computation

MDPI

Article

Arabic Temporal Common Sense Understanding

Reem Alqifari [1,2,*], Hend Al-Khalifa [1,†] and Simon O'Keefe [2,†]

[1] College of Computer and Information Sciences, King Saud University, Riyadh 11421, Saudi Arabia; hendk@ksu.edu.sa
[2] Department of Computer Science, University of York, York YO10 5GH, UK; simon.okeefe@york.ac.uk
* Correspondence: ra1003@york.ac.uk or ragifary@ksu.edu.sa
† These authors contributed equally to this work.

Abstract: Natural language understanding (NLU) includes temporal text understanding, which can be complex and encompasses temporal common sense understanding. There are many challenges in comprehending common sense within a text. Currently, there is a limited number of datasets containing temporal common sense in English and there is an absence of such datasets specifically for the Arabic language. In this study, an Arabic dataset was constructed based on an available English dataset. This dataset is considered a valuable resource for the Arabic community. Consequently, different multilingual pre-trained language models (PLMs) were applied to both the English and new Arabic datasets. Based on this, the effectiveness of these models in Arabic and English is compared and discussed. After analyzing the errors, a new categorization of errors was proposed. Finally, the ability of the PLMs to understand the input text and predict temporal features was evaluated. Through this detailed categorization of errors and classification of temporal elements, this study establishes a comprehensive framework aimed at clarifying the specific challenges encountered by PLMs in temporal common sense understanding (TCU). This methodology underscores the urgent need for further research on PLMs' capabilities for TCU tasks.

Keywords: common sense; temporal understanding; Arabic temporal understanding; natural language understanding; reading comprehension; transformers; transfer learning

Academic Editors: Filippo Palombi and Khaled Shaalan

Received: 29 October 2024
Revised: 17 December 2024
Accepted: 23 December 2024
Published: 28 December 2024

Citation: Alqifari, R.; Al-Khalifa, H.; O'Keefe, S. Arabic Temporal Common Sense Understanding. *Computation* **2025**, *13*, 5. https://doi.org/10.3390/computation13010005

1. Introduction

Temporal text refers to a text that contains temporal features (time, events, and relations). The presence of these features can be explicit or implicit, making the task more challenging. Moreover, comprehending temporal common sense is crucial for understanding the inherent temporal aspects of a text. The term "common sense" can be defined as "the basic level of practical knowledge and reasoning concerning everyday situations and events that are commonly shared among most people" [1]. Although humans possess this ability as part of their cognitive intelligence, it is challenging for machines to acquire it. For example, understanding temporal common sense enables people to understand the order in which events occur. Humans innately understand that certain events precede others, such as falling ill before dying. However, machines face difficulties in acquiring this understanding, as they require complex algorithms and programming to infer the chronological sequence of events. In addition to event ordering, temporal aspects include the event duration. Humans find it relatively easy to predict the duration of activities, such as eating, opening a door, or walking. However, the limited datasets about the temporal understanding or extraction that are currently available primarily involve unusual events

or unusual durations, posing a significant challenge for machines to predict the duration of routine events. There are more examples, including the frequency of an event; humans know that lunch happens once per day. Additional temporal aspects will be discussed later in this study, including the typical time that events occupy and the situation of an event. Therefore, the task of temporal common sense understanding (TCU) is complex. Figure 1 illustrates an example of a TCU challenge and how a model can fail to validate the correct answer.

Context	Question	Candidate answer	Label	Prediction
She added a special growing mix from the garden store to make the soil better.	How long did it take to add growing mix into the garden?	a couple of weeks	No ✓	Yes ✗
		30 minutes	Yes ✓	Yes ✓
	What happened after the growing mix was added?	the crops grew worse	No ✓	Yes ✗
		the plants grow better	Yes ✓	Yes ✓

Figure 1. Example of a TCU challenge showing a scenario where the model fails to validate the correct answer. The table highlights the scenario description, posed question, provided candidate answers, the correct label (marked with a ✓), and the model's incorrect prediction (marked with a ✗). This example illustrates the limitations of the model's temporal commonsense reasoning, emphasizing the need for better training or enhanced datasets tailored for temporal understanding.

Although progress has been made in TCU for English, little to no attention has been given to Arabic, a language spoken by over 400 million people globally and deeply rooted in diverse linguistic and cultural traditions. This study aims to address this gap by focusing on the development and evaluation of Arabic TCU, marking the first systematic exploration of this topic. Focusing on Arabic is essential not only due to the absence of existing datasets but also because of the unique linguistic and cultural challenges inherent to the language. Arabic is characterized by its rich morphological complexity, highly flexible syntactic structures, and distinctive temporal expressions, which differ significantly from those in English. These features introduce nuances that make temporal reasoning in Arabic both challenging and unique. These challenges will be discussed in greater detail in the subsection on challenges.

The scope of this research is TCU, and it is evaluated as textual reading comprehension by employing a multiple-choice format. Using the contextual information provided, a multiple-choice reading comprehension (MRC) system is responsible for selecting an appropriate response from a range of possible answers. To satisfy the requirements of the MRC task, which entails choosing the correct answer from a range of candidate alternatives, the suggested model must determine which answer is correct. In this study and based on the dataset, the model is asked to validate the plausibility of the answer, and each question might have more than one plausible answer. The effectiveness of deep learning in comprehending temporal common sense has been measured using a variety of models.

The model will take three inputs (context, questions, and answers), and it should learn to predict whether this answer is plausible. This model uses examples of textual training, where c is a context that is a passage of text, q is a question relevant to the context c, and a is an answer to question q. The model aims to learn a predictor l, which takes a context c, a corresponding question q, and a candidate answer a as inputs and predicts a score that will be high if the answer is likely plausible and low otherwise. This predictor model can be formulated with the following formula:

$$Score_i = l(\{(c_i, q_i, a_i)\})_{i=1}^{n} \in [0, 1] \quad (1)$$

This formula is applied so that the model can be applied to various MRC tasks with multiple correct answers.

1.1. Challenges

The challenges will be classified into three main categories.

1. **Implicit Temporal Features:** Temporal reasoning is complicated because some events are vague. Therefore, the task of extracting and annotating the temporal features is difficult [2]. For example, the sentence "She visited her friend after finishing her work" is ambiguous because the word "after" does not state whether the visit actually happened immediately after, a few hours later, or even the next day. According to that, the uncertainty about the precise timing of events makes it challenging for a model to accurately process temporal information.
   Implicit temporal features are another challenge. For instance, the temporal sequence in the sentence "Sara finished her breakfast and left for school" must be inferred because it suggests—without explicitly stating—that Sara left for school soon after finishing her breakfast. Similarly, the statement "He often travels for work" suggests a regularity of events without providing information about how frequently the travels take place. The process of temporal reasoning is made more difficult by these implicit temporal cues, which force models to infer the frequency and sequence of events from the context.
   In the MC-TACO dataset [3], if a question concerns an event's duration, all candidate responses belong to a duration type. The challenge is how to validate whether there is a logical duration for this event. Because each candidate's answer has a different duration, categorizing the answers based on the temporal type of the question—for instance, duration—will not eliminate answers that do not fit into the category. Thus, the model should acquire temporal common sense knowledge. For example, the model should know that 30 s would be an illogical illness duration—that is, in this case, 30 s would be a valid duration but not a logical answer. Acquiring this knowledge is expensive and difficult.
2. **Limited Data:** While there are a few datasets available in English, there is currently no dataset specifically designed for TCU in Arabic. One of the most widely used English datasets is MC-TACO, which is designed to evaluate models on TCU. MC-TACO is small, lacks a specific training split, and consists of only evaluation and test sets. In addition, the evaluation set is quite small and contains only 3783 question-answer pairs. Moreover, to the best of our knowledge, there is no dataset in English that is designed to cover all temporal features except MC-TACO. This scarcity of datasets significantly affects the development of models for TCU.
3. **Lack of Knowledge:** According to existing research, current language models lag behind human performance in the task of common sense understanding. For example, this is evident from the MC-TACO leaderboard. Numerous studies have shown that this performance gap can be overcome by relying on external sources that encapsulate common sense knowledge [1,4–6]. For instance, as previously discussed, temporal reasoning involves understanding sequences of events, durations, and implicit time-related features, which are often not fully captured by existing datasets [5,6]. As a result, models struggle to make accurate predictions. Therefore, insufficient data restrict the improvement of these models.
   Existing models still struggle to understand the varying lengths of different events, as the duration of a verb describing an event can change depending on the context. For example, regarding the duration of the verb "taking", the act of "taking a vacation" generally takes longer than "taking a shower". The latter usually lasts for only a

few minutes, whereas the former can last for several days or even weeks [3]. To address this issue, there should be a source of knowledge to help models accurately capture this temporal context. The existing corpus that can be used for this purpose is skewed towards uncommon or unexpected event durations and rare events [3]. For example, the duration of "opening a door" is not mentioned unless it is longer than usual. Determining the duration of various events manually is expensive and time-consuming. Addressing this gap requires the construction of comprehensive knowledge bases (KBs) specifically designed for temporal information or, alternatively, developing more advanced models and algorithms that can learn and infer temporal common sense from the limited data available.

