort the evidence events based on their relevance to the claim and obtain the top-k events with the highest score r^{e_j}. These top-k evidence events, along with the claim events, are then input into GPT to be reordered according to their chronological sequence. Given the limited number k of events, this reordering process requires minimal time and resources, making it practical for real-world applications.

The reordered sequence of claim events, denoted as $seqC = H^{c_1}_{CLS} \oplus \cdots \oplus H^{c_n}_{CLS}$ is passed to a Bi-LSTM to capture and embed the chronological order into the model. Similarly, the reordered sequence of evidence events, weighted by their relevance scores is given by

$$seqE = r^{e_1} H^{e_1}_{CLS} \oplus \cdots \oplus r^{e_m} H^{e_m}_{CLS}$$

This sequence is passed to a second Bi-LSTM. The outputs from the two Bi-LSTM, denoted as o^C for claim events and o^E for evidence events, are then fed into two fully connected layers followed by softmax to obtain the distribution z^o:

$$z^o = \mathrm{softmax}(FC_4(\mathrm{ReLU}(FC_3([o^C \oplus o^E]))))$$

where $z^o[0]$ is the probability that the chronological order of the claim events is supported by that of the evidence events, and $z^o[1]$ is the probability that the chronological order refutes the claim's timeline. The output of the chronological order classifier y^o is the label with the highest probability.

Claim Classifier. To predict the overall veracity of the claim, we concatenate $seqC$ and $seqE$, along with the distributions $z^{c_1} \cdots z^{c_n}, z^o$, and pass this vector to two fully connected layers followed by a softmax function to obtain the probability distribution z that the claim is SUP, REF or NEI. The label with the highest probability is depicted as y.

3.1 Model Training

We train the model using two losses $\mathcal{L}_{cross}$ and $\mathcal{L}_{soft}$. The first loss $\mathcal{L}_{cross}$ is defined as follows:

$$\mathcal{L}_{cross} = \sum_{i=1}^{n} F(g^{c_i}, z^{c_i}) + F(g^o, z^o) + F(g, z) \quad (2)$$

where $F(.)$ is a cross-entropy function, g^{c_i}, g^o, and g are the ground-truth labels for claim event c_i, chronological order, and overall claim label respectively.

The second loss ensures the consistency between the overall claim label, claim event labels, and chronological accuracy. In particular, we apply a set of logic rules based on the outcomes of the individual claim events and their chronological alignment with the evidence. Specifically, a claim is deemed supported if all its associated claim events are supported and their chronological sequence matches that of the evidence events. This is expressed using first-order logic:

$$y^{c_1} \wedge \cdots \wedge y^{c_n} \wedge y^o \implies y$$

This logical expression states that the overall veracity y of a claim is SUP if and only if each claim event y^{c_i} is supported and the chronological order y^o is consistent with the evidence. On the other hand, if any one of the claim event is refuted, or the chronological order does not align, then y is REF. Otherwise, y is NEI. We leverage Gödel t-norm to soften the hard reasoning rules in the claim classifier, and obtain the differentiable distribution z_{soft}:

$$\begin{aligned} z_{soft}[0] &= min(z^{c_1}[0], \cdots, z^{c_n}[0], z^o[0]) \\ z_{soft}[1] &= max(z^{c_1}[1], \cdots, z^{c_n}[1], z^o[1]) \\ z_{soft}[2] &= 1 - z_{soft}[0] - z_{soft}[1] \end{aligned} \quad (3)$$

With this, we define $\mathcal{L}_{soft}$ which ensures the consistency of the overall claim label with the claim events and their chronological order as follows:

$$\mathcal{L}_{soft} = D_{KL}(z || z_{soft}) \quad (4)$$

where D_{KL} is Kullback-Leibler divergence, which measures the difference between the predicted overall claim label distribution and the distribution derived from the soft logic.

			Train Set		Validation Set		Test Set	
			Support	Refute	Support	Refute	Support	Refute
Temporal Expression	Explicit	3 events	3,561	2,954	355	318	397	315
		4 events	3,574	3,376	354	282	359	262
		5 events	3,597	3,084	266	198	323	218
	Implicit	3 events	3,804	3,028	355	318	390	313
		4 events	3,453	3,306	355	278	361	263
		5 events	3,473	3,039	266	199	321	213
Temporal Category	Overlapping events		10,360	9,297	918	807	985	714
	Recurring events		8,914	6,507	729	376	865	361

Table 1: Characteristics of ChronoClaims dataset.

