Temporal Relation Processing: An XAI Perspective

Sofia Elena Terenziani (Author)
IT University of Copenhagen
sote@itu.dk

Anna Rogers (Supervisor)
IT University of Copenhagen
anro@itu.dk

Course Code: BIBAPRO1PE
Student Number: 19427
IT University of Copenhagen, Denmark
May 15 2024

Temporal Relation Processing: An XAI Perspective

Sofia Elena Terenziani
IT University of Copenhagen
sote@itu.dk

Abstract

Temporal annotations are used to identify and mark up temporal information in text. However, it remains unclear whether language models effectively rely and apply these temporal linguistic strategies when making decisions about time, especially from the perspective of explainability. This project investigates how BERT models of varying sizes handle temporal information in a temporal relation classification task. We define valid reasoning strategies based on the linguistic principles that guide commonly used temporal annotations. Using a combination of saliency-based and counterfactual explanations, we examine if the models' decisions are in line with these strategies. Our results indicate that the selected models do not rely on the expected linguistic cues for processing temporal information on a temporal relation classification task. [1]

1 Introduction

Interpreting and processing temporal information is a fundamental aspect of natural language, guiding comprehension and narrative flow (Mani et al., 2006; Uzzaman, 2012; ter Meulen, 1997). In the field of NLP, processing temporal information involves primarily the task of identifying temporal expressions (Verhagen et al., 2007a) and determining the temporal relations among actions and events (Verhagen et al., 2010; UzZaman et al., 2013a). These tasks enable a range of applications, from text summarisation and narrative analysis to more sensitive operations such as examining clinical records. Transformer-based pre-trained language models have shown impressive abilities in such applications (Xiong et al., 2024; Ko et al., 2023; Basyal and Sanghvi, 2023; Tai, 2024; Shi et al., 2023). Yet, their interpretation of time diverges from human understanding, making it challenging to evaluate their temporal processing capabilities, and whether they indeed possess the abilities to truthfully interpret temporal information (Qiu et al., 2023; Gurnee and Tegmark, 2024).

[1] https://github.com/seterenziani/TRC-XAI

> Leon **won** the Gusher Marathon only a few years after he **underwent** brain surgery in early 2011 .

Figure 1: A sample question from the MATRES (Ning et al., 2018a) dataset. A model is asked to predict the temporal relationship between winning a marathon and having brain surgery. Following annotation guidelines, Blue tokens mark temporal expressions, while orange tokens mark temporal signals that relate temporal expressions to events.

While temporal benchmarks (Tan et al., 2023a; Zhou et al., 2019; Ning et al., 2020; Zhou et al., 2021) have been extensively developed to evaluate models on their processing of temporal information, performance metrics alone do not reveal the underlying mechanisms or explain how conclusions are reached. They do not capture the full picture of how a model processes temporal information (Feng et al., 2023; Kumar and Talukdar, 2020; Aggarwal et al., 2021).

Explainability seeks to make AI decision-making more understandable and predictable to humans. It offers methods to reveal the reasons behind AI behaviors, prioritizing robustness, interpretability, transparency, and reliability to build trust in AI systems (Saeed and Omlin, 2021; Yang, 2023; Došilović et al., 2018; Zhao et al., 2023a; Danilevsky et al., 2020). Despite its promise, its application in understanding how language models process temporal information is still limited. Limiting factors in exploring temporal information include the complexity and variety of its expression in language (Rojat et al., 2021; Chu et al., 2023).

Standard annotations (Ning et al., 2018b; Verhagen et al., 2007b; Pustejovsky et al., 2003) have been developed to identify all temporal information present in text. These schemes outline the specific linguistic cues necessary for accurately processing temporal information. This project investigates whether pre-trained language models process temporal information using these linguistic cues. We draw from the line of research on being "right for the right reason". Specifically, we adopt the framework proposed by Ray Choudhury et al.

(2022) on a temporal relation classification task. We define valid reasoning strategies, and use a combination of saliency-based and example-based explainability methods to assess whether a model follows these strategies when making decisions. We examine models from the BERT-family of different sizes, with the aim on analysing whether larger models, which are extensively trained on more data, are also more likely to base their decision on valid information retrieval and processes. We find that despite showing better performance on tasks, even larger models do not follow the expected reasoning strategies. They might learn shortcuts or spurious correlations, rather than the underlying linguistic phenomena.

2 Related Works

Annotations & Related Tasks: Standardized annotations have been developed as formats to identify and describe elements related to time in natural text. Schemes such as TimeML (Mani et al., 2006) and ISO-TimeML (Pustejovsky et al., 2010) include annotations for some or all temporal expressions (`TIMEXes`), events, temporal relations (`T-LINKS`), temporal signals (`SIGNAL`) and temporal relation types. Annotation frameworks (Rogers et al., 2022; Bethard et al., 2012; Ning et al., 2018a) have been developed to provide guidelines for annotating large datasets on the basis of these standard schemes. Examples of large-scale annotated English datasets based on TimeML standards are TimeBank (Pustejovsky et al., 2003), TimeBankDense (Verhagen et al., 2007b), TDDiscourse (Naik et al., 2019) and MATRES (Ning et al., 2018a).

The TempEval (UzZaman et al., 2013b) workshops is part of the broader SemEval framework and focuses on advancing temporal information processing. Each iteration of TempEval (Verhagen et al., 2007a, 2010; UzZaman et al., 2013a) has introduced specific tasks. These primarily include temporal expression extraction and normalization, and temporal relation extraction.

Benchmarks: Benchmarks have been developed to assess the temporal processing abilities of large language models. Benchmarks range widely in format and scope. TimeQA (Chen et al., 2021) is designed to assess LLMs in question/answering tasks with special focus on temporal information. TEMPREASON (Tan et al., 2023b) adopts question/answering as a format and focuses specifically on time-event relations. TORQUE (Ning et al., 2020) is designed to evaluate models on a temporal reading comprehension task, adopting extractive question/answering format. MCTACO (Zhou et al., 2019), TRAM (Wang and Zhao, 2023) and TimeDial (Qin et al., 2021) are designed to evaluate temporal commonsense reasoning and adopt either a multiple-choice or multiple-choice cloze format.

Commonly used benchmarks have shown some limitations, also here ranging from task and scope. Models have shown to perform well on benchmarks related to commonsense reasoning due to their format rather than truthful reasoning abilities (Tan et al., 2023c), while benchmarks with special focus on temporal expressions, such as numeric years, have shown to not represent the full range of diversity of temporal expressions (Qin et al., 2021). Benchmarks that require language understanding are based on the naive assumption that performing better on such tests would inherently mean progress in models general language processing (Sugawara et al., 2019; Weston et al., 2015; Ray Choudhury et al., 2022). Generally speaking, performance on benchmarks alone, while useful, fails to capture the full picture of language models' processing capabilities, especially how a model reached its conclusions (Kumar and Talukdar, 2020).