Challenges of Arabic Temporal Text Understanding

The difference between how Arabic and English express temporal information can make it challenging to compare temporal common sense understanding in the two languages. The Arabic language is known for its richness and complexity. Below are some of the challenges that machines may face in understanding Arabic temporal text.

- Arabic is a complex language in which diacritics represent short vowels, but in MSA, they are often omitted. This lack of diacritics causes numerous ambiguities [7,8]. For instance, the same word "ذهب" without diacritics can have these two different meanings: "gold" if it is diacritized "ذَهْب" or go–went if it is diacritized "ذَهَب" [8]. This issue, which leads to ambiguity, is not present in English, as English does not use diacritics.
- Additionally, an Arabic date can be represented using the Gregorian calendar, the Hijri calendar, or both simultaneously. The Hijri calendar, also known as the Islamic calendar, is a lunar calendar that includes 12 months in a year with either 354 or 355 days. There are various methods of representing the Gregorian month names in Arabic, including using phonetically correct English or Arabic names [9]. For example, "January" can be written as "يناير" (Janāyer) or phonetically as "جانيواري" (Jānyuwārī).
- Another challenge arises from the dual usage of Hijri month names as personal names [9]. The names Rajab, Shaaban, and Ramadan can refer to either a month or a person. Additionally, Eid, which is an Islamic holiday, can also refer to a person's name. For example, in the provided sentence, the term "رمضان" is open to interpretation, as it can denote either an individual's name or the Islamic month of Ramadan: "The family is happy with the arrival of Ramadan" الأسرة سعيدة بدخول رمضان.
- Another difference between Arabic and English is the use of temporal adverbs. In this aspect, Arabic has a wider variety of temporal adverbs than English. For example, the Arabic adverb "قبل" ("before") can refer to events that happened before the present moment, events that occurred before a specific time, or events that happened before another event. For instance, see the following examples:
  - ذهبت إلى المدرسة قبل الساعة الثامنة ("I went to school before 8 o'clock");
  - أنهى واجبه قبل أن يصل والده ("He finished his homework before his father arrived");
  - كنت أعمل في الشركة قبل ثلاث سنوات ("I was working at the company three years ago");
  - قرأت هذا الكتاب قبل عدة أشهر ("I read this book several months ago").

  While the English adverb "before" can also order two events in the past or present, such as "I went to school before 8 o'clock" and "He finished his homework before his father arrived", it does not encapsulate all of the nuances and contexts that "قبل" can

in Arabic. For example, "before" does not express time periods without additional context as naturally, such as "I was working at the company three years ago" or "I read this book several months ago", whereas Arabic can use "قبل" directly to convey these time frames.

Allen's temporal relations [10] offer a formal structure for understanding temporal nuances in natural language. For example, the "Before" relation in Allen's framework corresponds to sentences like "I went to school before 8 o'clock", which depict a specific order of events. In Arabic, the adverb "قبل" adds further complexity as it can describe a wide range of temporal contexts, including intervals spanning past periods. This demonstrates the depth and flexibility of temporal expressions in Arabic, highlighting the need for more sophisticated natural language processing (NLP) approaches to accurately capture and interpret such temporal relationships across different languages.

Although English also shows ambiguity, it appears in a different way because of its sequential morphology and spelling patterns. An example of ambiguity in English is that identical words can possess multiple meanings.

1.2. List of Contributions

This study makes several significant contributions to the field of temporal common sense understanding (TCU) using deep learning models.

1. Construction of an Arabic TCU Dataset: An Arabic dataset is constructed to serve in TCU tasks. This construction will be highly impactful for the Arabic community, and it addresses the absence of such a resource. The dataset is based on an existing English dataset. The dataset and the code are available from the corresponding author upon request.
2. Benchmarking for Temporal Understanding: To evaluate the ability of PLMs to understand temporal features, a benchmark for temporal understanding was established.
3. Applying Multilingual Pre-Trained Language Models (PLMs): The effectiveness of different multilingual PLMs on MC-TACO (the original English dataset) and the Arabic dataset was examined. Each model was assessed in terms of each temporal aspect.
4. Analyzing Errors: By analyzing the errors, a new classification is suggested to identify specific issues, which will help improve the understanding of PLMs.

The rest of this article is structured as follows: Section 2 reviews and discusses related works. The construction of an Arabic dataset for the TCU task is presented in Section 3. Section 3 also provides a detailed overview of the dataset statistics. Section 4 presents the evaluation metrics used in this study. In Section 5, the applied PLMs for Arabic TCU are explored. A detailed analysis of the results of applying PLMs is presented in Section 6. Consequently, a methodology for evaluating the effectiveness of multilingual PLMs by analyzing and categorizing errors is proposed in Section 7. To understand the challenges affecting PLMs' performance in TCU, a benchmark for assessing PLMs in temporal classification is presented in Section 8. Finally, Section 9 provides the conclusion of the study and suggests potential directions for future research.

2. Related Works

The first application of a PLM to the MC-TACO dataset was in 2019, when Zhou et al. [3] applied BERT as a baseline model. The performance of BERT [11] fell significantly below human performance levels [3]. This led to the application of unit normalization for the inputs as a preprocessing step, resulting in slight improvements. Subsequently,

RoBERTa [12] was applied without unit normalization or any preprocessing, and its performance was better than that of BERT with unit normalization [3].

Duration normalization was also proposed as a preprocessing step to improve the results. In 2020, Kaddari et al. [13] applied duration normalization with T5 [14]. The result of the proposed model outperformed other models without duration normalization, although the improvement was marginal compared with T5 without the preprocessing step. State-of-the-art performance was achieved by applying DeBERTa-Large [15] without any preprocessing. It outperformed all models by a significant margin. This indicates that PLMs may be effective even without rule-based processing. Rule-based preprocessing, such as unit or duration normalization, is language-dependent, prone to errors, and labor-intensive due to manual coding.

A few studies have attempted to build specific language models for the TCU task, as proposed after the observed shortcomings of existing PLMs.

- TACOLM: This study highlights the inadequacies of PLMs in addressing TCU tasks, particularly in terms of their failure to recognize and learn from temporal dimensions. The study proposed an additional pre-training step designed to enrich models with time-related data by using two methodologies to construct the dataset for this enhanced pre-training phase [6].
- ECONET: This study aimed to develop temporal language models to improve event ordering tasks. Inspired by ELECTRA [16], this approach leveraged a targeted masking strategy to focus the model's learning on temporal aspects [17].
- A Third Language Model for TCU: This model also utilized a continual training approach, introducing a different target masking strategy and employing various time-related datasets. Unlike TACOLM and ECONET, this study did not construct its dataset but used pre-existing time-related datasets, offering comprehensive coverage of all temporal dimensions [18].

Virgo et al. [5] demonstrated that recent PLMs have yet to reach human performance levels in an event duration task. The limited training data, which cover only a finite number of events and their attributes, highlight the need to incorporate external event duration information to enhance effectiveness. A new QA dataset for event duration was constructed from an existing dataset and used for intermediate tasks in an adaptive fine-tuning approach. While Kimuar et al. [18] used the existing dataset as is, Virgo et al. [5] focused solely on event duration, whereas the authors of [18] studied all aspects of temporal understanding.

