Towards Robust Temporal Reasoning of Large Language Models via a Multi-Hop QA Dataset and Pseudo-Instruction Tuning

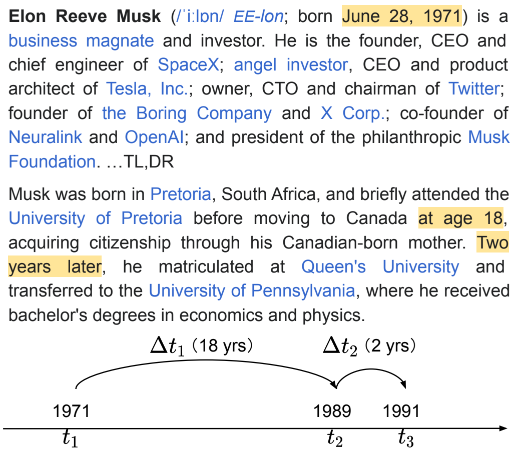
Qingyu Tan [*] [1, 2] **Hwee Tou Ng**[†] [2] **Lidong Bing**[1]
[1]DAMO Academy, Alibaba Group
[2]Department of Computer Science, National University of Singapore
{qingyu.tan,l.bing}@alibaba-inc.com
{qtan6,nght}@comp.nus.edu.sg

Abstract

Knowledge in the real world is being updated constantly. However, it is costly to frequently update large language models (LLMs). Therefore, it is crucial for LLMs to understand the concept of temporal knowledge. However, prior works on temporal question answering (TQA) did not emphasize multi-answer and multi-hop types of temporal reasoning. In this paper, we propose a complex temporal question-answering dataset **Complex-TR** that focuses on multi-answer and multi-hop temporal reasoning. Besides, we also propose a novel data augmentation strategy to improve the complex temporal reasoning capability and robustness of LLMs. We conducted experiments on multiple temporal QA datasets. Experimental results show that our method is able to improve LLMs' performance on temporal QA benchmarks by significant margins[1].

1 Introduction

Time is a fundamental aspect of the real world. Much information comes with an expiry date. Recent advances of large language models (LLMs) (Wei et al., 2022; Ouyang et al., 2022; Achiam et al., 2023) have demonstrated that LLMs can tackle many NLP tasks in a few-shot manner. However, preliminary studies showed that one of the key drawbacks of existing LLMs is the lack of temporal reasoning capability (Chen et al., 2021; Tan et al., 2023). The **SituatedQA** (Zhang and Choi, 2021) dataset was first proposed to incorporate extra-linguistic contexts to QA, which include temporal contexts and geographical contexts. Chen et al. (2021) proposed the **TimeQA** dataset and formulated temporal QA as an open-book QA task. Liska et al. (2022) proposed the **StreamingQA** dataset by WMT data from 2007 to 2020. Dhingra et al. (2022) constructed the **TempLAMA** dataset by the Wikidata Knowledge Base with facts from 2010 to 2020. Tan et al. (2023) proposed a temporal QA benchmark **TempReason** with coverage of long durations and divided temporal reasoning into three levels: time-time reasoning (**L1**), time-event (**L2**) reasoning, and event-event (**L3**) reasoning. The temporal question-answering (TQA) task is essentially answering questions with temporal constraints, and the answers to the questions are derived from time-dependent facts. An example query is $(s, r, ?, t_r)$, where s is the subject, r is the relation, t_r is the reference time for this question, and the answer denoted by $?$ is the object.



Elon Reeve Musk (/ˈiːlɒn/ *EE-lon*; born June 28, 1971) is a business magnate and investor. He is the founder, CEO and chief engineer of SpaceX; angel investor, CEO and product architect of Tesla, Inc.; owner, CTO and chairman of Twitter; founder of the Boring Company and X Corp.; co-founder of Neuralink and OpenAI; and president of the philanthropic Musk Foundation. ...TL,DR

Musk was born in Pretoria, South Africa, and briefly attended the University of Pretoria before moving to Canada at age 18, acquiring citizenship through his Canadian-born mother. Two years later, he matriculated at Queen's University and transferred to the University of Pennsylvania, where he received bachelor's degrees in economics and physics.

Figure 1: An example of a 3-hop temporal expression for t_3. The temporal expressions are highlighted in yellow in the paragraph. The temporal expressions include exact timestamps and time intervals. This example is taken from Elon Musk's Wikipedia page on 18 June 2023.

However, existing temporal question-answering datasets have several common drawbacks. The first drawback is that they fail to examine the temporal co-occurrence phenomenon. In the real world, multiple events can happen at the same time, and temporal questions can have multiple valid answers.

[*] Qingyu Tan is under the Joint PhD Program between Alibaba and NUS.

[†] Corresponding author.

[1] Our code and data are released at `https://github.com/nusnlp/complex-tr`

Findings of the Association for Computational Linguistics ACL 2024, pages 6272–6286
August 11-16, 2024 ©2024 Association for Computational Linguistics

For example, in Figure 1, we can see that Elon Musk is the chief executive officer of both Tesla and SpaceX as of June 2023. Nevertheless, all prior benchmarks followed the SQuAD (Rajpurkar et al., 2016) evaluation metrics, i.e., token-level F1 and exact match (EM) score. These two metrics take the max scores among all answers when there are multiple answers. Such metrics overestimate the performance of temporal QA and cannot properly evaluate questions with multiple answers.

The second drawback of existing temporal question-answering benchmarks is that the questions mainly focused on one-hop temporal reasoning, i.e., only one temporal expression is included in the question. For example, in the question "What team did Kobe Bryant play for in June 2010?", the temporal constraint refers to only one timestamp. In this paper, we define multi-hop temporal questions as questions that contain multiple temporal expressions. The temporal expression can be a timestamp t or a time interval Δt. In the real world, temporal concepts are often expressed by multiple time expressions. For example, in Figure 1, t_3 refers to the year 1991, and it can be explicitly expressed as a numerical value (1-hop expression) or implicitly as $t_1 + \Delta t_1 + \Delta t_2$, as shown in the paragraph. An example question for multi-hop temporal reasoning is "When did Elon Musk move to Canada?". In this case, t_1 is the birth date of Elon Musk and Δt_1 is 18 years. Multi-hop temporal expressions are common in the real world, whereas the study of multi-hop temporal reasoning is under-explored by prior temporal QA datasets. Note that the "multi-hop" concept in this paper refers to temporal hops (number of temporal expressions in a question) and they are different from the graphical hops used in QA over knowledge graphs (KGQA) (Lin et al., 2018; Saxena et al., 2020) and temporal knowledge graphs (Bai et al., 2021; Bai et al., 2023), where the number of graphical hops refers to the number of triples required to answer the question. To the best of our knowledge, we are the first to differentiate temporal hops from KG hops for temporal question answering.