Explainability: Explainability methods can account for some of the limitations that current benchmarks might have (Zhao et al., 2023a). Methods can help us understand a model's performance in more detail, by showing what a model pays attention to, or where a model fails to perform. Essentially, they can provide means to examine whether models are reliable, meaning that they not only perform right, but are consistently "right for the right reasons" (McCoy et al., 2019; Christianson, 2016). For this line of research, local and post-hoc explainable methods have been used to evaluate pre-trained language models on specific tasks. Ray Choudhury et al. (2022) apply a combination of these methods to analyze and evaluate models on two linguistic skills required for a reliable reading comprehension system. Du et al. (2021) apply similar methods to debug pre-trained language models in natural language understanding tasks. Both studies find that models use shortcuts rather than valid inference strategies.

In the context of large language models, explainability methods are both necessary[2] and challenging[3]. Research efforts are also put into examining the interpretability (González et al., 2021; Schuff et al., 2022) and the reliability (Harbecke and Alt, 2020; Spreitzer et al., 2022; Rahimi and Jain, 2022) of these methods.

Contribution: Evaluating temporal processing in NLP has primarily involved developing benchmarks to assess a model's ability to process temporal information, and implementing annotation schemes to mark all or some temporal information in text. Despite these developments, the use of explainability methods to examine how models process temporal data remains unexplored. This project aims to fill this gap. We utilize post-hoc and local explainability methods to examine whether models rely on valid linguistic cues when making decisions in a temporal relation classification task. Specifically, we use temporal annotation guidelines to define valid reasoning strategies and evaluate how well the models adhere to these strategies when making decisions.

[2]Often referred to as "black boxes", large language models are not inherently explainable.

[3]The complexities of large language models also makes the explanation themselves difficult to interpret.

3 Temporal Processing

Temporal processing in natural language involves understanding and interpreting temporal aspects of language. Given the variety of language structures that represent time, accurately understanding temporal dimensions requires a combination of linguistic, arithmetical, logical, factual, and commonsense reasoning. (Sanampudi and G.Vijaya, 2010; Wenzel and Jatowt, 2023).

Human understanding of time is deeply influenced by human experiences and cultural contexts (Callender, 2011). It is shaped by *experiencing* aspects like typical duration (e.g., how long it takes to eat a meal), causality (e.g., a door needs to be opened to pass through it), and frequency (e.g., how often you would buy groceries). Considering the example from Figure 1, knowledge of the typical recovery time helps us appreciate the challenge of winning a marathon after surgery. In contrast, language model's understanding of time is based on recognizing patterns and lexical cues within the data they have been trained on. This difference complicates the evaluation of machines' abilities to interpret temporal information. It raises questions about what it means to "understand" time (Rojat et al., 2021), and presents practical challenges in determining whether models fundamentally have the capabilities to interpret temporal information truthfully (Sanampudi and G.Vijaya, 2010).

While pre-trained models might not inherently understand time, they can be assessed on their ability to analyze linguistic structures to derive meaning about time. These considerations suggest that we might reconsider how we evaluate models on temporal tasks, focusing on their ability to draw temporal meaning from linguistic cues alone. This approach provides clear success criteria for temporal tasks: it allows a comparison between the properties a model has learned during training and inference strategies we expect the model to perform.

4 Processing Temporal Relations

4.1 Success Criteria for Temporal Relation Classification

Temporal relation classification involves identifying the temporal relationships between events or actions in a given text. That is determining whether they happen simultaneously, sequentially, or at different points in time (Ning, 2019). Focusing on the principle of performing "right for the right reasons", Ray Choudhury et al. (2022) defines three success criteria for NLP systems: a system should (a) accurately perform on a specific task, (b) do so by relying on valid reasoning steps, and (c) do so consistently under distribution shifts. We evaluate a model on the task of temporal relation classification based on the three success criteria: we define the expected reasoning steps for the model, and evaluate whether a model adheres to these strategies, and does so consistently.

4.2 What reasoning should a model perform?

For a model to correctly extract a temporal relation between two events or actions, it is expected to 1) be able to identify and interpret the linguistic features that express temporal information, 2) be able to map these features to the event or action that they are describing of modifying, and 3) to use that information to infer the temporal relationship between the events. We define these as valid reasoning steps (see Table 1).

TimeML (Mani et al., 2006) is a framework for annotating all temporal information in text. While we don't directly use TimeML annotations, we rely on its guidelines (Setzer, 2002) to identify the individual linguistic features essential for classifying temporal relations. We focus on three annotations: `TIMEX3` for explicit temporal expressions, `TLINK` for event relationships, and `SIGNAL` for cues that clarify the temporal relationships among events. For simplicity, we organize these tokens into five main groups (see Table 4).

- **Temporal Expressions:** Temporal expressions are explicit tokens that specify points or durations in time, such as dates (e.g. "January 1st"), times (e.g. "at noon"), duration (e.g. "two weeks") and frequencies (e.g. "weekly").
- **Temporal Prepositions and Adverbs:** Temporal prepositions (e.g. "in", "at") and adverbs (e.g. "recently") are tokens used to connect actions or events to specific times.
- **Temporal Conjunctions**: Temporal conjunctions are tokens used to relate events to each other, showing how events relate to each other in terms of sequencing (e.g. "then", "next"), simultaneity (e.g. "while"), or depending on each other (e.g. "before", "after").
- **Subordinate Conjunctions:** Subordinate conjunctions are tokens used to related events or actions in conditional (if-then relationship) or causal relationship (cause-effect relationship). They represent which condition must be met before an outcome can happen (e.g. "if", "then"), or which cause precedes its effect (e.g. "because", "therefore").
- **Verb Tenses and Aspects:** Tenses and aspects are properties of verbs, which provide cues for understanding the timing of actions or events. Tenses directly indicate when an event or action occurred, whether it happens in the past, present or future. Verb aspect indicate whether an action is completed (e.g. "has walked) or ongoing (e.g. "is walking").

4.3 What reasoning does a model perform?

Having established what reasoning steps a model should perform, the next step would be to determine whether a specific model follows these steps. Ray Choudhury et al. (2022) utilises a combination of example-based and saliency-based methods. These explainability methods are categorized as local and post-hoc: they focus

	Reasoning Step	Relevant Features
Leon won the Gusher Marathon only a few years after he underwent brain surgery in early 2011. ⟨won, **?**, underwent⟩	Identifying temporal information	Times: *years* Date: *2011* Conjunction: *after*
	Mapping temporal information to events	*underwent* := *2011* *won* := (*years, after*)
	Determining temporal relationship between events	*won* := *years after 2011* **Solution:** ⟨won, **AFTER**, underwent⟩

Table 1: Valid reasoning steps for determining the temporal relation between two events, having already identified a given event pair.