Several adaptive fine-tuning techniques were explored for the English MC-TACO dataset [19,20] and adversarial fine-tuning [21]. Despite exploring alternative training methodologies and constructing specialized datasets, the outcomes from these studies still fall short of the performance levels achieved by more advanced PLMs, such as DeBERTa [15]. According to the leaderboard (https://leaderboard.allenai.org/mctaco/submissions/public, accessed on 29 October 2024), these techniques perform worse than the DeBERTa-Large model [15], which uses the standard fine-tuning paradigm.

Although all suggested techniques have been surpassed by DeBERTa-v3, DeBERTa's performance still falls significantly short of human performance on the same task. This gap emphasizes the complexity of temporal reasoning in NLU and the ongoing need for research to refine and enhance the capabilities of language models in this critical area.

3. Dataset

TCU is an essential part of the larger field of natural language common sense comprehension. Despite the importance of TCU, the availability of resources dedicated to this aspect in English is limited. Remarkably, there are no datasets in Arabic that are tailored

to this specific domain. Currently, the only Arabic dataset that addresses common sense understanding is essentially an English translation that concentrates on common sense validation [22]. This disparity severely restricts the ability to assess the effectiveness of transformer-based models in understanding Arabic temporal common sense. The creation of a dataset is essential for the goals of this study, which include evaluating and enhancing the effectiveness of transformer-based models in Arabic TCU.

Despite the challenges of its construction, such a dataset promises to be a valuable addition to Arabic resources, allowing for a more sophisticated and culturally appropriate understanding of temporal common sense in Arabic. The construction of an Arabic TCU dataset is considered to have a crucial impact not only on the field of TCU but also on the Arabic community. This dataset would also greatly advance Arabic natural language understanding (NLU) by serving as a foundational resource for further research, in addition to the main goal, which is to evaluate transformer-based models in the Arabic linguistic context.

Dataset Construction

The construction of a dataset from scratch is particularly resource-intensive. This challenge is compounded in the context of Arabic, where there is a conspicuous absence of time-related datasets and a general scarcity of resources. Given these constraints, the decision to adapt an existing dataset from English to Arabic was motivated by both the practicality and the unique requirements of the focus of this study.

The MC-TACO dataset [3] was selected for translation into Arabic because of several key factors that align with the research objectives. First, MC-TACO is recognized, to the best of our knowledge, as the only dataset that encompasses a wide range of temporal characteristics, making it exceptionally relevant for our study of TCU. Second, the dataset's straightforward structure and use of simple sentences render it particularly amenable to translation, ensuring the preservation of semantic integrity during this process.

MC-TACO was designed as a multiple-choice reading comprehension (MRC) task. The input of the model from the dataset consists of three components: an abstract or context, a question, and a corresponding answer. The model requires the output of a prediction score based on a judgment of the plausibility value of the answer. The score should be close to one if the candidate answer is valid. The relatively concise nature of the information provided in the dataset, often encapsulated in three sentences, makes it feasible to employ a translation tool for the initial construction process. Google Translate was utilized for this purpose, with subsequent translations being subjected to a thorough review by two native Arabic speakers specialized in proofreading to ensure accuracy and natural language use. The reviewers, who examined all the inputs individually, were from different Arabic countries—specifically, Saudi Arabia and Morocco—as cultural differences might affect the understanding of the translated results. Finally, the overall results were reviewed to ensure consistency and accuracy.

The dataset encompasses approximately 13K question–answer pairs spanning five temporal dimensions, thereby offering a rich resource for exploring various aspects of temporal reasoning. The temporal dimensions included in MC-TACO are explained below.

1. Event duration: How long does an event last?
2. Temporal ordering: Typical order of events.
3. Typical time: When did an event occur?
4. Frequency: How often do events occur?
5. Stationarity: Is a state maintained in the long term or indefinitely?

Table 1 presents statistical information for both the English and Arabic versions of the dataset. Table 2 presents statistics for the temporal features. Furthermore, Figure 2 illustrates the distribution of question–answer pairs across different temporal aspects. The

dataset predominantly consists of question–answer pairs related to event duration, with frequency being the second most common aspect. On the other hand, the coverage of the stationarity feature is notably low, comprising only 870 pairs (7%).

Table 1. Dataset statistics.

Measures	Arabic	English
Number of unique questions	1893	1893
Number of unique question–answer pairs	13,225	13,225
Avg. context length	15.2	17.8
Avg. question length	6.5	8.2
Avg. answer length	3	3.3

Table 2. Temporal category statistics.

Category	Number of Unique Contexts	Number Unique Questions	Avg. Number of Candidates
Event Duration	135	440	9.4
Event Ordering	26	370	5.4
Frequency	229	433	8.5
Typical Time	43	371	6.8
Stationarity	73	279	3.1

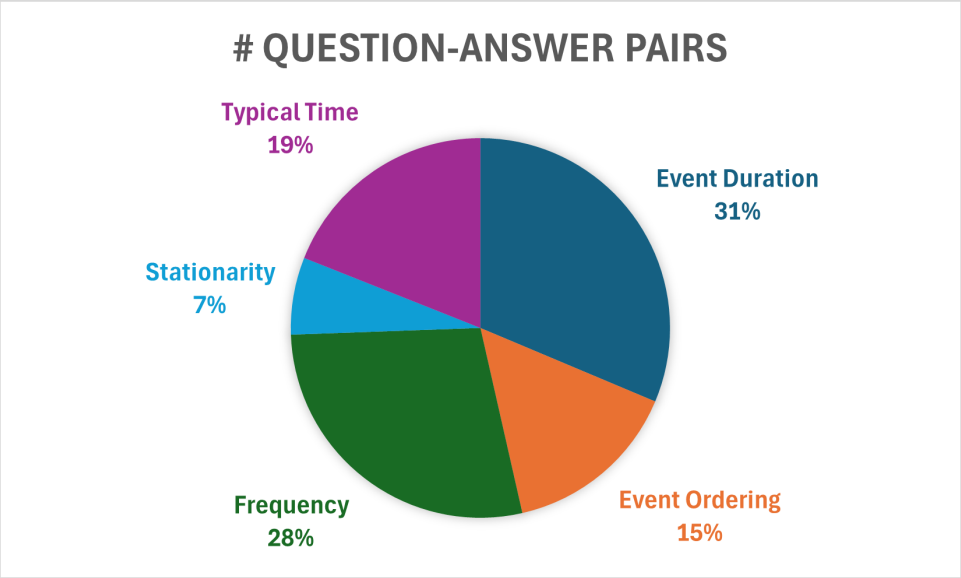

Figure 2. Percentage of the unique question–answer pairs in each temporal category.

A sample of the dataset is presented in Figure 3. This figure provides a comprehensive overview of different temporal categories, illustrating an example of each one. Additionally, this figure includes a question along with its corresponding set of answers. The correct answers are highlighted in bold.