To address the two shortcomings of the existing datasets, we created a temporal QA dataset **Complex TempReason** (**Complex-TR**) that emphasizes multi-hop and multi-answer temporal reasoning. We follow the logical breakdown of the TempReason dataset and focus on time-event (L2) and event-event (L3) reasoning, since these two reasoning types require grounding events to the time axis and are much more challenging than time-time (L1) reasoning. Besides the knowledge from Wikidata KB and Wikipedia articles, our dataset also includes external contexts from Google Custom Search API[2] for the open-domain QA (ODQA) setting. To examine the robustness of temporal reasoning, we only used questions before 2020/01/01 as training data. The examples after 2020/01/01 will be used as unseen future questions. Moreover, for our test set, we engaged college-educated human annotators to verify the correctness of the QA pairs. This human verification process ensures that our test set is of high quality for temporal QA research.

Besides the proposed dataset, we also proposed two methods to improve the performance of temporal QA. The first is Pseudo-Instruction Tuning (PIT), a data augmentation strategy to improve the robustness of temporal reasoning. The second is context refinement, an effective context selection strategy to accommodate long contexts for temporal ODQA. We conducted extensive experiments on our dataset in various TQA settings. Besides, we also conducted experiments on other TQA datasets. Experimental results show that our methods achieve significant performance gains over strong baselines, especially on the out-of-domain years and more complex questions.

In summary, our contributions are as follows:

- We are the first to study multi-hop and multi-answer questions for the temporal QA (TQA) task. We also found out that prior benchmarks for TQA adopted inappropriate evaluation metrics for multi-answer questions.
- We constructed a complex temporal QA dataset **Complex-TR** that covers diverse types of multi-hop temporal reasoning by distant supervision and human verification. Experimental results show that all LLMs perform significantly worse on multi-hop temporal questions.
- We propose a novel data augmentation strategy to create pseudo-instruction tuning data to improve the complex temporal reasoning capabilities and temporal robustness of LLMs. We also proposed an effective context refinement strategy for the long-context problem in ODQA. Extensive experimental results show that our methods significantly improve the performance over strong baselines.

[2] https://developers.google.com/custom-search/v1

Dataset	**Size**	**L2 1-hop**	**L2 M-hop**	**L3 1-hop**	**L3 M-hop**	**Δt Question**	**% M-answer**	**M-answer eval.**
TempLAMA	50K	✓	✗	✗	✗	✗	25.3%	✗
TimeQA	41.2K	✓	✓	✗	✗	✗	6.3%	✗
StreamingQA	147K	✓	✗	✗	✗	✗	25.0%	✗
SituatedQA	12.2K	✓	✗	✗	✗	✗	4.7%	✗
TempReason	52.8K	✓	✗	✓	✗	✗	8.6%	✗
Complex-TR (Ours)	10.8K	✓	✓	✓	✓	✓	23.5%	✓

Table 1: Comparison between Complex-TR and prior temporal QA datasets. The % M-answer column refers to the percentage of multi-answer questions. We can see that all prior datasets contain a considerable number of multi-answer questions, yet none of them used appropriate evaluation metrics for multi-answer questions.

2 Our Dataset

Tan et al. (2023) first proposed to divide temporal reasoning into three levels: time-time reasoning (**L1**), time-event reasoning (**L2**), and event-event reasoning (**L3**). In this paper, we focus on the harder temporal reasoning types: time-event reasoning (**L2**) and event-event reasoning (**L3**). We construct our dataset by the following steps:

Mining Temporal Facts in Wikidata We first extract all the knowledge triples that have temporal qualifiers (such as *start_time* and *end_time*) in the Wikidata (Vrandečić and Krötzsch, 2014) knowledge base (2023/03/20 dump). We then reformat the extracted triples into temporal quintuples (s, r, o, t_s, t_e), where s is the subject, r is the relation, o is the object, t_s is the start time, and t_e is the end time. We then group the quintuples with the same subject together, obtaining $S = \{(s, r_i, o_i, t_{si}, t_{ei}) | i \in 1...N\}$. Unlike prior works from Wikidata (Chen et al., 2021; Tan et al., 2023), where only one relation type is kept within one group, we include multiple relation types in one group, which adds more complexity to our dataset. For each group, we identify the most common relation as its representative relation. Since the relation distribution is highly imbalanced in the Wikidata KB, we set a ceiling of 250 groups for each representative relation type. In the end, we kept 2,000 temporal quintuple groups, and on average there are 9.2 temporal quintuples in each group. We divide the groups into training (1,000), development (500), and test (500) sets.

Creating Questions from Quintuples After obtaining the temporal quintuple groups, we create the question-answer pairs based on manually designed templates (details are shown in Appendix F). For L2 temporal questions, a 1-hop question can be expressed in the query $(s, r, ?, t_r)$, where t_r is the reference timestamp. We create the multi-hop L2 questions with two variations: (1) $(s, r, ?, t_{rs}, t_{re})$, where t_{rs} and t_{re} refer to the start and end time of the question respectively. (2) $(s, r, ?, t_r, \Delta t)$, where t_r is the reference time and Δt is the temporal difference between the reference time and the query time. The model is expected to infer the query time from Δt and t_r. As for the L3 questions, the number of temporal hops is dependent on the event-event temporal relation. For example, from the question "What team did Kobe Bryant play for before the LA Lakers?", we can imply that the query time is the start time of Kobe Bryant playing for the Lakers. Since only one timestamp is implied in the question, it is a 1-hop question. In contrast, for the question, "What awards did Kobe Bryant win when he was playing for the LA Lakers?", both the start time and end time have to be considered. Since Kobe Bryant had a 20-year career with the Lakers, it is a multi-hop question.

In order to have a fair comparison between 1-hop and multi-hop temporal reasoning, we also created a small number of 1-hop questions with the same temporal quintuples and contexts. We denote the timestamp mentioned in the question as the reference time. For L3 questions, we denote the start time of the reference event as the reference time. In order to examine the robustness of temporal reasoning in future years, we only use the questions with a reference time before 2020/01/01 for training, whereas the development and test sets contain questions both before and after 2020/01/01.

Open Domain Context Retrieval Previous temporal QA datasets typically rely on a fixed knowledge source. The TimeQA and TempReason datasets use Wikipedia articles as context for open-book QA setting. The StreamingQA dataset uses English news articles from the WMT challenge. TimeQA is a human-annotated dataset based on Wikipedia articles, and temporal facts not reflected in Wikipedia are deemed as "unanswerable". However, a single knowledge source may not be sufficient to answer temporal questions, as temporal facts are constantly evolving. To address this limitation, we construct our dataset in an open-domain QA (ODQA) fash-

Dataset	Type	Example Question	Answers
TempLAMA	L2 1-hop	In 2011, Tom Brady played for _X_.	**New England Patriots**
TimeQA	L2 M-hop	Which team did Olivier Bernard play for from 2000 to 2005?	**Newcastle United**
StreamingQA	L2 1-hop	Which player scored for St Mirren in November 2008?	**Franco Miranda**
SituatedQA	L2 1-hop	Who made the most free throws in NBA history as of 2020?	**Karl Malone**
TempReason	L2 1-hop	Where was Barack Obama educated in Apr 1981?	**Columbia University**
	L3 1-hop	Who was the chair of Swedish People's Party of Finland after Lars Erik Taxell?	**Jan-Magnus Jansson**
Complex-TR (Ours)	L2 M-hop	Who were the chairs of FC Barcelona from March 1984 to March 2003?	**Josep Lluís Núñez** and **Enric Reyna**
	L2 M-hop	Where was Lynne Bowker educated 15 years before June 2005?	**University of Ottawa**
	L3 M-hop	Which employer did Barack Obama work for 2 years after he/she studied at Occidental College?	**Business International Corporation**
	L3 M-hop	Who were the owners of Chelsea F.C. when Thomas Tuchel was the headcoach?	**Roman Abramovich** and **Todd Boehly**

Table 2: Example questions of prior temporal QA datasets and our Complex-TR dataset.

ion. For the context used in the ODQA setting, we first include the Wikipedia article on the question subject. We then use the Google Custom Search API with this question as the search query to retrieve the top 10 results. The searched web pages will be scraped by Trafilatura (Barbaresi, 2021), a state-of-the-art text extraction tool for NLP research. Due to the anti-scraping mechanism of certain websites, we cannot extract the contexts from all the retrieved web pages. On average, we have 6.7 articles for each question and the average total context length is 93K words. The long context from multiple sources introduces additional challenges for the temporal QA task.