	Purpose	# Docs	#Events	#TLinks
TimeBank	Training	162	6.6k	6.5k
Aquaint	Training	73	4,3k	6.4k
Platinum	Validation	20	748	837
Total		275	6k	13.5k

Table 2: Summary of purpose and key statistics for subsets of MATRES (Ning et al., 2018a) dataset.

Label	#	%
BEFORE	6.886	50%
AFTER	4.576	34%
VAGUE	1.644	12%
EQUAL	471	4%

Table 3: Label distribution for the entirety of the MATRES (Ning et al., 2018a) dataset.

on individual instances and they are applied after model has been trained.

Saliency-based Methods: Saliency-based methods are a family of methods that offer feature-centred explanations, focusing on how individual tokens affect a model's prediction (Molnar, 2022; Ding and Koehn, 2021a). These methods offer different ways of computing a score for each token in an input instance, showing how 'important' individual tokens are for a model's decision. By comparing the saliency scores to a predefined partition of tokens, these explanations can be used to determine whether a model is relying on the right information for correct predictions. Following Ray Choudhury et al. (2022), we define a partition of the token space as: the tokens a model should find important (positive), and the tokens a model should not find important (negative) (§ 5.5). From the example instance in Figure 1, important tokens includes `{years, after}`. These tokens are fundamental for a model to extract the temporal relationship `<won, after, underwent>`. If saliency scores show, that a particular model consistently has higher scores on the positive partition of tokens compared to the negative partition of tokens, it is likely that the model is focusing on the right features for making its decisions.

Counterfactual Explanation: Counterfactual explanations offer data-centred explanations by analyzing how changes in the input data can lead to different model predictions (Molnar, 2022). By changing parts of the input with alternative valid tokens that would alter the true label, these explanations can help determine if a model is relying on the right reasoning strategies (§ 5.6). From the example instance in Figure 1, altering the tokens from `{years, after}` to `{months, before}` would alter the relationship from `<won, after, underwent>` to `<won, before, underwent>`. If a model accurately predicts the temporal relationship of both the original instance and its counterfactual counterpart, it suggests that the model consistently relies on the correct information.

Explanation Alignment: For a model to demonstrate valid reasoning, both types of explanations, saliency-based and counterfactual, must show agreement across many instances. This explanation alignment is defined by two criteria: both the original and its counterfactual counterpart must have the correct prediction outcome, and these are obtained by relying on valid features, as indicated by the saliency scores.

5 Methodology

5.1 Data

The experiments are conducted using the MATRES dataset (Ning et al., 2018a). In total, MATRES includes 275 news articles from TempEval3 (UzZaman et al., 2013a) [4], annotated for temporal relations between pair of events and divided into three sections: TimeBank, Aquaint, and Platinum. For experimental consistency, we follow the original split for training and evaluation

[4]TempEval3 dataset was found to be inaccessible for online download. However, a portion of the dataset has been made available through the repository of MATRES (Ning et al., 2018a). All mentioned size and figures are accurate to the available source as of the conclusion of the project.

	Common Features	Examples
Temporal Expressions: Tokens that specify points in time	Absolute expressions, such as *December 2025, at 5PM*	She started a new job on September 1st, after moving to the city.
	Relative expressions, such as *week, Mondays, annually*	If it rains tomorrow, the picnic will be postponed until Sunday.
Temporal Prepositions and Adverbs: Tokens used to connect actions or events to specific times.	Prepositions such as *at, on, in, during, for, over, by*	She started a new job on September 1st, after moving to the city.
	Adverbials such as *again, late, now, then eventually, previously, recently*	Recently, he has taken up running before breakfast at 8AM.
Temporal Conjunctions: Tokens used to related events to each one another.	Conjunctions such as *before, after, while, until, since when, as soon as, as long as*	She started a new job on September 1st, just after moving to the city.
		Recently, he has taken up running before breakfast every morning.
Subordinate Conjunction: Tokens used to express conditional or causal relationship between events or actions.	References to causality such as *because, therefore, as*	Because you didn't reply in time, I only bought tickets for two.
	References to conditions such as *if, unless, then, so*	If it rains tomorrow, then the picnic will be postponed until Sunday at noon.
Verb Tenses, Aspects: Properties of verbs used to express the timing of actions and events.	Tense describing past, present, future: *walked, walks, will walk*	She ***started*** a new job on September 1st, just after ***moving*** to the city.
	Aspect describing ongoing or finished events: *is walking, have walked, have been walking*	Recently, he ***has taken*** up ***running*** before breakfast every morning.

Table 4: Categorisation of the essential linguistic features that express temporal information. Color coding follows the annotation guidelines from TimeML (Mani et al., 2006): orange is used for signal tokens (`SIGNAL`), providing cues for how events and temporal expressions are related to each other; blue is used for specific time expressions (`TIMEX3`).

(Ning et al., 2019) (see Table 2). For a temporal relation classification task, MATRES is annotated for four different temporal relation classes: BEFORE, AFTER, EQUAL and VAGUE. The VAGUE class is given to a pair of events, where the temporal relationship is either ambiguous or uncertain. In the context of temporal relation, there's a distinction made in how "events" are defined and annotated. MATRES focuses on actions expressed through verbs, particularly selecting the main verb as the event (e.g., in "I will go to the movies, after having eaten", "go" and "eaten" are identified as events).

5.2 Models

For some use cases, larger models have shown to generalise better compared to smaller models (Zhong et al., 2021; Desai and Durrett, 2020). Part of this project is set to investigate whether they are also more likely to rely on the right information when presented to temporal information. Therefore, the experiments are based on four transformer-based models from the BERT family of different sizes, namely BERT-base-uncased, BERT-large-uncased (Devlin et al., 2019), RoBERTa-base, and RoBERTa-large (Liu et al., 2019). The models differ mainly in scale, architecture and number of parameters, while RoBERTa also uses different optimisation techniques, is pre-trained on more data, and has a larger vocabulary compared to BERT. BERT is selected to ensure experimental consistency with the framework we are adopting (Ray Choudhury et al., 2022). We argue that a comparative analysis with more recent generative models is highly valuable. However, we leave this to further work.