The final loss function combines the soft logic loss $\mathcal{L}_{soft}$ with standard cross-entropy loss $\mathcal{L}_{cross}$, enabling the model to be trained both with supervision and structured constraints, with mu as the hyperparameter:

$$\mathcal{L} = (1-\mu)\mathcal{L}_{cross} + \mu\mathcal{L}_{soft} \tag{5}$$

4 ChronoClaims Dataset

We introduce a new benchmark dataset called `ChronoClaims` that is designed for enhancing the accuracy and complexity of timeline-based fact verification. Utilizing the November 2022 Wikidata snapshot [Vrandečić and Krötzsch, 2014], we preprocess and extract facts in the format <subject, relation, object, time_start, time_end>. To construct an evidence timeline, we organize all the facts that have the same subject in a chronological order. Then we randomly select N facts from this timeline and use GPT to transform these facts into coherent sentences.

For generating sentences with implicit temporal information, we omit the time_start and time_end from the fact, and prompt GPT to craft sentences that subtly embed the temporal context. Then we use a cloze-style template[1] to synthesize the sentences into a claim that preserves the chronological order. To generate claims that contradict the evidence timeline, we rearrange the order of these sentences. Finally, we use GPT to refine and rephrase the synthetic claim to make it more natural and fluent. Each claim is labeled as either SUP or REF, depending on whether it aligns or conflicts with the timeline.

We evaluate the quality of the `ChronoClaims` dataset by sampling 500 claims, ensuring equal distribution across different temporal expressions and event complexities (16% each type), and temporal categories (50% each category) to achieve representative coverage. Two human annotators are tasked with determining the labels of the generated claims based on the ground truth evidence timeline. The agreement rates are high, with 96.8% and 97% of the labels assigned by the annotators matching the labels of the generated claims.

We analyze the claims where the generated labels are different from the annotators, and discover that most of the discrepancies are due to errors in rephrasing For example, a claim "Nasrallah Peter Sfeir worked as a Catholic priest, and then Nasrallah Peter Sfeir was educated at Saint Joseph University," was rephrased to "Nasrallah Peter Sfeir pursued his education at Saint Joseph University before becoming a Catholic priest," which altered the original chronological sequence. To rectify this, we perform a second verification step using GPT to ensure that the semantic meanings of the original and rephrased claims remain consistent. Any claims where the meaning has been altered are discarded.

In total, we generated 40,249, 3,544 and 3,735 claims for the training, validation and test sets. Table 1 shows the detailed statistics of the `ChronoClaims` dataset.

[1] <sentence$_1$> and then <sentence$_2$> and then <sentence$_3$>

5 Performance Study

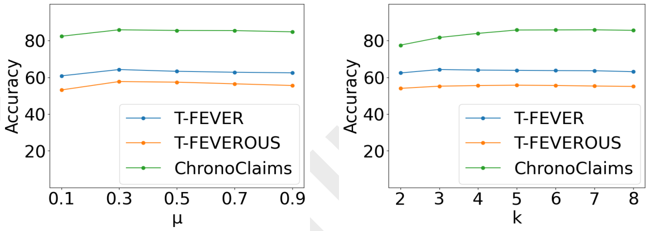
Datasets. Besides the `ChronoClaims` dataset, we also use the T-FEVER and T-FEVEROUS datasets [Barik *et al.*, 2024]. These datasets are derived from the benchmark fact verification datasets FEVER [Thorne *et al.*, 2018a], FEVER2.0 [Thorne *et al.*, 2018b] and FEVEROUS [Aly *et al.*, 2021] respectively such that the claims in T-FEVER and T-FEVEROUS contain temporal expressions. Each claim involves a maximum of 3 events and is labeled as SUP, REF and NEI. We also evaluate our method on the T-QuanTemp, a subset of the QuanTemp [Venktesh *et al.*, 2024] dataset focusing on real-world claims with temporal aspects. T-QuanTemp comprises of temporal claims containing temporal expressions or tagged as as temporal aspects by QuanTemp dataset. Table 2 shows the characteristics of T-FEVER, T-FEVEROUS and T-QuanTemp datasets. We use 80% of the data for training and 20% for testing. While T-QuanTemp comes with its own set of evidence, we rely on different knowledge sources for the others: Wikipedia for T-FEVER and T-FEVEROUS, and Wikidata for ChronoClaims.