	Arabic	English
Event Ordering	الفقرة: تدور أحداث فيلم لانق حول رجل يتجول في الشارع ليلة بعد ليلة . السؤال: ماذا يحدث بعد العثور على الرجل وهو يتجول في الشوارع ليلاً؟ • **أوقفت الشرطة الرجل عندما وجدته يتجول في الشوارع ليلاً** • **تم استجوابه من قبل الشرطة** • **تم استجوابه من قبل السلطات** • يأكل وجبة غداء	Context: Lang centers on a man who roams the street night after night. Question: What happens after the man is found to be roaming the streets at night? • **The man is stopped by police when he is found to be roaming the streets at night** • **He is questioned by police** • **He is questioned by authorities** • He eats lunch
Event Duration	الفقرة: توفي والد دورر عام 1502، وتوفيت والدته عام 1513. السؤال: كم من الوقت كانت والدته مريضة؟ • 30 ثانية • 6 قرون • 90 سنة • **6 أشهر** • **سنتين**	Context: Durer's father died in 1502, and his mother died in 1513. Question: How long was his mother ill? • 30 seconds • Six centuries • 90 years • **6 months** • **2 years**
Frequency	الفقرة: ذهب تومي وسوزي (أخ وأخت) إلى الملعب بعد ظهر أحد الأيام مع أمهما وأبيهما، جان ودين. السؤال: كم مرة ذهبوا إلى الملعب؟ • يذهبون للملعب مرتين في الليلة • مرتين في الدقيقة • **مرتان شهرياً** • **مرة في الأسبوع** • **يذهبون إلى الملعب مرة كل بضعة أيام**	Context: Tommy and Suzy (brother and sister) went to the playground one afternoon with their mom and dad, Jan and Dean. Question: How often did they go to the playground? • They go to the playground twice a night • Twice a minute • **Twice a month** • **Once a week** • **They go to the playground once every few days**
Stationarity	الفقرة: لقد كانت القضايا التي تعاملت معها على مر السنين تتعلق بمساعدة الناس في الحفاظ على أساسيات الحياة - المنزل والرعاية الصحية والوظائف والأسرة. السؤال: هل ما زالوا يساعدون الناس؟ • لا توقفوا قبل أسبوع • لا توقفوا بعد دقيقة • **أجل إنهم كذلك**	Context: The issues I've dealt with through the years have been on the side of helping people maintain the basics of life - home, health care, jobs and family Question: Are they still helping people? • No, they stopped before a week • No, they stopped after a minute • **Yes, they are**
Typical Time	الفقرة: على سبيل المثال، ماذا لو وضعت كعكة في الفرن وتركتها لفترة طويلة؟ السؤال: في أي وقت ستضع الكعكة في الفرن؟ • 3 صباحاً • 12 صباحاً • **2 ظهراً** • **3 مساءاً** • **12 ظهراً**	Context: For example, what if you place a cake in the oven and you leave it in too long? Question: What time would you put the cake in the oven? • 3:00 A.M. • 12 A.M. • **2 P.M.** • **3 P.M.** • **12 P.M.**

Figure 3. Sample of the dataset. Each row targets one temporal aspect from the five aspects covered by the original dataset. An example context for each aspect is provided from both the English and Arabic datasets. The English column is from the MC-TACO dataset and includes five different contexts, each representing one aspect. For each context, the question is provided along with all candidate answers, with the correct answers in bold. Note that there may be more than one correct answer for a question, and the number of answers for each question varies. The Arabic column is from the translated dataset.

4. Evaluations

Although the dataset can be viewed as a binary classification task, where accuracy is a commonly used metric, it may not be the most appropriate metric in this case. The distribution of labels for the candidate's responses is approximately one "no" to two

"yesses", implying that a high level of accuracy (or a low error rate) can be achieved even by a model without real skill that simply predicts the majority class. Consequently, the accuracy can be a misleading metric for this type of dataset.

To address this, we adopted the F1 score and Exact Match metrics based on recommendations in prior research by Zhou et al. [3]. The F1 score was chosen for its ability to balance precision and recall, which is critical for assessing nuanced tasks such as TCU, where errors often involve partial correctness. This ensures that the model's predictions are evaluated not just for their frequency of correctness but also for their completeness and consistency. Meanwhile, Exact Match serves as a stricter metric, providing insight into models' ability to produce fully correct outputs.

The F-measure is a single metric that trades precision for recall. This factor is the weighted harmonic mean of the precision and recall, where the weight is denoted by the variable β. The default balanced F-measure, where $\beta = 1$, is commonly written as $F1$, which is short for $F_{\beta=1}$.

$$F = \frac{(1+\beta^2) \times Precision \times Recall}{(\beta^2 \times Precision) + Recall} \tag{2}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Exact Match (EM) is a strict version of accuracy in which all labels must match exactly for the sample to be correctly classified. For MC-TACO, the model must correctly predict all answers to each question to be considered a correct prediction.

$$\text{EM} = \frac{\text{Total Number of Questions that are Predicted Correctly}}{\text{Total Number of Questions}} \tag{4}$$

5. Models

Various experiments were conducted using different PLMs. Multilingual PLMs and Arabic versions of BERT were applied to understand the Arabic dataset. Subsequently, a detailed comparison and analysis of the model results are presented. Two Arabic versions of BERT were adopted: AraBERTv2 and CAMeLBERT. AraBERTv2 is the latest version of AraBERT that was initially introduced by Antoun et al. [23]. This model was selected over other Arabic versions of BERT due to its superior performance, as evidenced by the model card on Hugging Face (https://huggingface.co/aubmindlab/bert-base-arabert, accessed on 1 May 2024) and research conducted by Alammary et al. [24]. This study involved text classification specifically for the Arabic language. In this study, AraBERTv2 exhibited better outcomes than XLM-RoBERTa. The CAMeLBERT model [25] (https://huggingface.co/CAMeL-Lab, accessed on 1 May 2024) was not included in the analysis conducted in [24]. It would be valuable to compare this model with the current leading Arabic BERT model. Furthermore, the performance of AraBERTv02 is better than that of CAMelBERT according to [25]. However, CAMeLBERT was the second-best model among all Arabic versions of BERT based on research conducted by CAMeL Lab [25]. Therefore, it is worthwhile to compare these two models for this task. CAMeLBERT has different versions. CAMeLBERT-msa was selected among all others because the target dataset was written in MSA Arabic. Notably, AraBERT and CAMeLBERT are different from multilingual models because they are tailored to Arabic. AraBERT, CAMelBERT, and multilingual BERT all have the same architecture because they are derived from the original BERT with some modifications. According to Inoue et al. [25], the pre-training data size may not be an important factor in fine-tuning performance.

In this study, multilingual BERT was selected because BERT was the baseline model for the original dataset and numerous versions of BERT have been designed specifically

for the Arabic language. Furthermore, mDeBERTav3 and XLM-RoBERTa were specifically chosen because of the effectiveness of their original models on the English dataset.

6. Results

This section presents the results of applying multilingual PLMs to Arabic TCU. To the best of our knowledge, this is the first study to explore this area. AraBERT and CAMeLBERT were pre-trained only on the Arabic dataset. Therefore, the Arabic versions of BERT were expected to outperform their multilingual counterpart and multilingual PLMs. The results of these experiments are presented in Table 3.

To assess the stability and variability of the system, additional runs were conducted, each with a different random seed. Three runs were conducted. Each run was independent, and its random processes, including data shuffling and GPU initialization, were influenced by its specific seed. From these three runs, the performance metrics were observed, and the standard error was calculated to deduce how much the performance varied with the change in seeds. Finally, the reported performance metrics were based on running the system with a default random seed, which was equal to 42.

Table 3. Results of applying PLMs on the Arabic dataset.

Model	F1	EM
mBERT	58.12	28
Arabert-v02	64.46	34.01
CAMeLBERT-msa	61.76	32.51
XLM-RoBERTa-Large	64.99	36.19
XLM-RoBERTa-base	61.77	31.53
mDeBERTa-v3	**67.98**	**38.66**

Based on the results presented in Table 3, multilingual DeBERTa-v3 achieved the best performance, followed by XLM-Roberta Large. Although AraBERTv02 and CAMeLBERT were trained on Arabic datasets, mDeBERTa-v3 outperformed them significantly. This may have occurred because the target task required common sense reasoning, suggesting that more advanced models such as mDeBERTa-v3 could be necessary, explaining the performance discrepancy. Factors that can be attributed to the superior performance of mDeBERTa-v3 compared with the other models are as follows:

1. The depth of the architecture in XLM-RoBERTa-large is 24 layers—twice the number of layers in mDeBERTa-v3, XLM-RoBERTa base, and mBERT, which all have 12 layers. This difference could indicate that the number of layers might not provide the best performance.
2. Although BERT and XLM-RoBERTa employ self-attention mechanisms, mDeBERTa-v3 may incorporate more advanced attention mechanisms, such as disentangled attention, which is specifically designed to capture precise linguistic dependencies. Consequently, this model can surpass the others in tasks requiring extensive linguistic analysis, such as TCU.
3. This is also evidenced by applying English DeBERTa-v3 Large to MC-TACO, which achieved state-of-the-art results. The success of DeBERTa-v3 in TCU demonstrates the effectiveness of transfer learning, which overcomes the problems of limited datasets and extensive labeling by leveraging the large knowledge base acquired during the pre-training process.