Human Verification To ensure that our dataset is of high quality and aligned with the retrieved contexts, we engaged college-educated human annotators to verify the correctness of the QA pairs and retrieved contexts. The annotators are given a temporal QA pair and its corresponding articles, and they are asked to read through the articles and judge whether the QA pair is correct or not. Due to cost limitations, we randomly sample 500 QA pairs from the test set for human verification. Among the 500 QA pairs, 100 of them are annotated by two annotators to measure the inter-annotator agreement. The Fleiss Kappa coefficient of this subset is 0.71, which implies a substantial agreement level. The conflicts are resolved by a third annotator. The hourly pay of the annotators is above 22 USD, which is significantly higher than the local minimum wage. In the end, 329 QA pairs are deemed as correctly reflected in the contexts, we then used these 329 QA pairs as our gold test set in the experiments. We name our dataset Complex-TempReason (**Complex-TR**). Our detailed dataset statistics are shown in Table 3. The"Pseudo" column refers to the pseudo-data that we generated in Section 3.1. Besides, we also compare our questions and reasoning types with prior temporal QA datasets in Table 1. We can see that none of the prior datasets included L3 multi-hop temporal reasoning, or questions with time interval Δt. Besides, all prior temporal QA datasets contain a considerable number of questions with multiple answers, yet none of them adopted any evaluation metrics for multi-answer questions. We provide a more detailed question comparison of our dataset and prior works in Table 2.

	Pseudo	Training	Dev	Test (Gold)
Time Coverage	1021-2040	1529-2019	1254-2023	1659-2023
L2 1-hop	11,389	1,109	670	72
L2 M-hop	13,919	2,407	1,487	106
L3 1-hop	15,205	1,236	690	71
L3 M-hop	10,938	1,759	1,146	80
#Total	51,451	6,511	3,993	329
Avg. Facts	10.8	9.0	10.0	8.6
Avg. Contexts	-	6.6	6.8	6.7
Avg. Word Len.	-	92K	99K	94K

Table 3: Statistics of our dataset. Note that we do not include questions after December 2019 in our training set. The **Avg. Word Len.** row refers to the average number of words in all contexts for the ODQA setting.

3 Methodology

In this paper, we also propose two strategies to improve the robustness and the capabilities of LLMs for temporal reasoning, which are Pseudo-Instruction Tuning and Context Refinement. Pseudo-Instruction Tuning uses pseudo data to alleviate the data scarcity and data imbalance problems of temporal QA, which can improve LLMs' temporal reasoning capability and robustness. The improved reasoning capability also has a positive impact in the ODQA setting. On the other hand, Context Refinement is used to address the long context problem in the ODQA setting. In this setting, the total context can reach over 100K tokens, which makes it infeasible to feed the context to existing QA models. The Fusion-in-Decoder (FiD; Izacard and Grave, 2021) model was proposed for the QA task to process long contexts. It breaks a long context into multiple smaller paragraphs to avoid the quadratic computation of self-attention. However, even FiD can only process up to 10K tokens, which is still insufficient for ODQA. As such, we use sentence embedding models to refine the paragraphs and only keep the most relevant paragraphs to the question for our QA models. We will introduce each strategy in detail in the following subsections.

3.1 Pseudo-Instruction Tuning

One of the main challenges of temporal QA is that the labeled data are typically concentrated on recent years, and some datasets (Liska et al., 2022; Dhingra et al., 2022) only contain data from 2000-2020. The data imbalance is caused by data distribution in the Wikidata knowledge base. As a result, LLMs trained by such data tend to be biased towards recent years (Tan et al., 2023). To overcome this challenge, we aim to create artificial data with an emphasis on low-frequency time periods.

Pseudo-Data Generation Data augmentation has proven to be effective in many NLP tasks (Zhou et al., 2021; Ding et al., 2020; Cao et al., 2023). For TQA, we have the training group S of subject s as:

$$S = \{(s, r_i, o_i, t_{si}, t_{ei}) | i \in 1...N\} \tag{1}$$

We then shift S by Δt for every fact within that group, obtaining:

$$S_p = \{(s, r_i, o_i, t_{si} + \Delta t, t_{ei} + \Delta t) | i \in 1...N\} \tag{2}$$

where $-100 \leq \Delta t \leq +20$ is a random temporal shift with a maximum of 20 years going forward and a maximum of 100 years going backward, and S_p is the shifted pseudo-group. Since shifting temporal facts introduces temporally augmented facts, we replace all the subjects and objects with fictional entities to avoid conflicts. We used ChatGPT to generate multiple types of fictional entities, such as person names and sports teams. This process can be repeated multiple times. In this way, we can create large amounts of artificial temporal quintuples. We then follow the question templates in Section 2 to generate question-answer pairs using fictional facts, and hence generate temporal reasoning data without human annotation. Examples of fictional entities and fictional temporal facts are shown in Appendix G.

Temporal Resampling To accommodate data imbalance in the Wikidata KB, especially for future years, we resample the generated pseudo-data by time intervals. Specifically, we divide time into 20-year intervals from 1900 to 2020, and then count all the examples in our training dataset within each time interval, obtaining $\{n_i | i \in 1..k\}$. Examples before 1900 will be treated as one group, so there are 7 counts in total and $k = 7$. Note that our training data does not contain any questions after December 31, 2019. Hence, examples after December 31, 2019 in the development and test sets will be used to simulate future data. We calculate probabilities for resampling by:

$$P_i = 1 - \frac{n_i}{max(\{n_i | i \in 1..k\})} \tag{3}$$

We then sample the generated questions with probability P_i. For pseudo-data after 2019, we set the sampling probability to be 1. In this way, we will be able to de-bias the temporal distribution in the Wikidata KB and let the models focus on improving the performance on low-frequency years.

Training for PIT After we obtained the resampled pseudo-data, we followed the instruction templates for QA tasks from FLAN (Wei et al., 2022) to fine-tune the LLMs. We name this process Pseudo-Instruction Tuning (**PIT**). The final size and statistics of PIT are shown in the "Pseudo" column in Table 3. The instruction-tuned LLMs will then be used to fine-tune on the task-specific data. We believe that improving temporal reasoning capability can have a positive impact on downstream tasks such as ODQA.