Dataset	Support	Refute	NEI
T-FEVER	11,799	9,292	3,975
T-FEVEROUS	33,357	28,959	1,266
T-QuanTemp	1,261	3,470	1,349

Table 2: Dataset characteristics.

Implementation Details. We implement the `ChronoFact` framework using Hugging Face Transformers Library with PyTorch. The Event Encoder use flan-T5 base [Chung *et al.*, 2024] with a hidden size of 768. In the Multi-level Attention Encoder, the token-level, event-level, and time-level representations pass through a linear layer of dimension 768 to calculate attention scores. The hidden size of the fully connected layers is set to 192. The Chronological Order Classifier uses two layers of Bi-LSTM, each with a hidden size of 768, and the fully connected layers have a hidden size of 192, matching those in the Claim Classifier.

We train the model using Adafactor with a batch size of 8 and a learning rate of 5e-5 for 5 epochs on each dataset.



(a) Effect on μ (b) Effect on k

Figure 3: Sensitivity Experiments on `ChronoFact`.

ChronoFact is trained to predict REF and SUP labels on ChronoClaims, as the dataset does not have NEI labels. For T-FEVER and T-FEVEROUS, the model is trained to predict SUP, REF, and NEI. We report the macro F1 score of the best performing model on the test set.

5.1 Sensitivity Experiments

We first conduct sensitivity experiments to obtain the optimal value of the parameter μ in Eqn 5. Figure 3(a) shows the label accuracy as we vary the value of μ from 0.1 to 0.9. The performance improves when μ increases from 0.1 to 0.3, indicating that the model benefits from incorporating $\mathcal{L}_{soft}$. The best performance is achieved when $\mu = 0.3$ across all datasets and we use this value for the rest of the experiments.

We also vary the number of top-k evidence events in the classifier module to examine its effect on `ChronoFact`'s performance. Figure 3(b) shows that the optimal performance for T-FEVER, T-FEVEROUS, and `ChronoClaims` was achieved when k is 3, 5, and 7 respectively, and we use these values in our experiments. Note that ChronoClaims generally achieves higher performance because it has only two class labels (SUPPORTS, REFUTES), whereas T-FEVER and T-FEVEROUS include an additional label ('NOT ENOUGH INFO').

5.2 Comparative Experiments

We compare `ChronoFact` with the following state-of-the-art evidence-based fact verification baselines:

- KGAT [Liu *et al.*, 2020]. This uses a transformer to obtain claim-sentence representations and a graph attention network to aggregate the evidence for claim verification.
- CGAT [Barik *et al.*, 2022]. This method incorporates external knowledge from ConceptNet to enrich the contextual representations of claim and evidence sentences. It then employs graph attention networks to propagate the information among the evidence sentences to verify the claim veracity.
- ITR [Allein *et al.*, 2023]. This also employs a transformer to obtain the claim and evidence representations which are augmented with the publication dates of the claim and evidence for temporal reasoning.
- FAITH [Jia *et al.*, 2024]. We adapt this temporal QA model for temporal claim verification by using GPT to generate relevant temporal questions for FAITH, then prompting GPT to verify the claim based on FAITH's generated answers.
- GPT-4o in a zero-shot setting. Given the claim and retrieved evidence using TACV retrieval method, we prompt GPT-4o to predict the claim's label.
- TACV [Barik *et al.*, 2024]. This is an end-to-end solution for temporal claim verification that considers the temporal information in claims to obtain relevant evidence sentences and uses LLM for temporal reasoning.

Table 3 shows the macro F1 score of the various methods on the ChronoClaims dataset, with our proposed `ChronoFact` achieving the highest score and significantly outperforming the baseline models[2]. Further analysis in terms of the number of events shows that `ChronoFact` is better in managing complex claims with multiple events. This is evident in the smaller decline in the performance as the number of events increases, compared to other baselines. We also analyze `ChronoFact`'s ability to handle complex temporal event types where the timelines of different events may overlap or the events may recur at multiple time points. Once again, there is a significant gap in the macro F1 score compared to other methods, demonstrating `ChronoFact`'s superior performance in managing such complexities.