This study compared the performance of AraBERT-v02 and CAMeLBERT, which are both designed for the Arabic language. AraBERT-v02 showed better results than CAMeLBERT, which was possibly due to the former's vocabulary size, which is twice that of CAMeLBERT.

Figure 4 presents the outcomes of the various models when applied to the Arabic dataset. The F1 score is displayed for each temporal aspect, facilitating an assessment of the effectiveness of each model. The figure reveals several significant findings, which are summarized below:

- The strength of all models is the stationarity aspect. All of the models scored above 74 in this aspect. mDEBERTa-v3 and XLM-RoBERTa Large scored the same. Upon analyzing the data, it appears that this particular aspect may be less difficult than other aspects. This is mainly because the majority of the responses for this feature were either yes or no. Furthermore, certain responses are evidently unrelated and are easily dismissed by the models.
- The event duration was the most challenging feature for all models, with a mean discrepancy of 10 units less than the overall F1 score of each model. Due to the challenging nature of this aspect, some studies, including that of Virgo et al. [5], suggested that an external source is required.
- Overall, mDeBERTa-v3 is the most effective model, but it did not outperform all models in all aspects, and mDeBERTa-v3 demonstrated superiority in event duration and frequency.
- AraBERt-v02 and CAMeL-msa demonstrated superior performance to that of all other models in the Typical Time feature. Notably, the overall effectiveness of CAMeL was lower than that of the other models. Thus, it was necessary to identify the distinguishing factor of the typical time feature that made the models trained on Arabic datasets perform better than the other models. This might be because the typical time is closely related to the nature of the culture, making models fully trained on Arabic datasets more effective.
- mBERT had the lowest performance in all categories.

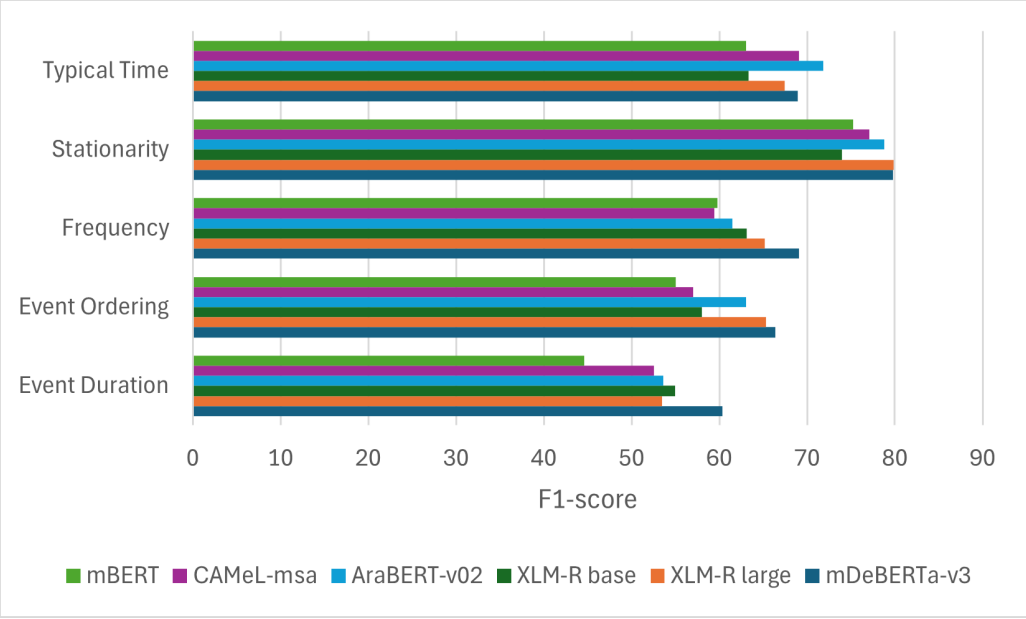

Figure 4. Model results: F1 score for each temporal aspect.

Although the AraBERT and CAMeL models have different performance levels, it is important to analyze the differences in their predictions because they have numerous similarities in their model architecture and pre-training datasets. The distribution of the prediction percentages across the different temporal features is illustrated in Figure 5. The prediction of the probable order of events varied significantly between the two models, with most predictions differing by approximately 30%. Moreover, the percentage of identically incorrect predictions was the highest. The most similar predictions were those of frequency and event duration.

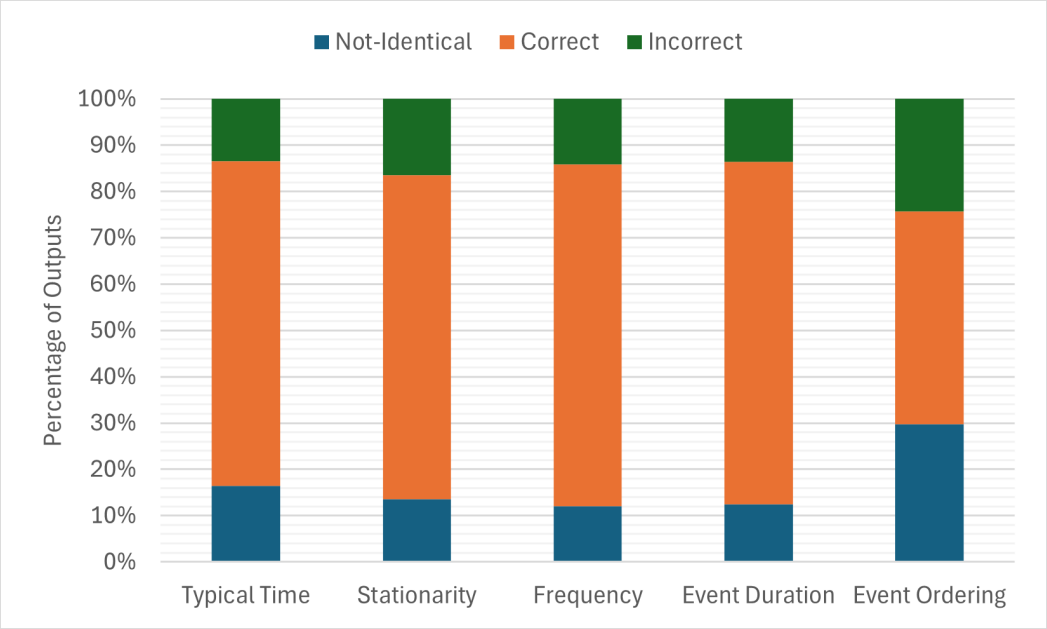

Figure 5. Predictions of AraBERT vs. CAMeLBERT.

Hyperparameter Settings

The hyperparameters used for training and evaluating all models in this study were kept consistent to ensure a fair comparison across the different pre-trained language models. These settings were chosen based on preliminary experiments and followed best practices from related work in temporal reasoning tasks. Table 4 summarizes the hyperparameters used across all models.

Table 4. Hyperparameters used to train and evaluate all models in this study. These hyperparameters were consistent across all models to ensure a fair comparison.

Parameter	Value
Learning Rate	2×10^{-5}
Batch Size	32
Number of Epochs	10
Max Sequence Length	128

These hyperparameters were applied uniformly to all models to ensure that differences in performance were solely attributable to the model architecture and not the configuration.

Computational Environment: All experiments were conducted on Google Colab using an A100 GPU, and the model implementation was based on PyTorch (version 2.4.0) and the Hugging Face Transformers library (version 4.45.0).

7. Error Analysis

The task undertaken in TCU presents significant challenges, as evidenced by the results of the models in the previous section. This led to a manual investigation of the inputs to understand the challenges. Thus, by studying the MC-TACO dataset, certain questions pose difficulties, even for humans, which is substantiated by the dataset's human performance metrics (87%). Consequently, errors encountered in this context can be classified into two main categories: human-challenging errors and linguistic and task complexity errors. Human-challenging errors are instances where both the model and humans struggle to validate the given answers. On the other hand, linguistic and task complexity errors denote situations in which the model fails to grasp the intricacies of language use or validation of answer plausibility. The following sections present each type of error in detail and provide a more detailed analysis.