3.2 Context Refinement

In the ODQA setting, the contexts are from multiple sources. We first denote our context set as C, and follow the pre-processing protocol of FiD to split all the articles into 100-word paragraphs:

$$C = \{p_1, p_2, p_3, ..., p_m\} \tag{4}$$

where p_i refers to a paragraph in the context set. We then encode the temporal question q and the paragraphs by a sentence embedding model f and then calculate their cosine similarity:

$$z_i = Cos(f(q), f(p_i)) \tag{5}$$

z_i is used as a relevance score to re-rank all the paragraphs. Due to computation constraints, we only use the top k paragraphs as the contexts for the FiD model. In this way, we can refine the extra-long context to an acceptable level. We examined multiple sentence embedding models and chose `bge-base-en-v1.5` (Xiao et al., 2023) as our sentence encoder f. It is an advanced embedding model on the MTEB (Muennighoff et al., 2023) benchmark and works best in our experiments We show the ablation studies of different re-rankers in Appendix C.

4 Experiments

4.1 Experimental Setup

In this paper, we focus on two temporal QA settings. The first is open-domain QA (**ODQA**), where the models are provided with multiple retrieved articles as context. The second is the **ReasonQA** setting proposed by Tan et al. (2023), where all the relevant structured knowledge quintuples to answer a question are provided as context. This is because we aim to study the reasoning aspect of temporal QA. We elaborate on the problem settings in Appendix B in greater detail.

Baselines (1) **FLAN-T5-XL** (Wei et al., 2022) This model is an instruction-tuned encoder-decoder model with 3B parameters. It achieves respectable few-shot performance on the MMLU benchmark. (2) **GPT-3.5** (Ouyang et al., 2022) We used `gpt3.5-turbo` as our baseline. (3) **GPT-4** (Achiam et al., 2023) This model is the most advanced LLM in the market. It achieves strong zero-shot performance on many NLP tasks. Since the model is constantly being updated, we used the `gpt4-0613` model for consistent evaluation. We evaluate the one-shot performance of the first three LLMs. (4) **T5-SFT** (Raffel et al., 2020) This model is the supervised fine-tuned model with our labeled training data. We used the T5 models as our backbone model. We conducted experiments on T5-base (**T5-B**) and T5-large (**T5-L**). In the ODQA setting, we truncate the context to 1,024 tokens due to GPU memory constraint. (5) **T5-PIT-SFT** This model is first instruction-tuned by pseudo data and then further fine-tuned with labeled data. (6) **FiD** is the supervised fine-tuned baseline for FiD. It uses T5 as its backbone model and splits a long context into smaller paragraphs. Hence, the maximum context window of FiD is significantly larger than T5 under the same memory constraint. (7) **FiD-PIT** combines T5-PIT initialization with FiD training strategy. (8) **FiD-PIT-Refined** is the FiD-PIT model with context refinement described in Section 3.2. For SFT models, we report the average result of three random runs.

4.2 Evaluation Metrics

As mentioned in Section 1, all of the prior TQA benchmarks followed SQuAD (Rajpurkar et al., 2016). However, the metrics in SQuAD are computed by using the maximum scores with all references, which significantly overestimate the performance for multi-answer questions. Therefore, we adopted two stricter metrics for our experiments. The first metric is set-level accuracy (**Set Acc.**; Zhong et al., 2023). This metric will only return correct if the prediction set is identical to the ground truth set. The second additional metric is answer-level F1 (**Ans. F1**; Amouyal et al., 2022). Unlike token-level F1 in SQuAD, **Ans. F1** counts true positives only when there is an exact match in the answer set. Besides, if the prediction contains extra answers, it will also be penalized. The upper bound of these two metrics is the EM score. We also further analyze the four metrics in Section 5.2.

	Single-hop		**Multi-hop**	
	Set Acc.	**Ans. F1**	**Set Acc.**	**Ans. F1**
FLAN-T5-XL	61.5	64.1	35.5	49.7
GPT-3.5	28.0	45.3	31.2	51.8
GPT-4	67.1	80.2	51.6	65.4
T5-*base*				
SFT	80.4	83.3	59.1	65.1
PIT-SFT (Ours)	**91.6**	**93.8**	**78.0**	**82.4**
T5-*large*				
SFT	86.0	88.1	71.0	76.4
PIT-SFT (Ours)	**95.1**	**95.6**	**85.0**	**89.5**

Table 4: ReasonQA experimental results (in %) that compare single-hop and multi-hop temporal reasoning. The context used in this setting is structured facts in KB.

4.3 Experimental Results

In Table 4, we can see that multi-hop temporal reasoning is much more challenging compared to single-hop reasoning, and most tested models have lower performance on multi-hop temporal reasoning. This shows that multi-hop temporal reasoning is a common weakness for current LLMs. The GPT-4 model can achieve relatively good results for the answer F1 metric, whereas its set accuracy scores are still significantly below our supervised models. This indicates that GPT-4 can find valid answers under time constraints, but it cannot find all correct answers when multiple answers are presented. We provided the example errors by GPT-4 in Appendix E.

The experimental results of open-domain QA are shown in Table 5. The ODQA setting is a much

	Single-Hop		Multi-hop	
	Set Acc.	Ans. F1	Set Acc.	Ans. F1
FLAN-T5-XL	30.1	31.5	14.5	20.3
GPT-3.5	17.5	26.3	9.7	23.0
GPT-4	19.6	35.1	14.0	37.2
T5-*base*				
SFT	33.6	35.6	17.7	26.6
PIT-SFT (Ours)	39.9	40.8	22.6	30.1
FiD	33.6	35.6	17.7	26.0
FiD-PIT (Ours)	39.2	40.1	23.1	30.2
+Refine (Ours)	**42.7**	**44.3**	**24.7**	**31.2**
T5-*large*				
SFT	35.0	35.7	23.1	32.1
PIT-SFT (Ours)	39.2	40.1	25.3	32.8
FiD	46.2	46.6	27.4	37.3
FiD-PIT (Ours)	46.9	47.8	29.0	37.5
+Refine (Ours)	**49.0**	**49.7**	**31.2**	**39.1**

Table 5: ODQA experimental results (in %). The contexts used in this setting are multiple web articles.

harder QA setting. We can see that all models perform significantly worse than in the ReasonQA setting. Nevertheless, our **PIT-SFT** and **FiD-PIT** models still outperform their corresponding baselines (**SFT** and **FiD**) significantly. This experimental result verified our assumption that improving temporal reasoning capability can also improve the downstream performance of ODQA. We also show the experimental results of **PIT** on the TimeQA dataset in Appendix D. We find that **PIT** also has a positive impact on TimeQA, which demonstrates the generalizability of **PIT**.