Table 4 shows the performance on T-FEVER and T-FEVEROUS datasets, where `ChronoFact` is the best performer on both datasets. Similar trend is observed when comparing the performance in terms of the number of events. Table 4 also shows `ChronoFact`'s robustness on the real-world T-QuanTemp, despite noisy date information.

Error Analysis. We conduct an error analysis on 50 randomly incorrectly predicted claims on ChronoClaim dataset. We find that 82% of these errors are due to claims containing implicit temporal information. For example, the claim "Greg Clark first held the position of Minister of State for Decentralisation, thereafter became the Secretary of State for Business and Trade, and later served as the Secretary of State for Levelling Up, Housing and Communities." lacks explicit dates or clear temporal markers. This ambiguity poses a challenge for the model to accurately encode and reason about the sequence of events. The remaining 18% of the errors comes from claims with events that have multiple dates and varying levels of granularity, such as "Nicola Fratoianni joined the Movement for the Left from 2009 until October 22, 2010". The complexity of this claims hinders the model's ability to encode the temporal information.

5.3 Ablation Studies

We examine the effectiveness of the key modules in `ChronoFact` by implementing the following variants:

- `ChronoFact` without multi-level attention encoder module. In this variant, we set the final attention scores between each claim event c_i and evidence event e_j to 1.
- `ChronoFact` without claim event classifier. Here, the input to the claim classifier is the concatenation of $seqC$, $seqE$, and the distribution of chronological order z^o.
- `ChronoFact` without chronological order classifier. This variant does not assess the consistency of the chronological order of the claim events with that of the evidence events. Therefore, the input to the Claim Classifier is the concatenation of $seqC$, $seqE$, and the probability distribution of the claim events $z^{c_1} \cdots z^{c_n}$.

[2]Results of micro F1 scores are provided in https://arxiv.org/pdf/2410.14964

Method	**Overall**	**Number of Events**			**Event Types**	
		3 events	4 events	5 events	Overlapping	Recurring
KGAT	50.56±3.52	52.61±3.20	49.33±4.06	47.17±5.57	50.14±3.88	47.43±1.68
CGAT	56.69±0.36	58.11±1.20	55.98±0.57	54.75±0.58	56.56±0.22	56.95±0.97
ITR	58.34±6.44	63.86±3.49	61.52±0.27	51.15±2.41	57.87±5.25	55.50±6.31
FAITH	60.84±0.11	61.64±0.89	61.48±0.30	59.03±1.31	61.03±0.21	52.42±0.16
GPT4o	61.93±0.30	70.06±0.33	57.65±0.54	55.51±0.48	61.86±0.22	55.08±0.18
TACV	63.42±0.60	67.45±0.87	63.02±0.78	61.81±0.56	62.37±0.69	65.84±5.04
`ChronoFact`	**85.49±0.53**	**86.33±0.46**	**85.55±0.31**	**84.86±0.43**	**84.86±0.76**	**80.89±0.55**

Table 3: Comparison of macro F1 on ChronoClaims.

Method	**T-FEVER**			**T-FEVEROUS**				**T-QuanTemp**
	Overall	**Number of Events**		**Overall**	**Number of Events**			**Overall**
		1 event	2 events		1 event	2 event	3 event	
KGAT	40.66±1.04	40.82±1.21	37.85±2.40	17.66±1.21	21.95±3.03	17.85±1.39	16.40±0.82	39.82±0.46
CGAT	42.31±2.47	42.47±2.61	39.20±0.85	19.62±1.92	23.78±1.66	19.30±1.85	18.54±1.95	42.01±0.25
ITR	45.21±3.86	45.62±3.89	36.77±2.36	27.62±2.99	30.19±4.50	27.64±3.39	26.99±2.55	44.90±1.27
FAITH	49.03±0.79	49.42±0.57	42.65±4.06	38.63±0.02	41.89±0.15	40.17±0.27	37.75±0.15	49.65±0.24
GPT4o	53.77±0.23	53.54±0.30	57.19±1.29	43.54±0.19	44.28±0.48	44.11±0.22	43.00±0.27	53.12±0.21
TACV	49.86±0.40	50.25±0.41	43.64±0.01	39.84±0.60	43.72±0.84	39.78±0.93	37.72±0.82	49.69±0.47
`ChronoFact`	**56.29±1.50**	**56.14±1.41**	**57.34±2.39**	**47.78±0.98**	**48.27±1.62**	**48.23±2.41**	**47.88±2.21**	**65.67±0.20**

Table 4: Comparison of macro F1 on T-FEVER, T-FEVEROUS, and T-QuanTemp.