5.3 Fine-Tuning

Each model is fine-tuned for a temporal relation classification task using the architecture and tokenization strategy proposed by Yanko et al. (2023) and Baldini Soares et al. (2019). The strategy consists in explicitly marking the boundaries of each action in an input sentence with special tokens. We define these special tokens as `[a1]`, `[/a1]`, `[a2]`, `[/a2]` and process each input sentence as following:

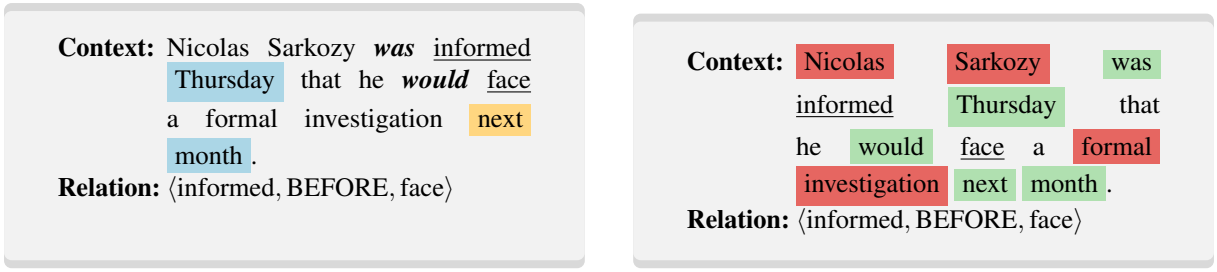

Figure 2: An instance from the MATRES dataset. Positive (green) and negative (red) partition of tokens is defined by the individual features defined as relevant for defining the temporal relation between two events.

Leon [a1] won [a1] the Marathon a few years after he [a2] underwent [/a2] brain surgery.

These special tokens are used to help the model focus on the event pair. As an input is processed by each encoder, the embedding of these special tokens are adjusted based on surrounding tokens. This results in a context-specific representation for each marked event. We concatenate the embedding vectors of the special tokens and use them for classification by feeding them into a simple linear layer on top of each encoder (see Appendix A for details on configurations).

5.4 Evaluation Metrics

We evaluate each model using standard evaluation metrics for classification: F1 (the harmonic mean of precision and recall) and Exact-Match (proportion of total correct predictions out of all predictions made). The MATRES dataset shows considerable class imbalance (see Table 3). To address this, the F1-score is particularly important. We report both the weighted and macro-average F1-score. While Exact-Match might be misleading in the context of imbalanced datasets, it is reported because of its straightforward interpretability.

5.5 Token partition

We have previously defined specific linguistic features used for expressing information about a temporal relationship (§ 4.2). Token partitioning is guided by this definition. The positive token partition is defined as all individual tokens that express or clarify the temporal relationship between two events. This includes specific linguistic elements such as temporal expressions, prepositions, conjunctions, and the tense and aspect of verbs. The negative token partition is defined as tokens that are not part of the positive partition and do not match the queried tokens for the event pair. These are tokens that are defined as irrelevant for capturing the temporal relationship.

Figure 2 shows the relevant tokens for an instance, including temporal expressions, temporal conjunctions and verbs, and how these tokens are used to define the partition of tokens into positive and negative features.

5.6 Counterfactual Explanations

Counterfactual explanations are manually generated from 200 instances randomly selected from the validation (Platinum) dataset. Counterfactual explanation instances are manually generated with minimal alteration to the original input structure. The event pair relevant for the temporal relation remains unchanged[5], and the alterations are solely performed on the context surrounding them. The development process involves a two-stage approach: (a) identifying the positive partition of tokens (§ 5.5), likely to impact predictions significantly, and (b) modifying these tokens to reverse or change the temporal relationship and prediction outcome.

We identify four types of possible and semantically correct alterations, each dependent on the nature of the relationship.

1. We consider simple temporal relationships those that contain explicit temporal conjunctions (e.g. "before", "after" and "while"). For simple temporal relationships, revering the temporal conjunction and changing verb tenses were sufficient as semantically correct alterations (see Table 3.a). This strategy most often resulted in reversing BEFORE and AFTER relationships.

2. For instances where a direct reversal of temporal conjunction or verb tense change was not possible, temporal conjunctions or adverbs (e.g. "subsequently", "already") and temporal expressions (e.g. "months", "years") were added or removed (see Table 3.b). This strategy often resulted in altering BEFORE or AFTER relationships to an EQUAL relationship, or vice-versa.

[5]The event pairs are defined by verbs, and most often alterations required reversing the verb tenses of these verbs. Given how verb tenses are constructed in the English language and that the event pair for the most part represented verb base form, alterations were possible for almost every randomly selected instance without changing the original pair. One drawback is that it is not always possible to switch to perfect tenses, which is useful to indicate the completion of an action (e.g. *"I will finish"*, *"I had finished"*)

Original: Leon won the Gusher Marathon a few days before he underwent brain surgery.
Relation: ⟨won, BEFORE, underwent⟩

Altered: Leon won the Gusher Marathon a few years after he underwent brain surgery.
Relation: ⟨won, AFTER, underwent⟩

(a) Example of reversal of temporal conjunctions.

Original: A computer that is about to be deployed, is taking computing into (...)
Relation: ⟨deployed, AFTER, taking⟩

Altered: A computer that had already been deployed for some months, is taking computing into (...)
Relation: ⟨deployed, BEFORE, taking⟩

(b) Example where direct reversal was not possible.

Original: But if it performs as (...), the design could be used to charge even the most powerful systems.
Relation: ⟨performs, BEFORE, used⟩

Altered: But if it is used to change (...), it indicates that the system currently performs as (...).
Relation: ⟨performs, EQUAL, used⟩

(c) Example with a conditional relationship.

Original: Lowe took part in the trans-antarctic expedition. He also made expeditions to (...).
Relation: ⟨took, VAGUE, made⟩

Altered: He also made expeditions to (...). Lowe later took part in the trans-antarctic expedition.
Relation: ⟨took, AFTER, made⟩

(d) Example with reordering of sentences.

Figure 3: Examples of counterfactual altered instances, with alteration highlighted in yellow.

3. We consider more complex relationships those that include conditional or causal relationships between the two events. Focus was put in not altering the nature of such relationships. For these cases, reversing the temporal relationship involved reversing the cause with the effect or vice-versa [6](see Table 3.c).

4. For actions described in separate sentences, reordering the sentences was considered as a valid semantic alteration. This alteration is possible and particularly relevant for the dataset at hand, which is based on news snippets. For the news domain, the order of mention often dictates the sequence of events. This strategy often resulted in altering to or from a VAGUE relationship. Reordering sentences within the text, by placing them closer or further apart, either increased or decreased the contextual dependency between a pair of actions (see Table 3.d).