Since our dataset contains multiple articles for ODQA, the total context length can easily exceed FiD's limit. However, with our proposed context refinement strategy, we can further improve our best model **FiD-PIT** (large) by 2.1% in set accuracy for single-hop questions and 2.2% in set accuracy for multi-hop questions

	In-domain	Future	Overall
		Set Acc.	
FLAN-T5-XL	46.5	54.5	46.8
GPT-3.5	30.2	18.2	28.0
GPT-4	58.2	63.6	58.4
T5-B	69.2	45.5	68.4
T5-B-PIT	84.9	54.5	83.9
T5-L	78.0	63.6	77.5
T5-L-PIT	**89.3**	**90.9**	**89.4**

Table 6: Analysis of the ReasonQA performance for the in-domain years (before 2020/01/01) and out-of-domain future years. The numbers reported in this table are the **Set Acc.** scores.

5 Analysis

5.1 Robustness of Temporal Reasoning

In this section, we analyze temporal reasoning performance by time periods. In an ideal scenario, temporal reasoning capability should generalize to unseen time periods. In Table 6, we show the experimental results for the in-domain and the out-of-domain (OOD) subsets. We treat the questions after 2020/01/01 as OOD (future examples). We believe that this is a valid assumption because our training data do not contain such questions and our backbone model T5 was released in October 2019. From Table 6, we can see that **FLAN-T5-XL** and **GPT-4** have higher performance on the future subset. This could be because these two models are fine-tuned on data after 2020. For the supervised models, we can see that **T5-B**, **T5-B-PIT**, and **T5-L** all suffered from severe performance degradation on the future subset, whereas **T5-L-PIT** can achieve similar performance on in-domain years and future years. This implies that achieving robust temporal reasoning not only requires de-biased pseudo-data but also a good capability of the base language model.

	Single-Answer			
	Set Acc.	Ans. F1	EM	Tok. F1
FLAN-T5-XL	59.5	59.5	59.9	70.7
GPT3.5	29.0	48.5	53.7	62.7
GPT-4	60.6	71.2	73.8	79.0
T5-B	71.8	72.1	72.2	78.0
T5-B-PIT	86.5	87.0	86.5	89.0
T5-L	79.9	81.2	80.7	84.4
T5-L-PIT	**90.7**	**91.8**	**91.5**	**93.9**
	Multi-Answer			
	Set Acc.	**Ans. F1**	**EM**	**Tok. F1**
FLAN-T5-XL	0.0	43.1	65.7	81.7
GPT3.5	32.9	50.8	62.9	75.2
GPT-4	50.0	74.2	91.4	92.7
T5-B	55.7	76.5	87.1	91.5
T5-B-PIT	74.3	88.6	94.3	96.2
T5-L	68.6	82.4	90.0	94.1
T5-L-PIT	**84.3**	**93.5**	**98.6**	**99.8**

Table 7: Experimental comparison between the performances on single-answer questions and multi-answer questions of selected models. The experiments are conducted on the Complex-TR dataset in the ReasonQA setting.

5.2 Analysis of Multi-Answer Questions

The ability to understand co-occurring events is a crucial aspect of temporal reasoning. In this section, we analyze the reasoning performance on single-answer and multi-answer questions. We report their experimental results separately in Table 7. Besides the set-level accuracy (**Set Acc.**) and answer-level F1 (**Ans. F1**) reported in the main experiments, we also include exact match (**EM**) and token-level F1 (**Tok. F1**). The **EM** and **Tok. F1** scores are adopted by all prior temporal QA benchmarks. Both metrics take the maximum scores with all possible answers. In the multi-answer scenario, **EM** and **Tok. F1** are generally much higher than **Set Acc.** and **Ans. F1**. **Tok. F1** scores for multi-answer questions are even higher than those of single-answer questions. This indicates that **EM** and **Tok. F1** can significantly overestimate the performance of multi-answer questions. Therefore, it is better for temporal QA benchmarks to adopt **Set Acc.** and **Ans. F1** for evaluation.

6 Related Work

Temporal Information Extraction Early studies of temporal research in NLP focused on studying temporal relations of short-term events. The TimeBank (Pustejovsky et al., 2003) dataset was first proposed as a benchmark for the temporal information extraction (TIE) task. It is a human-annotated dataset with annotated events, temporal expressions, and temporal relations (such as *before*, *after*, and *contains*). The TempEval challenges (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013) were later proposed for the TIE task. The schema of TempEval is similar to that of TimeBank. Cassidy et al. (2014) found that prior TIE datasets were not exhaustively annotated and introduced a dense annotation schema for event ordering. They also released the Timebank-Dense dataset, which is a more complete version of TimeBank. Han et al. (2019) proposed an end-to-end framework to jointly extract events and temporal relations. However, TIE research focused on studying the order of short-term events within a specific context. On the other hand, the focus of our paper is studying temporal reasoning with factual grounding on the global time axis.

Temporal Question Answering Since time is a fundamental aspect in real-life applications, numerous efforts have been made to study the temporal reasoning problem in question answering (QA). The first line of work in this field worked on QA over temporal knowledge graphs (TKGs). The TempQuestions (Zhen et al., 2018) dataset was introduced to extend the KGQA task to TKGs. Jia et al. (2021) introduced the TimeQuestions dataset as an extension of TempQuestions. The CronQuestions (Saxena et al., 2021) dataset is a large-scale QA dataset over TKG with more complex questions. Shang et al. (2022) proposed a temporal contrastive learning approach to improve the time-sensitivity for the TKGQA task. However, the TKGQA task is not the focus of our paper, and it requires models to rank all the nodes (including entities and timestamps) for each question. That is, the TKGQA task assumes that all nodes are known to the model, whereas our focus is on performing temporal reasoning over raw natural language text.

The temporal QA task for LLMs was derived from conventional QA. Even though there is a subset of time-related questions in popular QA datasets such as SQuAD (Rajpurkar et al., 2016), the questions are usually asking for a temporal expression present in the context without temporal reasoning. The MC-TACO (Zhou et al., 2019) dataset was later proposed to study temporal commonsense reasoning. However, MC-TACO did not study the evolving aspect of temporal reasoning, which is crucial for LLMs' continual learning. SituatedQA (Zhang and Choi, 2021) first introduced extra-linguistic context to conventional QA, and it contains evolving temporal questions. Chen et al. (2021) proposed the TimeQA dataset with Wikipedia articles and the Wikidata knowledge base. The TempLAMA (Dhingra et al., 2022) dataset was later constructed similarly, but it focused on the close-book QA setting. Kasai et al. (2023) proposed a real-time QA benchmark that updates questions weekly. Tan et al. (2023) systematically tackled the TQA problem by breaking down temporal reasoning into three levels. However, even though most of the previously proposed datasets contain questions with multiple answers, none of them studied multi-answer and multi-hop temporal reasoning. The concept of "multi-hop" is commonly used in QA to describe complex questions. The term "multi-hop" has different meanings in different contexts. Yang et al. (2018) used the term to describe questions that can only be answered by multiple paragraphs. In the prior works of KGQA (Lin et al., 2018; Saxena et al., 2020), "multi-hop" describes questions that can only be answered by multiple knowledge triples. Since our paper focuses on temporal reasoning, we use "multi-hop" to describe questions that contain multiple temporal expressions. A temporal expression can be a specific timestamp or a time interval.