7.1. Human-Challenging Errors

As previously mentioned, the first type of error, characterized by its complexity and the challenges it poses to both humans and models, can be approached with a degree of acceptance or tolerance compared with the second type. These errors fall into an "uncertain zone", where they are neither fully correct nor entirely incorrect. For instance, from MC-TACO, consider the question related to the duration of an illness based on a provided context: "Dürer's father passed away in 1502, and his mother died in 1513. How long was his mother ill?". One of the candidate answers is that she was ill for 30 years. Even though the gold label for this answer is "no", meaning that the answer is incorrect and considered an unlikely duration for an illness, it is also possible to argue "yes", indicating that a 30-year duration could indeed represent a period of illness. This example highlights the subjective nature of certain inputs, emphasizing the complexity and potential ambiguity inherent in evaluating the validity of the answers provided for the TCU task.

The human performance metric displayed on the leaderboard of MC-TACO—specifically, the F1 score—was 87.1%, and the Exact Match (EM) rate was 75.8%. However, it is important to note that these measures, which were derived from a subset of the dataset, may not fully encapsulate accuracy. Despite this limitation, it can be assumed that human-challenging errors account for approximately 22% of the total error rate, providing insight into the extent of the challenges posed by these types of questions.

Capturing inputs considered challenging for humans within a test set comprising 9442 items is an arduous task. Implementing a voting mechanism across the entire test set could offer insight into the difficulty level of each question. However, the benchmark was derived from a sample of the total data [3], suggesting that a rough sampling of instances falling into this challenging category would be both practical and acceptable. This approach allows for the identification and analysis of particularly complex cases without the need to exhaustively review every item in the dataset, thereby providing a feasible method for gauging the extent of challenging human errors within the dataset.

Additionally, it is important to recognize that some answers may be culturally dependent, meaning that for certain cultures, an event or concept might be considered plausible, whereas for others, it might not. This variability introduces another layer of complexity, classifying instances as human-challenging errors. This cultural dimension underscores the necessity of incorporating a diverse perspective when evaluating answers, as it highlights the subjective nature of understanding and interpreting information. Recognizing the influence of cultural context on what is deemed correct or incorrect is crucial for accurately assessing the scope of human-challenging errors within the dataset. For instance, within the dataset, a question regarding the appropriate time for an interview illustrated the impact of cultural differences. In Saudi culture, interviews can be scheduled on Sundays, a practice

that might differ from norms in other cultures, where the workweek typically begins on Monday. Furthermore, the start time for schools in Saudi Arabia is earlier than that in many other countries, reflecting another aspect of cultural variance. In addition, during Ramadan, eating late at night is very common to accommodate the fasting schedule. This contrasts with the dining habits of cultures that do not observe Ramadan. These examples highlight how cultural contexts significantly influence the interpretation of what constitutes a correct or plausible answer, thereby contributing to the categorization of such instances as human-challenging errors within the dataset.

Some research has focused on the grounding of time expressions as a culturally dependent aspect. One notable study by Shwartz [26] analyzed time expressions across 27 languages, although Arabic was not included. This study aimed to define how the conceptual range of time periods, such as morning and noon, can vary significantly across different cultures. Additionally, it explores the impact of these cultural variations on PLMs. The findings of such studies are crucial because they highlight the challenges of PLMs in accurately understanding and generating context-appropriate responses to time-related queries [26]. These variations in the perception of time can affect a model's ability to provide correct and culturally sensitive answers, underscoring the importance of incorporating diverse cultural understanding into the development and training of language models.

Accordingly, human-challenging errors can have two subcategories: "cultural temporal interpretation" and "subjective event understanding".

In the exploration of PLMs to navigate the complexities of culturally dependent temporal expressions and subjective understandings, it is crucial to examine the specific instances where errors occur. The following examples provide a comprehensive overview of various error types identified in this category, showcasing sample inputs alongside the expected responses and comparing these with the outputs generated by the PLMs. This comparative analysis not only highlights the discrepancies between expected and actual responses but also offers insights into the models' underlying challenges with cultural nuances and subjective interpretations. This examination can shed light on areas of improvement in PLMs for the TCU task. Moreover, creating a dataset that can address this issue has great potential.

Examples

Table 5 presents the various scenarios that can be classified as subjective event understanding. The analysis of each example is discussed in the following list, which is ordered according to the same sequence as the examples in the table.

1. All of the PLMs in English and Arabic predicted "no". This outcome is indicative of a challenge for PLMs and touches upon human cognitive processes. The specificity of the duration—7.5 min—represents an atypical time frame that is not commonly associated with the activity described.
2. The response of all of the PLMs in English and Arabic is "yes". This input could be considered a human-challenging error, as it involves understanding the legislative process and the realistic pace at which laws and initiatives are typically passed, which vary significantly across different jurisdictions and over time.
3. Not all of the PLMs are able to predict the correct label. The scenario involving steam rising from a wet road after a summer rainstorm, with the duration specified as "steam rises for 30 min off a wet road before a summer rainstorm", presents a nuanced challenge that tests both temporal reasoning and contextual understanding. The question is inherently human-challenging, not just because of the temporal aspect—quantifying the duration for which steam rises—but also because of the contextual misunderstanding in the provided response. The mention of steam rising "before" a rainstorm

contradicts the common observation and understanding that steam typically rises "after" rain has fallen and is heated by the warm road, creating a visual phenomenon observed by many.

Table 5. Examples of errors in subjective event understanding.

Context	Question	Answer	Label
It was huge and inefficient, and she should never have spent so many pesos on a toy, but Papa would not let her return it.	How long did she spend at the store buying the toy?	She spent 7.5 min at the store buying the toy	yes
California was first to require smog checks for clean air and to pass anti-tobacco initiatives and bike helmet laws.	How often are such initiatives passed?	One a month	no
Most of us have seen steam rising off a wet road after a summer rainstorm.	How long does steam rise after a summer rainstorm?	Steam rises for 30 min off a wet road before a summer rainstorm.	no

Table 6 presents various scenarios that may be classified as cultural temporal interpretation. The analysis of each example is discussed in the following list, which is ordered according to the same sequence as the examples in the table.

1. All PLMs, in English and Arabic, predicted "no". In fact, this case touches on cultural or geographical dependency. The perception of how often it rains in the summer varies significantly across different regions and climates, which makes this question inherently dependent on the cultural and geographic context. This information is not provided in the context.
2. Labeling certain answers as "no" can be culturally dependent, which touches on broader themes of consumer behavior, store operating hours, and possibly societal norms regarding appropriate times for shopping. The assumption that purchasing a toy at midnight is unusual or incorrect may indeed vary by culture and locale. In some regions or during certain times (such as holidays or special sale events), late-night shopping can be common, while in others, it might be seen as atypical due to differing social norms and operational hours of businesses.
3. Similarly to the previous case, in some cultures and during holidays, baking is possible at 12 a.m.

Table 6. Examples of errors in cultural temporal interpretation.

Context	Question	Answer	Label
Most of us have seen steam rising off a wet road after a summer rainstorm.	How often does it rain in the summer?	A couple times every month	yes
It was huge and inefficient, and she should never have spent so many pesos on a toy, but Papa would not let her return it.	What time did she purchase the toy at the store?	Midnight	no
For example, what if you place a cake in the oven and you leave it in too long?	What time would you put the cake in the oven?	12:00 a.m.	no

7.2. Linguistic and Task Complexity Errors

This category encompasses a broad range of challenges encountered by language models, making it the most diverse category of errors. These errors are predominantly language-dependent, and they are significantly influenced by the specific tokenization methods used during model training. Additionally, linguistic features such as morphology and overall complexity contribute to the difficulties faced by models in processing and understanding language inputs accurately. Finally, validating the likelihood of provided answers can be challenging and requires common sense knowledge.

A key aspect of linguistic complexity errors is their relationship with the structure and nuances of language, including how words are formed and combined. The way a model tokenizes input—breaking down sentences into words, subwords, or characters—affects its ability to understand and generate coherent responses. Morphological complexity, which involves the structure of words and their relationship with one another within a language, further complicates comprehension, especially in languages with rich inflectional systems.