Variants	**Datasets**			
	ChronoClaims	T-FEVER	T-FEVEROUS	T-QuanTemp
w/o multi-level attention encoder	84.65±0.30	53.60±0.91	42.01±0.55	63.53±1.51
w/o claim event classifier	83.72±0.57	54.55±1.10	41.24±2.28	63.95±0.98
w/o chronological order classifier	76.88±2.14	55.31±0.32	40.75±1.14	64.35±0.43
`ChronoFact`	**85.49±0.53**	**56.29±1.50**	**47.78±0.98**	**65.67±0.20**

Table 5: Macro F1 score of ablation studies.

Table 5 shows that the largest performance drop occurs when when we exclude the chronological order classifier, emphasizing that predicting chronological order enhances the model prediction. This effect is particularly evident in the ChronoClaims dataset, which is specifically designed to test the model's ability to reason over the chronological order of events. Its effect is less noticeable on the T-FEVER dataset as it mainly consists of single-event claims. The next largest drop in macro F1 score is when we exclude the claim event classifier, highlighting the importance of predicting individual claim events for the overall claim prediction.

Excluding the multi-level attention encoder leads to a decrease in macro F1 score across all datasets, indicating the role of considering the relevance of evidence events in claim verification. We conduct a manual analysis of the attention scores associated with the relevant evidence events for 50 claim events. For each claim event, we examine whether the evidence event with the highest event-level and token-level attention scores is semantically related, and whether the evidence event with the highest time-level attention score is temporally related. Our findings indicate that, in every case, the evidence event with the highest score is indeed semantically or temporally relevant to the corresponding claim event.

6 Case Studies

Finally, we present case studies to illustrate the importance of considering the chronological order in verifying temporal claims. Table 6 shows a T-FEVEROUS claim with three events involving Yossi Yona: "studying for a PhD", "become a Professor", and "join the left camp". The temporal context provided by the word "before" is lost between the events "became a Professor" and "joined the Left Camp", resulting in each individual claim event being supported by the retrieved evidence. However, the chronological order of the claim events is incorrect because evidence shows that "joined the left camp" *before* "became a Professor". `ChronoFact` addresses this issue by inferring the missing temporal relationships from the timeline of events, and correctly predicts the label as REF. In contrast, TACV evaluates the claim events independently and mistakenly predicts the claim as SUP.

Table 7 shows a claim from ChronoClaim dataset. The claim has five events, where the first event "Davie was a member of Dundee United F.C" overlaps with the second event "Davie joined Arbroath F.C.". TACV does not accommodate the possibility of overlapping events and assumes that "join" implies a full transfer or departure, and conclude that the evidence refutes the claim event c2. In contrast, `ChronoFact` can handle overlapping events and correctly conclude that c2 is supported by the evidence.

Table 8 shows another claim which involves a recurring event where Andonov was a member of Botev Polediv from 2002 to 2006, left, and rejoined in 2009. TACV fails to recognize this recurrence, mistakenly using evidence of Andonov's initial membership period (2002-2006) to refute the claim of his 2009 return. In contrast, `ChronoFact` organizes the evidence chronologically, recognizing Georgi Andonov's return to Botev Plovdiv in 2009 after a hiatus, yielding a correct prediction.