7 Conclusions

In this paper, we studied the under-explored multi-hop temporal reasoning problem in temporal QA. We proposed a novel dataset Complex-TR that covers multi-hop temporal reasoning. Besides, we found out that all prior temporal reasoning benchmarks used inappropriate evaluation metrics (exact match and token F1) for this task. In addition, we proposed Pseudo-Instruction Tuning to enhance the robustness of temporal reasoning and Context Refinement to alleviate the long-context problem in ODQA. Extensive experimental results showed that our methods are significantly better than strong baseline methods.

8 Limitations

Since our dataset is constructed from the Wikidata knowledge base, it may retain some errors present in the Wikidata KB. That is, the training and validation sets of our dataset may contain factual errors. However, we ensured the high quality of our gold test set by a rigorous human verification process. The other limitation, for the experiments on Open-domain QA, we leveraged the Wikipedia page and the retrieved results from Google Custom Search API on 2023/12/10. The retrieval results may be different as the internet evolves. Our experimental results are also dependent on the retrieval performance of Google Custom Search API. Nevertheless, it is by far the best-performing retrieval tool for the ODQA task as shown in other QA works (Kasai et al., 2023; Zhao et al., 2023).

9 Ethics Statement

We created our Complex TempReason (Complex-TR) dataset from the Wikidata knowledge base. Wikidata is open-source and under the Creative Commons CC0 License[3] and Wikipedia articles are under the Creative Commons AttributionShare-Alike 3.0 License[4] (CC BY-SA). Therefore, these data can be re-engineered to construct the Complex-TR dataset. Besides, we also engaged college-educated human annotators to verify our test data. We offer the annotators competitive compensation with more than 22 USD in hourly pay, which is significantly higher than the local minimum wage. We also open-source our data and code under the CC BY-SA license. Complex-TR is meant for academic research of LLMs' temporal robustness and reasoning capabilities. However, the retrieved content from Wikipedia and Google Custom Search may contain inappropriate language. Besides, our data augmentation method is based on fictional entities generated by the free version of the ChatGPT[5] model. Some of the fictional entities may overlap with real entity names or people's names by coincidence. The generated entities are purely used to improve the temporal reasoning and robustness of LLMs. The authors of this paper hold neutral views toward the generated entities and the contents retrieved from Google Custom Search.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Samuel Joseph Amouyal, Ohad Rubin, Ori Yoran, Tomer Wolfson, Jonathan Herzig, and Jonathan Berant. 2022. QAMPARI: an open-domain question answering benchmark for questions with many answers from multiple paragraphs. *arXiv preprint arXiv:2205.12665*.

Luyi Bai, Mingzhuo Chen, Lin Zhu, and Xiangxi Meng. 2023. Multi-hop temporal knowledge graph reasoning with temporal path rules guidance. *Expert Systems with Applications*.

Luyi Bai, Wenting Yu, Mingzhuo Chen, and Xiangnan Ma. 2021. Multi-hop reasoning over paths in temporal knowledge graphs using reinforcement learning. *Applied Soft Computing*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. MS MARCO: a human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of ACL: Demo*. Association for Computational Linguistics.

Hannan Cao, Wenmian Yang, and Hwee Tou Ng. 2023. Mitigating exposure bias in grammatical error correction with data augmentation and reweighting. In *Proceedings of EACL*.

[3] https://www.wikidata.org/wiki/Wikidata:Licensing

[4] https://en.wikipedia.org/wiki/Wikipedia:Copyrights

[5] https://chat.openai.com/

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of ACL*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Proceedings of NeurIPS*.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-Aware Language Models as Temporal Knowledge Bases. *Transactions of ACL*.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of EMNLP*.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of EMNLP*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of EACL*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. ATLAS: few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of CIKM*.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. RealTime QA: What's the answer right now? In *Proceedings of NeurIPS*.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. 2018. Multi-hop knowledge graph reasoning with reward shaping. In *Proceedings of EMNLP*.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. 2022. StreamingQA: a benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of ICML*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of EACL*.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Proceedigs of NeurIPS*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. In *Proceedings of ACL*.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of ACL*.

Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. 2022. Improving time sensitivity for question answering over temporal knowledge graphs. In *Proceedings of ACL*.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of ACL*.

Naushad UzZaman, Hector Llorens, Leon Derczynski, Marc Verhagen, James Allen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of SemEval*.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of SemEval*.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of SemEval*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *CACM*, pages 78–85.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *Proceedings of ICLR*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. C-Pack: Packed resources for general Chinese embeddings. *arXiv preprint arXiv:2309.07597*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of EMNLP*.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of EMNLP*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of ACL*.

Jia Zhen, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strotgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Proceedings of WWW*.

Victor Zhong, Weijia Shi, Wen-tau Yih, and Luke Zettlemoyer. 2023. RoMQA: a benchmark for robust, multi-evidence, multi-answer question answering. *Findings of EMNLP 2023*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. “going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding. In *Proceedings of EMNLP*.

Ran Zhou, Ruidan He, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2021. MELM: data augmentation with masked entity language modeling for cross-lingual NER. *arXiv preprint arXiv:2108.13655*.

A Implementation Details

For the fine-tuning experiments, we used NVIDIA-V100 GPUs. The experiments on **T5-L** were conducted on 1 40GB-A100 GPU and experiments on **T5-B** were conducted on 1 12GB-Titan X GPU. For the **FLAN-T5-XL** experiments, inference was conducted on 1 40GB-A100 GPU. For the **T5-B** experiments, the number of training epochs for **PIT** was 10 and for **SFT** 15. For **T5-L** experiments, the number of training epochs was 10 for both **PIT** and **SFT**. For the implementation of the FiD model, we followed the GitHub repository of ATLAS[6] (Izacard et al., 2023). This is because this implementation is compatible with later transformer versions and includes many memory-saving functions, such as gradient checkpointing, which suits our computation budget. We use the top 100 ($k = 100$) paragraphs for the FiD experiments. The learning rate was set to 1e-5 for all fine-tuning experiments. For the questions with multiple answers, we join all the answers by a special connector “and”. The predicted string of the models is also split by this connector. For the prompting experiments on **GPT-3.5** and **GPT-4**, we used the OpenAI API. The estimated total cost for reproducing all our experiments is 120 USD.

B Problem Settings

In this section, we elaborate on the two problem settings in detail. In Table 8, we show an example of a ReasonQA context. The example is about the politician Layla Moran and all related temporal knowledge of Layla Moran is provided as context. In this scenario, a human is able to perform temporal reasoning easily by searching for answers based on the time constraints in the questions. This capability should not be affected by the change of time points. However, in Table 6, we see that LLMs have large performance variations over the years. The results on future years are generally worse than average. However, constantly updating LLMs with new data can be costly and susceptible to catastrophic forgetting. Therefore, it is crucial to adapt LLMs to unseen temporal expressions, e.g., future year tokens.

On the other hand, the context we used for the ODQA setting is the Wikipedia article on the subject. In a more general open-domain QA setting, models can leverage off-the-shelf retrieval modules to extract more relevant contexts from all over the web. However, the focus of our work is to study the temporal reasoning capability and robustness of LLMs on the temporal QA task. Hence, we leave the open-domain QA setting for future research.