5.7 Saliency Scores

We obtain saliency scores from two different methods: Occlusion and Integrated Gradients.

Occlusion (DeYoung et al., 2020) is a perturbation-based method. It works by systematically replacing the input token with a baseline token and observing the changes in the model's output probabilities. The occlusion score for each token represents the change in the model's output probability when the token is occluded. We select `[MASK]` as the baseline token to represent the absence of a specific feature. By replacing each token one at a time with `[MASK]`, we remove the specific information provided by that specific token and observe how its absence affects the model's output.

Integrated gradient (Sundararajan et al., 2017; Molnar, 2022) is a gradient-based method. This family of methods work by quantifying how much each token in the input contributes to the gradient being propagated downstream. Tokens that have a larger impact on the output will impact the gradient more, and are considered more influential. Integrated gradients in particular work by comparing the actual input against a baseline input. We again select `[MASK]` as the baseline token, and create baseline inputs based on the length of the original input. Gradients are computed along a linear path, from the baseline to actual input. This path represents a transition from absence of features to the actual input. The gradients are accumulated at multiple steps along the linear path. The results is a vector for each token, representing a separate gradient value for each of a feature's dimension. To convert these vectors to a single saliency score per token, we apply normalization using $L2$ norm (Ray Choudhury et al., 2022).

[6]For example, "If the train will arrive late, I would have to take a taxi." becomes "If I take a taxi, the train will probably arrive on time.", and the temporal relationship between "take" and "arrive" would be correctly reversed.

	F1 M/avg	F1 W/avg	EM
RoBERTa $_{\text{large}}$	0.60	0.68	0.67
RoBERTa $_{\text{base}}$	0.59	0.67	0.67
BERT $_{\text{large-uncased}}$	0.56	0.65	0.63
BERT $_{\text{base-cased}}$	0.54	0.61	0.60

Table 5: Performance of different models from the BERT-family on the MATRES (Ning et al., 2018a) dataset.

	Original		Counterfactual	
	F1	EM	F1	EM
RoBERTa $_{\text{large}}$	0.64	0.63	0.49	0.49
RoBERTa $_{\text{base}}$	0.57	0.58	0.41	0.40
BERT $_{\text{large-uncased}}$	0.54	0.53	0.36	0.37
BERT $_{\text{base-cased}}$	0.54	0.55	0.34	0.36

Table 6: Performance of models from the BERT-family on 200 counterfactual instances over their original counterpart.

Special tokens `[a1]`, `[/a1]`, `[a2]`, `[/a2]`, which we introduced during fine-tuning (Section 5.3), must be carefully considered when applying both methods. For Integrated Gradients, the special tokens are included in the baseline inputs, to ensure the integrity of the input. For Occlusion, the special tokens are not occluded/perturbed. Essentially, the occlusion scores measure the impact of each regular token on how these special tokens are represented, which in turn affects the model's predictions.

For the temporal relation classification task on the MATRES dataset, applying each saliency method results in four scores per token. These represent the individual token's impact on a specific label (BEFORE, AFTER, EQUAL and VAGUE). We aggregate these scores into a single value by summing over each score. This score represents the each token's overall significance for the model's predictions across all temporal classes. [7].

5.8 Explanation Alignment Score

Recalling § 4.3, explanation alignment is defined as when a model is (a) accurately predicting both counterfactual and original instance, and (b) is doing so by relying on the right information, as reflected in saliency scores. For each model, an alignment score is computed as a measure of point (b). We compute the alignment score for the instances in the validation dataset (Platinum) where both original and its counterfactual counterpart are correctly predicted. For a single instance with a random partition of tokens, the positive and negative partitions should have similar saliency scores. For a dataset this translates to them being significantly different in $\approx 0\%$ of cases. Explanation alignment score for a dataset is defined as the percentage of instances where the positive partition of tokens has a statistically significant higher average saliency score than the negative partition of tokens. For computing the statistical significance test, a one-tailed independence T-test is used, with p-value equal to 0.05. We follow Ray Choudhury et al. (2022) and formulate the null hypothesis as: *"the positive partition of tokens does not have higher average saliency score than the negative partition"*.

6 Results & Analysis

6.1 Model Evaluation

Table 5 shows the performance of each fine-tuned model on the MATRES dataset. Across all models, we consistently see that weighted average F1-scores exceed macro average F1-scores. Macro-average F1-score is calculated by taking the F1-score for each class independently, regardless of its size, and average these scores. This trend indicates difficulties in predicting the minority classes, such as VAGUE [8].

The performance scores indicate that larger models generally outperform smaller models. However, the advantage of RoBERTa-large over the smallest model is relatively low, showing 0.07 improvement on weighted average F1-score. This suggests that the notion, that larger models might perform better for some use cases (Zhong et al., 2021; Desai and Durrett, 2020), only partially holds true for a larger pre-trained encoders fine-tuned for a temporal relation classification task on the MATRES dataset.

The moderate performance in the temporal relation classification task is not unique to these experiments. This task has posed challenges across various benchmarks (Yuan et al., 2023; Galvan et al., 2018; Mathur et al., 2021), and even human annotators struggle with it, with evidenced of low annotation agreement rates (Derczynski, 2016; Jin, 2022).

6.2 Counterfactual Evaluation

Table 6 shows the performance of 200 counterfactual explanation instances compared to their original counterparts, using weighted average F1-score and Exact-Match metrics. For all models, we observe a significant decrease in performance on counterfactual instances compared to the original instances, both in terms of F1-score and Exact-Match. The average performance decline is 18%. Counterfactual explanations test a model's ability to apply valid reasoning strategies consistently

[7] Summing or averaging are common approaches for aggregating the influence of the token across classes (Molnar, 2022; Atanasova et al., 2020a). Important to note is that both might overlook the importance of tokens that are particularly influential for a specific class.

[8] To account for class imbalance in evaluation, Yanko et al. (2023) proposes a relaxed F1-score, which is reported in more detail in the Appendix B

	Counterfactual	
	IG	Occlusion
RoBERTa $_{\text{large}}$	0.16	0.62
RoBERTa $_{\text{base}}$	0.17	0.64
BERT $_{\text{large-uncased}}$	0.33	0.55
BERT $_{\text{base-cased}}$	0.39	0.45

Table 7: Explanation alignment score between the portion of the counterfactual instance correctly predicted, and selected saliency-based method (Integrated Gradients and Occlusion). Alignment score is shown for each selected model.