Claim: Yossi Yona began studying for a PhD and went on to become a Professor of philosophy of education at Ben - Gurion University before he joined the Left Camp of Israel party . Label: REF				
Claim Events	c1: Yossi Yona began studying for a PhD c2: Yossi Yona become a Professor of philosophy of education at Ben - Gurion University c3: Yossi Yona joined the Left Camp of Israel party	Claim Event Label	Chrono. Order	Final Claim
TACV	**Evidence Events:**			
	• After graduating in 1979, he began studying for a PhD, graduating from the University of Pennsylvania in Philadelphia	SUP c1	N.A.	SUP
	• Yossi Yona became a Professor of philosophy of education at Ben-Gurion University	SUP c2		
	• He subsequently joined the Education Department at Ben-Gurion University of the Negev.			
	• Whilst at university he joined the Left Camp of Israel party.	SUP c3		
`ChronoFact`	**Evidence Events in Chronological Order:**			
	1. Whilst at university, Yossi Yona joined the Left Camp of Israel party	SUP c3	REF	REF
	2. After graduating in 1979, Yossi Yona began studying for a PhD, graduating from the University of Pennsylvania in Philadelphia	SUP c1		
	3. He subsequently joined the Education Department at Ben-Gurion University of the Negev.			
	4. Yossi Yona became a Professor of philosophy of education at Ben-Gurion University	SUP c2		

Table 6: Sample T-FEVEROUS claim where chronological order of claim events is inconsistent with that of evidence events.

Claim: Davie Dodds was a member of Dundee United F.C. before joining Arbroath F.C., thereafter becoming part of the Scotland national football team, then moving to Neuchâtel Xamax, and finally playing for Aberdeen F.C. Label: SUP				
Claim Events	c1: Davie Dodds was a member of Dundee United F.C. c2: Davie Dodds joined Arbroath F.C. after joining Dundee United F.C. c3: Davie Dodds became part of the Scotland national football team c4: Davie Dodds moved to Neuchâtel Xamax c5: Davie Dodds played for Aberdeen F.C.	Claim Event Label	Chrono. Order	Final Claim
TACV	**Evidence Events:**			
	• Davie Dodds is a member of the Dundee United F.C. from 1975 until 1986	SUP c1	N.A.	REF
	• Davie Dodds is a member of the Arbroath F.C. from 1977 until 1978	REF c2		
	• Davie Dodds is a member of the Scotland national football from 1983 until 1983.	SUP c3		
	• Davie Dodds is a member of the Neuchâtel Xamax from 1986 until 1986	SUP c4		
	• Davie Dodds is a member of the Aberdeen F.C. from 1986 until 1989	SUP c5		
`ChronoFact`	**Evidence Events in Chronological Order:**			
	1. Davie Dodds is a member of the Dundee United F.C. from 1975 until 1986	SUP c1	SUP	SUP
	2. Davie Dodds is a member of the Arbroath F.C. from 1977 until 1978	SUP c2		
	3. Davie Dodds is a member of the Scotland national football from 1983 until 1983	SUP c3		
	4. Davie Dodds is a member of the Neuchâtel Xamax from 1986 until 1986	SUP c4		
	5. Davie Dodds is a member of the Aberdeen F.C. from 1986 until 1989	SUP c5		

Table 7: Sample claim in ChronoClaims dataset involving overlapping events.

Claim: Georgi Andonov was a member of the Botev Plovdiv from 2002 to 2006, joined the Bulgaria national under-21 team from 2003 to 2005, returned to Botev Plovdiv in 2009, played for PSFC Chernomorets Burgas from 2010 to 2012, and then became a member of PFC Beroe Stara Zagora starting in 2015. Label: SUP				
Claim Events	c1: Georgi Andonov was a member of the Botev Plovdiv from 2002 to 2006 c2: Georgi Andonov joined the Bulgaria national under-21 team from 2003 to 2005. c3: Georgi Andonov returned to Botev Plovdiv in 2009 c4: Georgi Andonov played for PSFC Chernomorets Burgas from 2010 to 2012 c5: Georgi Andonov became a member of PFC Beroe Stara Zagora starting in 2015	Claim Event Label	Chrono. Order	Final Claim
TACV	**Evidence Events:**			
	• Georgi Andonov is a member of the Botev Plovdiv from 2002 until 2006	SUP c1 REF c3	N.A.	REF
	• Georgi Andonov is a member of the Bulgaria national under-21 team from 2003 until 2005	SUP c2		
	• Georgi Andonov is a member of the PSFC Chernomorets Burgas from 2010 until 2012	SUP c4		
	• Georgi Andonov is a member of the PFC Beroe Stara Zagora from 2015	SUP c5		
	• Georgi Andonov is member of Botev Plovdiv from 2009 until 2009			
`ChronoFact`	**Evidence Events in Chronological Order:**			
	1. Georgi Andonov is a member of the Botev Plovdiv from 2002 until 2006	SUP c1	SUP	SUP
	2. Georgi Andonov is a member of the Bulgaria national under-21 team from 2003 until 2005	SUP c2		
	3. Georgi Andonov is a member of the Botev Plovdiv from 2009 until 2009	SUP c3		
	4. Georgi Andonov is a member of the PSFC Chernomorets Burgas from 2010 until 2012	SUP c4		
	5. Georgi Andonov is a member of the PFC Beroe Stara Zagora from 2015	SUP c5		

Table 8: Sample claim in ChronoClaims dataset involving recurring event.