C Ablation Studies

C.1 Pseudo-Instruction Tuning

In Table 10, we show the ablation studies of pseudo-instruction tuning (**PIT-SFT** models). We examine the two components of **PIT** in the ReasonQA

[6] `https://github.com/facebookresearch/atlas`

ReasonQA Context	Reason Type	Example Question	Answers
Layla Moran held the position of Member of the 57th Parliament of the United Kingdom from June 2017 to November 2019.	L2 1-Hop	Where was Layla Moran educated in November 2005?	Brunel University
Layla Moran studied at UCL Institute of Education from September 2007 to September 2008.	L2 M-Hop	Where was Layla Moran educated from May 2003 to July 2006	Imperial College London, Brunel University
Layla Moran held the position of Member of the 58th Parliament of the United Kingdom from December 2019 to May 2023.	L2 M-Hop	Where was Layla Moran educated 6 years and 2 months after May 2002?	UCL Institute of Education
Layla Moran studied at Brunel University from September 2005 to March 2007.	L3 1-Hop	Where was Layla Moran educated before she studied at Brunel University?	Imperial College London
Layla Moran studied at Imperial College London from September 2000 to August 2003.	L3 M-Hop	Where was Layla Moran educated 4 years and 11 months after he/she studied at Imperial College London	UCL Institute of Education

Table 8: An example of a ReasonQA context, where the subject is **Layla Moran**. All information in the **ReasonQA Context** column is provided to the model along with the question. For the ODQA experiments in our paper, the context will be changed to the Wikipedia article of the subject and the web-retrieved articles based on the question.

	Single Hop		Multi-hop		Overall	
	Set Acc.	Ans. F1	Set Acc.	Ans. F1	Set Acc.	Ans. F1
Baseline	46.9	47.8	29.0	37.5	36.8	42.1
Contriever	45.5	46.7	26.9	35.4	35.0	40.3
Contriever-MSMARCO	47.6	48.7	30.7	**39.4**	38.0	43.4
GTE	**50.4**	**51.1**	29.0	36.3	38.3	42.7
BGE	49.0	49.7	**31.2**	39.1	**38.9**	**43.7**

Table 9: Comparison of sentence embedding models for context refinement. The baseline model refers to the **FiD-PIT** model with the T5-large encoder and simply takes the top 100 passages as context.

	L2 1-Hop Set Acc.	L2 M-Hop Set Acc.	L3 1-Hop Set Acc.	L3 M-Hop Set Acc.
T5-B-PIT-SFT	87.0	71.9	90.5	69.1
-Resampling	85.1	69.7	88.7	66.7
-Fictional	83.9	68.6	87.9	65.5
T5-L-PIT-SFT	91.3	72.7	91.1	71.9
-Resampling	89.8	71.1	89.3	69.8
-Fictional	88.7	69.8	87.7	68.4

Table 10: Ablation studies of our **T5-B-PIT-SFT** and **T5-L-PIT-SFT** models on the validation set of Complex-TR in the ReasonQA setting.

setting. The first is temporal resampling, and the second is the usage of fictional names. We created two other pseudo-instruction training sets of the same size and examined the final **PIT-SFT** performance. From Table 10, we can see that removing temporal resampling leads to 2.2 set accuracy drop for L2 M-hop questions and 2.4 for L3 M-hop questions (for the **T5-B** model). This shows that for the same amount of data, emphasizing the data of low-frequency years leads to better performance. On the other hand, if we use real-world data with shifted temporal information, the performance drop is significant. For the **T5-B-PIT-SFT** model, changing the fictional names to real-world data can lead to 3.3 and 3.6 set accuracy drop for L2 and L3 multi-hop questions. This performance drop could have resulted from the lack of entity diversity from the temporally-shifted data.

C.2 Comparison of Context Refinement Models

Since the context length of our ODQA experiment exceeds the limit of most LLMs, the context refinement process is highly important for the performance of ODQA. In this section, we show the performances of different sentence embedding models for the ODQA setting. We experimented with several popular sentence embedding models for information retrieval and leading open-source models on the Massive Text Embedding Benchmark (MTEB, Muennighoff et al., 2023).

The models include: (1) **Contriever** (Izacard et al., 2022) This model is a popular embedding model trained on contrastive learning in an unsupervised manner. (2) **Contriever-MSMARCO** (Izacard et al., 2022) This model is the contriever model further fine-tuned on the massive MS MARCO (Bajaj et al., 2016) dataset, which drastically improved the information retrieval performance of contriever. (3) **GTE** The General Text Embedding model was trained by multi-stage contrastive learning and demonstrated strong performance in various sentence embedding tasks, such as MNLI. (4) **BGE**

(Xiao et al., 2023) This model is short for BAAI General Embedding. It includes a family of sentence embedding models. It demonstrates strong performance on the MTEB leaderboard and we used `bge-base-en-v1.5` as our context refinement model.

In Table 9, we show the ODQA experiments of the FiD-PIT-Large model with different embedding models as the re-rankers. Interestingly, we can see that the embedding models affect single-hop and multi-hop performance differently. Contriever-MSMARCO and BGE show a more positive impact on the multi-hop questions whereas GTE has the best performance for single-hop temporal questions. In terms of overall performance, BGE is the best among all the sentence embedding models.

	Set Acc.	**Ans. F1**	**EM**	**Tok. F1**
T5-B-FiD†	-	-	10.3	19.7
T5-B	32.9	34.9	37.3	46.8
T5-B-PIT	34.2	36.4	39.0	48.4
T5-B-FiD	39.4	41.7	44.3	53.2
T5-B-FiD-PIT	41.1	43.3	46.0	54.7
T5-L-FiD	45.1	47.6	50.5	59.8
T5-L-FiD-PIT	47.3	49.8	52.7	61.0

Table 11: Experiments on fine-tuning on TimeQA-Hard. We follow the default OBQA setting of their paper. Results with † are taken from Chen et al. (2021).

D TimeQA Experiments

In this section, we evaluate the **PIT** strategy on prior temporal QA datasets. We used **T5-B-PIT** as initialization, and then fine-tuned on the relevant task datasets for the experiments in this section. In Table 11, we show the experimental results of **PIT** on TimeQA (Chen et al., 2021). We only conducted experiments on the Hard subset since the temporal reasoning involved in the Easy subset is too simple. Fusion-in-Decoder (FiD) (Izacard and Grave, 2021) was used as the baseline method for TimeQA. Our re-implemented FiD baseline achieved significantly higher results compared to the results of Chen et al. (2021). We can see that **PIT** is able to improve the performance of both conventional T5-SFT and the more advanced FiD model. This result indicates that the improved temporal reasoning capability of PIT can be transferred to the downstream open-book QA setting, and confirms the generalizability of the PIT method.