Moreover, these types of errors can often be mitigated through strategic preprocessing steps, such as time normalization (converting time expressions to a standard format) [13] or unit normalization (standardizing measurements) [3]. Additionally, models vary in their ability to interpret numbers, whether presented in word form or as numerals, which can lead to inconsistencies in understanding and answering questions accurately.

Some models demonstrate proficiency in identifying the type of answer required (e.g., a date or a quantity) but falter when it comes to providing the correct specific response. This discrepancy may stem from the models' training datasets, which might not adequately prepare them for the breadth of task complexity that they encounter in real-world applications. This limitation suggests the need for more comprehensive training approaches that better encapsulate the linguistic diversity and complexity inherent in natural languages. Moreover, enhancing the current dataset can assist the model in comprehending this complex task.

Table 7 illustrates a sample of errors in the complexity of the language and tasks. The analysis of each example is discussed in the following list, which is ordered according to the same sequence as the examples in the table.

1. mBERT failed to predict this in both. For the Arabic dataset, XLM-R base, AraBERT, and CAMeL failed, while a human can easily validate the answer as "no". However, this represents a more complex challenge for PLMs. The models must not only process the natural language of the question but also apply logical reasoning and background knowledge to identify the irrationality of the premise that a historical figure could die multiple times. This difficulty arises from the models' reliance on patterns and data within their training corpus, which does not explicitly cover every aspect of common sense or logical reasoning needed to immediately flag the question's premise as impossible.
2. All of the PLMs in English and Arabic predicted "no". The difficulty lies not only in interpreting historical events and their timelines but also in the nuanced understanding of the term "postwar slump" and its impact over time. The term "postwar slump" refers to the economic downturn following a significant conflict—in this case, likely World War I, considering the reference to the 1930s. The incorrect labeling of "decades" as a possible answer by all PLMs could be attributed to different reasons. The model's failure to recognize "decades" as a plausible duration may indicate a gap in understanding the prolonged effects of postwar economic conditions or the specific historical context.
3. All of the PLMs in English and Arabic predicted "no". Correctly interpreting and validating answers regarding the frequency of activities also involves common sense

understanding and world knowledge, such as the typical behaviors of families with young children. PLMs must leverage this broader knowledge to make informed inferences about habitual actions.

Table 7. Linguistic and task complexity errors.

Context	Question	Answer	Label
However, more recently, it has been suggested that it may date from earlier than Abdalonymus' death.	How often did Abdalonymus die?	every two years	no
Setbacks in the 1930s caused by the European postwar slump were only a spur to redouble efforts by diversifying heavy industry into the machine-making, metallurgical, and chemical sectors.	How long did the postwar slump last?	decades	yes
Tommy and Suzy (brother and sister) went to the playground one afternoon with their mom and dad, Jan and Dean.	How often did they go to the playground?	twice a month	yes

Adopting error categorization can enhance our understanding of PLMs' behavior and pinpoint the sources of errors. This insight lays the foundation for future research focused on improving these models in several key areas.

- Domain-Specific Pre-Training: Pre-training models on specialized temporal corpora, such as Arabic news archives or domain-specific datasets, can improve their understanding of temporal expressions across diverse contexts. This approach would enable models to better grasp nuances specific to temporal reasoning in the Arabic language.
- Data Augmentation: Introducing diverse temporal reasoning examples, such as historical timelines, weather forecasts, or event-driven narratives, can broaden models' ability to generalize. By enriching the dataset with varied temporal scenarios, we can enhance a model's robustness and its capacity to handle different forms of temporal reasoning.

8. PLM Evaluation

In this section, after identifying the errors in the previous section, the investigation focuses on determining whether the failure stems from a lack of understanding of temporal features or the inherent challenges of the TCU task. Additionally, a comparison between the two languages is conducted to analyze this aspect.

A novel hypothesis proposes the assessment of PLMs' proficiency in TCU tasks by categorizing inputs into five temporal dimensions. Within this framework, a model's ability to categorize questions correctly suggests that it grasps the temporal features and understands the type of temporal question being asked. However, failing to provide the correct answer implies that the model struggles to apply logical reasoning or lacks the necessary world knowledge to determine whether an answer is plausible within the given temporal context.

Drawing inspiration from traditional question–answer (QA) systems, where question classification is crucial for high performance, based on Huange et al. [27] and Kolomiyets et al. [28], our approach adapts this methodology for a modern context. Unlike QA systems, where classification directly aids in generating answers, here, it serves as a diagnostic tool. This distinction is key, emphasizing that our goal is not to classify for the sake of classification but to deepen our understanding of PLMs' handling of temporal information.

Implementing this strategy requires the utilization of all PLMs applied to the dataset for temporal classification. This foundational step ensures that the models understand the temporal dimension of queries. Following classification, the models' performance on this task was compared with their TCU task performance, focusing on each temporal aspect. This comparison was designed to reveal correlations or discrepancies, providing insight into the models' capabilities and areas that need refinement.

For example, a model proficient in classifying questions but faltering in providing accurate answers might indicate an understanding of temporal concepts but a lack of application in complex scenarios. Such findings highlight the importance of enhancing models' reasoning capabilities and understanding temporal contexts.

Our methodology not only offers a nuanced assessment of PLMs' handling of temporal information but also points to potential improvements. By identifying specific weaknesses, this approach can guide the development of targeted training or fine-tuning strategies, thereby enhancing PLMs' effectiveness in TCU tasks and beyond.

Results and Discussion

The classification task was implemented in both monolingual and cross-lingual settings, examining the models' performance across different languages. In monolingual environments, specifically for Arabic and English, the classification accuracy reached an impressive rate of approximately 97%. This high level of accuracy underscores the effectiveness of PLMs in understanding temporal information in a single-language context.

Figure 6 distinctly showcases the significant gap between classification accuracy and performance on the TCU task. This discrepancy is notably pronounced, emphasizing the challenges that PLMs face when transitioning from understanding temporal categories to applying this understanding in more complex TCU scenarios.

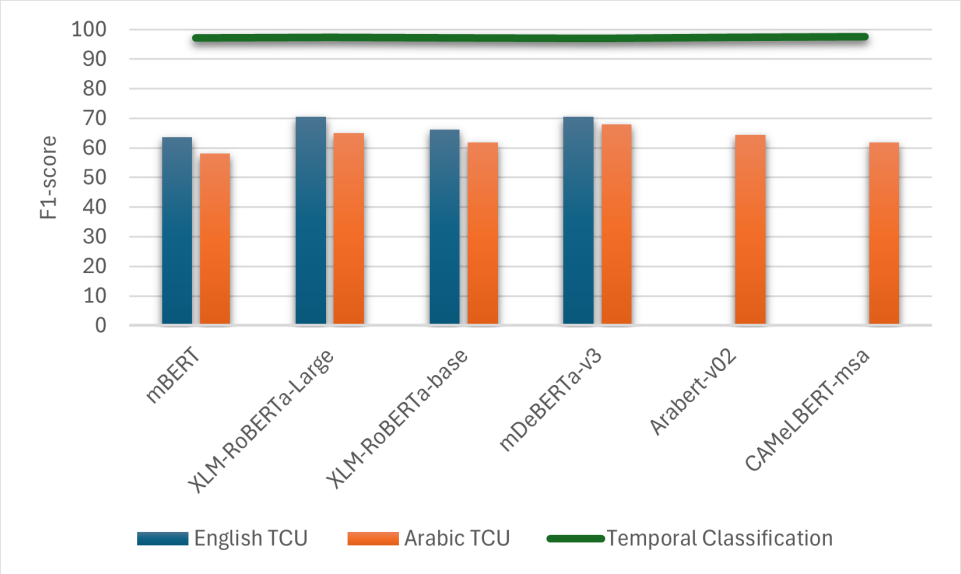

Figure 6. The results of TCU and temporal classification for Arabic and English.

Given the similarity in outcomes between the Arabic and English datasets, the analysis focused in depth on the Arabic dataset to explore this phenomenon further. This targeted approach allows for a nuanced examination of where and how PLMs struggle, particularly in the context of language-specific nuances and temporal reasoning.