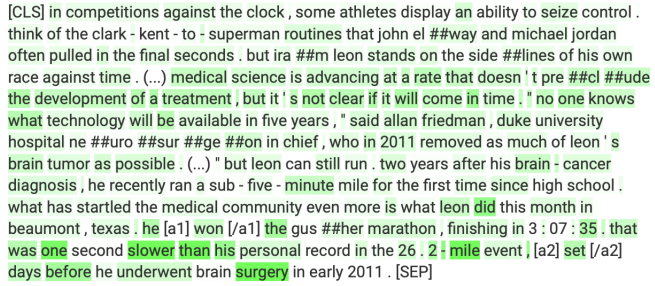

Figure 4: Visualisation of the occlusion saliency scores for one instance, performed on BERT-large. The model correctly predicts the temporal relation between the events "won" and "set". More saturated tokens have higher saliency scores.

across different possible scenarios. A decline in performance on counterfactual compared to original instances shows that models fail to generalize the expected reasoning strategies to altered scenarios.

Larger models show slightly better resilience in weighted F1-score and EM scores compared to their smaller counterparts. However, they also present challenges in performance on the counterfactual instances, which indicates that even larger models are not likely to follow valid reasoning strategies.

Future work could explore the possibility of loosening the criteria that a model's prediction on a counterfactual scenario must exactly match the true class. By examining the prediction probabilities, we might find that while models do not always match the true class exactly, they adjust their class probabilities appropriately in response to changes in the counterfactual scenarios. This is particularly valuable for classification with unbalanced distribution of labels (Molnar, 2022).

6.3 Explanation Alignment

Table 7 shows the explanation alignment score between correctly predicted counterfactual instances against the two selected saliency-based methods. We observe that Integrated Gradients and Occlusion do not agree on the alignment scores, and that this trend is especially present for larger models. This lack of alignment between these two methods is consistent with previous findings (Ray Choudhury et al., 2022; Atanasova et al., 2020b), and it must be addressed to draw appropriate conclusions.

Evaluation of the counterfactual instances suggest that larger models are more likely to perform the expected reasoning operations compared to smaller models (see Figure 6). This trend is consistent with the alignment score presented by Occlusion, with larger models having higher alignment score than smaller models. The explanation alignment with Occlusion suggests that, when larger models make the right decision (correct prediction) and are able to generalise well (correct prediction for both original and counterfactual), they are also more likely to rely their decisions on the right information compared to smaller models.

However, the explanation alignment scores provided with Integrated Gradients shows the opposite trend. Potential nterpretations have been suggested (Ray Choudhury et al., 2022; Harbecke and Alt, 2020). One interpretation is that Integrated Gradients simply do not offer a reliable measure for computing saliency scores, and Occlusion does a better job. Harbecke and Alt (2020) argues that Integrated Gradients fail to account for the discrete nature of text data. The intermediate representations of text that are needed to compute the scores do not fit the discrete word embedding well (Zhao et al., 2023b), and therefore the computed gradients might not produce truthful saliency scores.

Another possible interpretation is that Integrated Gradients are in fact more faithful (Ray Choudhury et al., 2022). The trend shown by Integrated Gradients suggests that as model's size increases, the features we define as important do not align with the model's strategies for correct predictions. Larger models, with their increased capacity, might be more likely to learn complex statistical patterns in the training data, including spurious ones. If the training data contain many of such correlations, a larger model might be more prone to learn them and use them for predictions (Linzen, 2020). This could explain the higher accuracy of larger models compared to smaller ones (§ 6.1), but also indicates that these models might depend on these correlations instead of the "right" information (essentially, performing right for the wrong reasons).

Overall, the findings support the concerns raised by Ray Choudhury et al. (2022) about the disagreement of the selected saliency-based methods. We question their faithfulness and utility in evaluating models for valid reasoning strategies.

7 Discussion

This project explores explainability in pre-trained language models, focusing on how temporal information is processed for a temporal relation classification task. Various benchmarks have been developed to evaluate mod-

els on their temporal processing abilities, with higher performance being interpreted as better temporal processing abilities (Sugawara et al., 2019; Weston et al., 2015). These benchmarks, while useful, do not provide insight into the underlying strategies used by models. In this study, we opted to evaluate temporal processing through the perspective of explainability, guided by the principle of "being right for the right reason". The experiment shows that selected models do not follow the expected reasoning in a temporal relation classification task.

Post-hoc and local explainability methods are often used to evaluate if model decisions are justifiable from a human perspective, but their reliability is often questioned (Dasgupta et al., 2022; Saini and Prasad, 2022). Our project uses a combination of saliency-based and counterfactual explanations. Counterfactual explanations are considered more truthful (Zhao et al., 2023b) since they identify input features that impact predictions. However, they must be carefully created to avoid incorrect or unreliable conclusions. Saliency scores are not always faithful (Jukić et al., 2023; Ding and Koehn, 2021b; Atanasova et al., 2020c). Different saliency methods can produce conflicting results, meaning that they inconsistently reflect the model's decision process. Moreover, there is no ground truth for evaluating saliency methods, making it challenging to evaluate whether they are correctly approximating the model's processes. As explainability is an active research area, we could expect better explanations for these disagreements of saliency methods and more reliable methods to be developed soon.

Having addressed the truthfulness of these methods, the question of their utility remains. For this project, we must conclude that the models follow some other strategy for correct prediction (rather than relying on the expected linguistic cues). Explainability methods aim to make a model's decisions understandable to humans, but this becomes challenging when a model's reasoning doesn't align with human reasoning.

González et al. (2021) shows that humans struggle to predict the answer for poorly performing models even when saliency explanations are provided. Figure 4 [9] illustrates this with saliency scores obtained from Occlusion for a correct prediction using BERT-large. From this example, it is not clear how and why a model arrived at its conclusions. Discovering alternative reasoning strategies or shortcuts through these explanations is challenging because they are not necessarily human interpretable. This raises questions about the practical value of these methods, as they provide only a partial interpretable view of a model's processes.

[9]Research that has used a similar framework (Ray Choudhury et al., 2022; Du et al., 2021) has reported both negative and positive saliency scores. For our calculation of saliency scores, both increase and decrease in probability are treated as a positive contribution to the saliency score.

8 Limitations & Further Work

This project has several limitations that could use further exploration. Firstly, a larger number of counterfactual/original pair of instances is necessary for more definitive conclusions. Moreover, our approach to saliency-scores may need reevaluation. Currently, the saliency-scores do not consider the potential negative impact of individual tokens on predictions, and they are obtained by aggregating all scores without identifying tokens that are particularly influential for a specific class. Exploring alternative saliency-based techniques, such as LIME (Ribeiro et al., 2016), could improve our understanding, especially because of the misalignment between the methods that we are currently using. Our analysis focuses only on non-autoregressive transformer-based models. Autoregressive models process inputs sequentially and might process temporal information in ways more closely aligned with human/expected reasoning strategies (Zhao et al., 2023a). Extending the analysis to include autoregressive models could provide insights into how different model architecture influence how temporal information is processed.