7 Conclusion

We have introduced a framework for temporal claim verification that incorporates temporal reasoning based on the chronological order of events. By utilizing a multi-level attention encoder, `ChronoFact` effectively captures the relevance of evidence events to claim events. This enables `ChronoFact` to accurately predict to claim event labels and verify the chronological order consistency between claim and evidence events before determining the overall claim label. We have curated a new benchmark dataset that involves complex claims with multiple events that may overlap or recur. Extensive experiments on multiple datasets have shown that `ChronoFact` outperforms state-of-the-art models.

Acknowledgments

This work is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001).

References

[Allein *et al.*, 2021] Liesbeth Allein, Isabelle Augenstein, and Marie-Francine Moens. Time-aware evidence ranking for fact-checking. *Journal of Web Semantics*, 71:100663, 2021.

[Allein *et al.*, 2023] Liesbeth Allein, Marlon Saelens, Ruben Cartuyvels, and Marie Francine Moens. Implicit temporal reasoning for evidence-based fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 176–189, 2023.

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

[Aly *et al.*, 2021] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. FEVEROUS: Fact extraction and VERification over unstructured and structured information. 2021.

[Barik *et al.*, 2022] Anab Maulana Barik, Wynne Hsu, and Mong Li Lee. Incorporating external knowledge for evidence-based fact verification. In *Companion Proceedings of the Web Conference 2022*, pages 429–437, 2022.

[Barik *et al.*, 2024] Anab Barik, Wynne Hsu, and Mong-Li Lee. Time matters: An end-to-end solution for temporal claim verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 657–664, 2024.

[Chung *et al.*, 2024] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[De Cao *et al.*, 2020] N De Cao, G Izacard, S Riedel, and F Petroni. Autoregressive entity retrieval. In *ICLR 2021-9th International Conference on Learning Representations*, volume 2021. ICLR, 2020.

[Jia *et al.*, 2024] Zhen Jia, Philipp Christmann, and Gerhard Weikum. Faithful temporal question answering over heterogeneous sources. In *Proceedings of the ACM on Web Conference 2024*, pages 2052–2063, 2024.

[Liu *et al.*, 2020] Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, 2020.

[Mori *et al.*, 2022] Marco Mori, Paolo Papotti, Luigi Bellomarini, and Oliver Giudice. Neural machine translation for fact-checking temporal claims. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 78–82, 2022.

[Pan *et al.*, 2023] Liangming Pan, Xinyuan Lu, Min-Yen Kan, and Preslav Nakov. Qacheck: A demonstration system for question-guided multi-hop fact-checking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 264–273, 2023.

[Qudus *et al.*, 2023] Umair Qudus, Michael Röder, Sabrina Kirrane, and Axel-Cyrille Ngonga Ngomo. Temporalfc: A temporal fact checking approach over knowledge graphs. In *International Semantic Web Conference*, pages 465–483. Springer, 2023.

[Soleimani *et al.*, 2020] A Soleimani, C Monz, and M Worring. Bert for evidence retrieval and claim verification. *Advances in Information Retrieval*, 12036:359–366, 2020.

[Stammbach and Neumann, 2019] Dominik Stammbach and Guenter Neumann. Team domlin: Exploiting evidence enhancement for the fever shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109, 2019.

[Thorne *et al.*, 2018a] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, 2018.

[Thorne *et al.*, 2018b] James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Venktesh *et al.*, 2024] V Venktesh, Abhijit Anand, Avishek Anand, and Vinay Setty. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. In *47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024*, pages 650–660. Association for Computing Machinery (ACM), 2024.

[Vrandečić and Krötzsch, 2014] Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[Zhong *et al.*, 2020] Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, 2020.

[Zhou *et al.*, 2019] Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun.

Gear: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, 2019.