Figure 7 breaks down the gap across various temporal aspects, shedding light on the specific areas where discrepancies in performance are most stark. This visual representation serves as a critical tool for identifying the dimensions of temporal understanding

that require further refinement in PLMs, offering insights into potential focus areas for improving model training and development.

Figure 7. Results of Arabic temporal classification in comparison with TCU.

9. Conclusions

This study has made several contributions to the field of TCU, especially for Arabic. The scarcity of datasets has been addressed by constructing an Arabic version from the English MC-TACO dataset. Furthermore, this study examined the performance of several multilingual PLMs across both datasets. As a result, the overall performance exhibited a significant gap between Arabic and English. Moreover, the results showed that, for the English dataset, models originally designed for English revealed a significant performance advantage over their multilingual counterparts. However, this is not the case for Arabic models with an Arabic dataset. This study highlighted the inherent complexity of the Arabic language and emphasized the significant scarcity of suitable datasets for this research area for the Arabic language. Addressing this gap will require a collaborative effort.

To evaluate PLMs' understanding of TCU tasks, a new hypothesis was introduced. Based on this analysis, it was found that the challenge leading to the failure of PLMs lies in the complexity of common sense reasoning rather than in understanding temporal features.

Several issues must be addressed and considered as limitations of this study. First, the size of MC-TACO is quite small, and the split that has been suggested might be adapted to enhance the performance of PLMs in both languages. Additionally, the dataset contains inputs that might conflict with Arabic cultural standards, which might have limited the efficacy of the models with the Arabic dataset.

To address these issues, MC-TACO can be augmented using a temporal common sense dataset specifically designed for Arabic. This can enhance the model performance

by overcoming cultural disparities. Addressing cultural issues could significantly reduce the potential for errors and misunderstandings in the model's output. The construction of an extensive Arabic dataset is the optimal solution. However, this effort could require substantial investment in both resources and specialized knowledge.

In addition, we acknowledge the imbalance in the distribution of temporal categories, particularly the underrepresentation of the stationarity feature. This distribution reflects the natural occurrence of these features in the collected texts. While balancing the dataset could improve evaluations, it may compromise the realism and authenticity of the dataset. As such, this trade-off remains a critical consideration for future work in this domain.

This study serves as a foundation for advancing temporal reasoning in Arabic, offering valuable insights for further research on underrepresented languages and real-world applications such as healthcare, legal systems, and digital assistants.

Author Contributions: This study is part of a PhD research project that is primarily being conducted by the first author R.A. The research was supervised and guided by H.A.-K. and S.O., who provided valuable insights and feedback throughout the process. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset and the code are available from the corresponding author upon request.

Acknowledgments: The proofreaders who assisted in reviewing the dataset are appreciated for their contributions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
DeBERTa	Decoding-enhanced BERT with disentangled attention
EM	Exact Match
NLU	Natural Language Understanding
MRC	Multiple-Choice Reading Comprehension
MSA	Modern Standard Arabic
PLM	Pre-trained Language Model
RoBERTa	Robustly optimized BERT approach
TCU	Temporal Common Sense Understanding
XLM-RoBERTa	Multilingual version of RoBERTa

References

1. Sap, M.; Shwartz, V.; Bosselut, A.; Choi, Y.; Roth, D. Commonsense Reasoning for Natural Language Processing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, Online, 5 July 2020; pp. 27–33. [CrossRef]
2. Schockaert, S.; Ahn, D.; Cock, M.D.; Kerre, E.E. Question Answering with Imperfect Temporal Information. In *Proceedings of the Flexible Query Answering Systems*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Geramny, 2006; pp. 647–658. [CrossRef]
3. Zhou, B.; Khashabi, D.; Ning, Q.; Roth, D. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3363–3369. [CrossRef]

4. Yang, Z.; Du, X.; Rush, A.; Cardie, C. Improving Event Duration Prediction via Time-aware Pre-training. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 3370–3378. [CrossRef]
5. Virgo, F.; Cheng, F.; Kurohashi, S. Improving Event Duration Question Answering by Leveraging Existing Temporal Information Extraction Data. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 4451–4457.
6. Zhou, B.; Ning, Q.; Khashabi, D.; Roth, D. Temporal Common Sense Acquisition with Minimal Supervision. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7579–7589. [CrossRef]
7. Habash, N. *Introduction to Arabic Natural Language Processing*; Synthesis Lectures on Human Language Technologies; Morgan & Claypool Publishers: San Rafael, CA, USA, 2010. [CrossRef]
8. Bousmaha, K.Z.; Rahmouni, M.K.; Kouninef, B.; Hadrich, L.B. A Hybrid Approach for the Morpho-Lexical Disambiguation of Arabic. *J. Inf. Process. Syst.* **2016**, *12*, 358–380.
9. Boudaa, T.; El Marouani, M.; Enneya, N. Arabic Temporal Expression Tagging and Normalization. In Proceedings of the Big Data, Cloud and Applications, Kenitra, Morocco, 4–5 April 2018; pp. 546–557. [CrossRef]
10. Allen, J.F. Maintaining Knowledge about Temporal Intervals. *Commun. ACM* **1983**, *26*, 832–843. [CrossRef]
11. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
12. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
13. Kaddari, Z.; Mellah, Y.; Berrich, J.; Bouchentouf, T.; Belkasmi, M.G. Applying the T5 language model and duration units normalization to address temporal common sense understanding on the MCTACO dataset. In Proceedings of the 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 9–11 June 2020; pp. 1–4. [CrossRef]
14. Raffel, C.; Shazeer, N.M.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2019**, *21*, 140.
15. He, P.; Gao, J.; Chen, W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In Proceedings of the Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, 1–5 May 2023.
16. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
17. Han, R.; Ren, X.; Peng, N. ECONET: Effective Continual Pretraining of Language Models for Event Temporal Reasoning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021; pp. 5367–5380. [CrossRef]
18. Kimura, M.; Pereira, L.K.; Kobayashi, I. Effective Masked Language Modeling for Temporal Commonsense Reasoning. In Proceedings of the 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS), Ise, Japan, 29 November–2 December 2022; pp. 1–4. [CrossRef]
19. Kimura, M.; Kanashiro Pereira, L.; Kobayashi, I. Toward Building a Language Model for Understanding Temporal Commonsense. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop, Online, 20 November 2022; pp. 17–24.
20. Kimura, M.; Kanashiro Pereira, L.; Kobayashi, I. Towards a Language Model for Temporal Commonsense Reasoning. In Proceedings of the Student Research Workshop Associated with RANLP 2021, Online, 2–3 September 2021; pp. 78–84.
21. Pereira, L.; Cheng, F.; Asahara, M.; Kobayashi, I. ALICE++: Adversarial Training for Robust and Effective Temporal Reasoning. In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, Shanghai, China, 5–7 November 2021; pp. 373–382.
22. Tawalbeh, S.; AL-Smadi, M. Is this sentence valid? An Arabic Dataset for Commonsense Validation. *arXiv* **2020**, arXiv:2008.10873. [CrossRef]
23. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based Model for Arabic Language Understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, 12 May 2020; pp. 9–15.
24. Alammary, A.S. BERT Models for Arabic Text Classification: A Systematic Review. *Appl. Sci.* **2022**, *12*, 5720. [CrossRef]

25. Inoue, G.; Alhafni, B.; Baimukan, N.; Bouamor, H.; Habash, N. The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine (Virtual), 9 April 2021; pp. 92–104.
26. Shwartz, V. Good Night at 4 pm?! Time Expressions in Different Cultures. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 2842–2853. [CrossRef]
27. Huang, P.; Bu, J.; Chen, C.; Kang, Z. Question Classification via Multiclass Kernel-based Vector Machines. In Proceedings of the 2007 International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, 30 August–1 September 2007; pp. 336–341. [CrossRef]
28. Kolomiyets, O.; Moens, M.F. A survey on question answering technology from an information retrieval perspective. *Inf. Sci.* **2011**, *181*, 5412–5434. [CrossRef]