9 Conclusion

Temporal annotations are used to identify and mark some or all linguistic features that convey temporal information within a text. We analyze whether models rely on the linguistic strategies outlined by these annotation guidelines when presented with temporal information. We adopt from the line of research of "being right for the right reason" and evaluate BERT model of varying sizes on a temporal relation classification task using explainability techniques. We define expected reasoning strategies, based on the linguistic features that form the base of temporal annotation systems. Analysis involve a combination of counterfactual explanations and saliency-based methods: saliency-based explanations determine whether a model is relying on the right information to make decisions, while counterfactual explanations evaluate whether a model uses these right information consistently. A high alignment between these two explanations, indicate that a model is following a valid processing strategy. We find this is not the case for the selected models, meaning that they might learn spurious correlations and shortcuts rather than learning the linguistic phenomena that form temporal meaning. Further work is needed to refine the success criteria in temporal tasks, as well as developing more faithful and useful explainability methods.

References

Aggarwal, S., Mandowara, D., Agrawal, V., Khandelwal, D., Singla, P., and Garg, D. (2021). Explanations for CommonsenseQA: New Dataset and Models. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2020a). A diagnostic study of explainability techniques for text classification. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2020b). A diagnostic study of explainability techniques for text classification. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. (2020c). A diagnostic study of explainability techniques for text classification. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Baldini Soares, L., FitzGerald, N., Ling, J., and Kwiatkowski, T. (2019). Matching the blanks: Distributional similarity for relation learning. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Basyal, L. and Sanghvi, M. (2023). Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chatgpt models.

Bethard, S., Kolomiyets, O., and Moens, M.-F. (2012). Annotating story timelines as temporal dependency structures. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2721–2726, Istanbul, Turkey. European Language Resources Association (ELRA).

Callender, C. (2011). *The Oxford Handbook of Philosophy of Time*.

Chen, W., Wang, X., and Wang, W. Y. (2021). A dataset for answering time-sensitive questions.

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good enough, underspecified, or shallow language processing. *Quarterly journal of experimental psychology (2006)*, 69:1–29.

Chu, Z., Chen, J., Chen, Q., Yu, W., Wang, H., Liu, M., and Qin, B. (2023). Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. (2020). A survey of the state of explainable AI for natural language processing. In Wong, K.-F., Knight, K., and Wu, H., editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Dasgupta, S., Frost, N., and Moshkovitz, M. (2022). Framework for evaluating faithfulness of local explanations.

Derczynski, L. (2016). Representation and learning of temporal relations. In Matsumoto, Y. and Prasad, R., editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1937–1948, Osaka, Japan. The COLING 2016 Organizing Committee.

Desai, S. and Durrett, G. (2020). Calibration of pre-trained transformers.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. (2020). ERASER: A benchmark to evaluate rationalized NLP models. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Ding, S. and Koehn, P. (2021a). Evaluating saliency methods for neural language models.

Ding, S. and Koehn, P. (2021b). Evaluating saliency methods for neural language models.

Došilović, F. K., Brčić, M., and Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215.

Du, M., Manjunatha, V., Jain, R., Deshpande, R., Dernoncourt, F., Gu, J., Sun, T., and Hu, X. (2021). Towards interpreting and mitigating shortcut learning behavior of nlu models.

Feng, Y., Zhou, B., Wang, H., Jin, H., and Roth, D. (2023). Generic temporal reasoning with differential analysis and explanation.

Galvan, D., Okazaki, N., Matsuda, K., and Inui, K. (2018). Investigating the challenges of temporal relation extraction from clinical text. In Lavelli, A., Minard, A.-L., and Rinaldi, F., editors, *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 55–64, Brussels, Belgium. Association for Computational Linguistics.

González, A. V., Rogers, A., and Søgaard, A. (2021). On the interaction of belief bias and explanations. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2930–2942, Online. Association for Computational Linguistics.

Gurnee, W. and Tegmark, M. (2024). Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Harbecke, D. and Alt, C. (2020). Considering likelihood in NLP classification explanations with occlusion and language modeling. In Rijhwani, S., Liu, J., Wang, Y., and Dror, R., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 111–117, Online. Association for Computational Linguistics.

Jin, P. e. a. (2022). Temporal relation extraction with joint semantic and syntactic attention.

Jukić, J., Tutek, M., and Šnajder, J. (2023). Easy to decide, hard to agree: Reducing disagreements between saliency methods.

Ko, D., Lee, J., Kang, W.-Y., Roh, B., and Kim, H. (2023). Large language models are temporal and causal reasoners for video question answering. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316, Singapore. Association for Computational Linguistics.

Kumar, S. and Talukdar, P. (2020). NILE : Natural language inference with faithful natural language explanations. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization? In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Mani, I., Verhagen, M., Wellner, B., Lee, C. M., and Pustejovsky, J. (2006). Machine learning of temporal relations. In Calzolari, N., Cardie, C., and Isabelle, P., editors, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760, Sydney, Australia. Association for Computational Linguistics.

Mathur, P., Jain, R., Dernoncourt, F., Morariu, V., Tran, Q. H., and Manocha, D. (2021). TIMERS: Document-level temporal relation extraction. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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.

McCoy, R. T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference.

Molnar, C. (2022). *Interpretable Machine Learning*. LeanPub.

Naik, A., Breitfeller, L., and Rose, C. (2019). TDDiscourse: A dataset for discourse-level temporal ordering of events. In Nakamura, S., Gasic, M., Zukerman, I., Skantze, G., Nakano, M., Papangelis, A., Ultes, S., and Yoshino, K., editors, *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249, Stockholm, Sweden. Association for Computational Linguistics.

Ning, Q. (2019). Understanding Time in Natural Language Text.

Ning, Q., Subramanian, S., and Roth, D. (2019). An improved neural baseline for temporal relation extraction. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *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.

Ning, Q., Wu, H., Han, R., Peng, N., Gardner, M., and Roth, D. (2020). TORQUE: A reading comprehension dataset of temporal ordering questions. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Ning, Q., Wu, H., and Roth, D. (2018a). A multi-axis annotation scheme for event temporal relations. In Gurevych, I. and Miyao, Y., editors, *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.

Ning, Q., Wu, H., and Roth, D. (2018b). A multi-axis annotation scheme for event temporal relations. In Gurevych, I. and Miyao, Y., editors, *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.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003). The timebank corpus. *Proceedings of Corpus Linguistics*.

Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). ISO-TimeML: An international standard for semantic annotation. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Qin, L., Gupta, A., Upadhyay, S., He, L., Choi, Y., and Faruqui, M. (2021). Timedial: Temporal commonsense reasoning in dialog.

Qiu, Y., Zhao, Z., Ziser, Y., Korhonen, A., Ponti, E. M., and Cohen, S. B. (2023). Are large language models temporally grounded?

Rahimi, A. and Jain, S. (2022). Testing the effectiveness of saliency-based explainability in nlp using randomized survey-based experiments. *ArXiv*, abs/2211.15351.

Ray Choudhury, S., Rogers, A., and Augenstein, I. (2022). Machine reading, fast and slow: When do models "understand" language? In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier.

Rogers, A., Karpinska, M., Gupta, A., Lialin, V., Smelkov, G., and Rumshisky, A. (2022). Narrativetime: Dense temporal annotation on a timeline.

Rojat, T., Puget, R., Filliat, D., Ser, J. D., Gelin, R., and Díaz-Rodríguez, N. (2021). Explainable artificial intelligence (xai) on timeseries data: A survey.

Saeed, W. and Omlin, C. (2021). Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities.

Saini, A. and Prasad, R. (2022). Select wisely and explain: Active learning and probabilistic local post-hoc explainability.

Sanampudi, S. and G.Vijaya, K. (2010). Temporal reasoning in natural language processing: A survey. *International Journal of Computer Applications*, 1.

Schuff, H., Jacovi, A., Adel, H., Goldberg, Y., and Vu, N. T. (2022). Human interpretation of saliency-based explanation over text. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM.

Setzer, A. (2002). Temporal information in newswire articles : an annotation scheme and corpus study.

Shi, X., Xue, S., Wang, K., Zhou, F., Zhang, J. Y., Zhou, J., Tan, C., and Mei, H. (2023). Language models can improve event prediction by few-shot abductive reasoning.

Spreitzer, N., Haned, H., and van der Linden, I. (2022). Evaluating the practicality of counterfactual explanations. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

Sugawara, S., Stenetorp, P., Inui, K., and Aizawa, A. (2019). Assessing the benchmarking capacity of machine reading comprehension datasets.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks.

Tai, Bentley, X. S. F. C.-M. . M. (2024). An examination of the use of large language models to aid analysis of textual data.

Tan, Q., Ng, H. T., and Bing, L. (2023a). Towards benchmarking and improving the temporal reasoning capability of large language models.

Tan, Q., Ng, H. T., and Bing, L. (2023b). Towards benchmarking and improving the temporal reasoning capability of large language models.

Tan, Q., Ng, H. T., and Bing, L. (2023c). Towards benchmarking and improving the temporal reasoning capability of large language models.

ter Meulen, A. G. B. (1997). *Representing Time in Natural Language: The Dynamic Interpretation of Tense and Aspect*. MIT Press.

Uzzaman, N. (2012). *Interpreting the temporal aspects of language*. PhD thesis, USA. AAI3543329.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013a). SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In Manandhar, S. and Yuret, D., editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013b). SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In Manandhar,

S. and Yuret, D., editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007a). SemEval-2007 task 15: TempEval temporal relation identification. In Agirre, E., Màrquez, L., and Wicentowski, R., editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007b). SemEval-2007 task 15: TempEval temporal relation identification. In Agirre, E., Màrquez, L., and Wicentowski, R., editors, *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.

Verhagen, M., Saurí, R., Caselli, T., and Pustejovsky, J. (2010). SemEval-2010 task 13: TempEval-2. In Erk, K. and Strapparava, C., editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.

Wang, Y. and Zhao, Y. (2023). Tram: Benchmarking temporal reasoning for large language models.

Wenzel, G. and Jatowt, A. (2023). An overview of temporal commonsense reasoning and acquisition.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., and Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks.

Xiong, S., Payani, A., Kompella, R., and Fekri, F. (2024). Large language models can learn temporal reasoning.

Yang, Wei, W. e. a. (2023). Survey on explainable ai: From approaches, limitations and applications aspects.

Yanko, G., Pariente, S., and Bar, K. (2023). Temporal relation classification in Hebrew. In Park, J. C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., and Krisnadhi, A. A., editors, *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 261–267, Nusa Dua, Bali. Association for Computational Linguistics.

Yuan, C., Xie, Q., and Ananiadou, S. (2023). Zero-shot temporal relation extraction with ChatGPT. In Demner-fushman, D., Ananiadou, S., and Cohen, K., editors, *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. (2023a). Explainability for large language models: A survey.

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. (2023b). Explainability for large language models: A survey.

Zhong, R., Ghosh, D., Klein, D., and Steinhardt, J. (2021). Are larger pretrained language models uniformly better? comparing performance at the instance level. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3813–3827, Online. Association for Computational Linguistics.

Zhou, B., Khashabi, D., Ning, Q., and Roth, D. (2019). "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Zhou, B., Richardson, K., Ning, Q., Khot, T., Sabharwal, A., and Roth, D. (2021). Temporal reasoning on implicit events from distant supervision. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371, Online. Association for Computational Linguistics.

Appendix A

Each encoder model is fine-tuned for the task of temporal relation classification using the architectural and tokenisation strategies presented by Yanko et al. (2023) and Baldini Soares et al. (2019).

All models are fine-tuned for the duration of 10 epochs with a batch-size of 8, using AdamW optimizer (batch-size was lowered to 4 for RoBERTA-large, given constrains on the available GPU). The learning rate ranged from 3e-05 for BERT-base and BERT-large to 2e-05 for RoBERTa-base and RoBERTa-large.

All base models were sourced from the Hugging Face Transformers library.

Appendix B

Yanko et al. (2023) proposes a "related F1" metric, to account for the complexities with the VAGUE class. The proposed evaluation metric does not account for the mistakes of non-VAGUE predictions on VAGUE samples. This significantly impacts the analysis, as the VAGUE class poses specific challenges. The paper argues that VAGUE inherently incorporates both temporal directions (BEFORE and AFTER), and therefore errors in this class can be discarded to some degree.

	Relaxed F1 M/avg	Relaxed F1 W/avg	EM
RoBERTa $_{\text{large}}$	0.67	0.80	0.78
RoBERTa $_{\text{base}}$	0.69	0.80	0.78
BERT $_{\text{large-uncased}}$	0.64	0.77	0.74
BERT $_{\text{base-cased}}$	0.63	0.77	0.76

Table 8: Performance evaluation of different models from the BERT-family on MATRES (Ning et al., 2018a) dataset, using the "relaxed" F1 metric proposed by Yanko et al. (2023).