Example 1
Error Cause: Misunderstanding of temporal overlap.
Question: Which employer did Hans Kramers work for in September 1931?
Context: Hans Kramers worked for:
Leiden University from January 1934 to January 1952.
Utrecht University from January 1926 to January 1934.
Delft University of Technology from January 1931 to January 1952.
......
GPT-4's Prediction: **Delft University of Technology** and **Leiden University**
Ground Truth: **Delft University of Technology** and **Utrecht University**

Example 2
Error Cause: Misunderstanding of temporal containment.
Question: Which employer did Elon Musk work for 3 years and 6 months before he/she was living in Boca Chica (Texas)?
Context: Elon Musk worked for:
OpenAI from December 2015 to January 2019.
SpaceX from June 2002 to Oct 2023.
Neuralink from July 2016 to Oct 2023.
The Boring Company from December 2016 to Oct 2023.
Tesla Inc. from April 2004 to Oct 2023.
......
Elon Musk lived in:
Boca Chica (Texas) from June 2021 to Oct 2023.
......
GPT-4's Prediction: **The Boring Company** and **Neuralink**
Ground Truth: **The Boring Company** and **Neuralink** and **OpenAI** and **Tesla Inc.** and **SpaceX.**

Figure 2: Examples of GPT-4's erroneous temporal reasoning in the ReasonQA setting.

E Error Analysis

In this section, we aim to analyze some mistakes by LLMs in temporal reasoning. We mainly investigate the errors made by GPT-4, because this model has demonstrated excellent performance on various professional and academic benchmarks. We find that GPT-4 still makes mistakes in temporal reasoning. In Figure 2, we can see that in Sept. 1931, Hans Kramers was working for both Delft University of Technology and Utrecht University (January 1926 – January 1934). GPT-4 did a good job of finding the answer Delft University of Technology, but it failed to find the other answer Utrecht University and instead gave the wrong answer Leiden University. This shows that in the multi-answer scenario, GPT-4 can find a good answer but struggles to find all answers. This can also be seen from Table 7, where GPT-4 has a significantly higher answer-F1 score and a much lower set accuracy score for multi-answer questions.

For the second example in Figure 2, we need to first find the starting time of Elon Musk living in Boca Chica (June 2021) and perform time deduction with respect to that time point. The inferred

Property	Type	Template
P54 *member of sports team*	L2 M-hop	Which team did **subject** play for from t_1 to t_2?
P39 *position held*	L2 M-hop	Which position did **subject** hold Δt before t_1
P108 *employer*	L3 M-hop	Which employer did **subject** work for Δt after he/she studied at **object?**
P102 *member of political party*	L2 1-hop	Which political party did **subject** belong to in t_1?
P286 *head coach*	L2 M-hop	Who was the head coach of **subject** from $t1$ to $t2$?
P69 *educated at*	L3 M-hop	Where was **subject** educated when he/she was living in **object**?
P488 *chairperson*	L2 M-hop	Who was the chair of **subject** Δt before t_1?
P6 *head of government*	L2 M-hop	Who was the head of **subject** from t_1 to t_2?
P35 *head of state*	L3 1-hop	Who was the head of state of **subject** after **object**'s term of head of state?
P127 *owned by*	L3 M-hop	Who was the owner of **subject** when **object** was the chair?
P26 *spouse*	L3 M-hop	Which team did **subject** play for when he/she was married to **object**
P166 *award received*	L3 M-hop	Which award did **subject** receive when he/she was working for **object**?
P937 *work location*	L3 M-hop	Where did **subject** work when he/she was married to **object**
P551 *residence*	L2 1-Hop	What was the residence of **subject** in t_1

Table 12: Example templates for Wikidata properties for our Complex-TR dataset.

time of interest is December 2017. GPT-4 was only able to determine that Elon Musk worked for the Boring Company and Neurallink, perhaps because these starting times are closer to December 2017. On a longer time horizon, Musk has been working for Tesla and SpaceX since the early 2000s, but the model failed to include these companies in its answers.

F Question Templates

In this section, we show examples of our templates to create the Complex-TR dataset. We used 14 temporally related properties in the Wikidata KB. An example template for each property is shown in Table 12.

G Examples of Fictional Entities

In Table 13, we show some example fictional entities that we used to construct the pseudo-data for **PIT**. We also show an actual group of fictional data in Table 14.

H Annotation Interface

To enhance the accessibility and clarity of the human verification process, we hosted a user-friendly interface on Heroku[7]. Our interface was built on a popular open-source data annotation Github repository named doccano[8] (Nakayama et al., 2018). A screenshot of the annotation interface is given in Figure 3.

[7] https://www.heroku.com

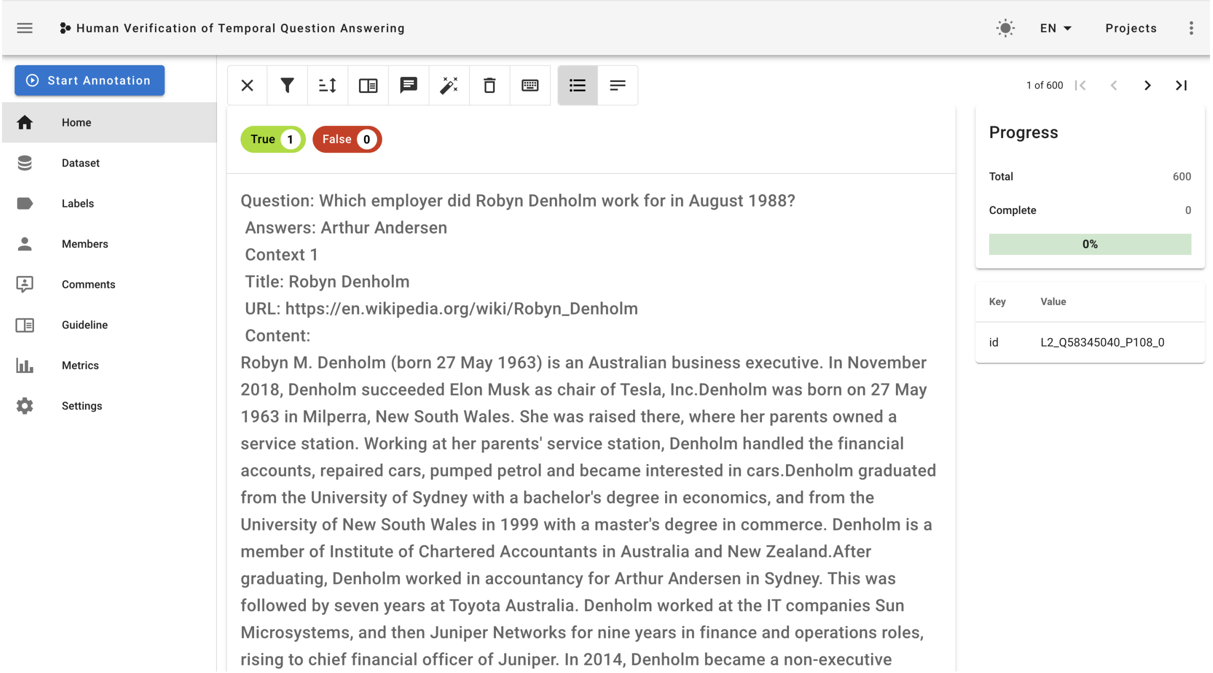
[8] https://github.com/doccano/doccano

Types	Number	Examples
countries	100	Unatin, Lislands Ofnited, Dencuslandsand, Djisvalwan, New Saintco Moazer, Nuazbe
companies	240	BrightBoost, AzureAlly, VitalVisionary, LuminaryLogic, NightOwl, AquaAdventures
teams	152	Polar Bears, Ice Warriors, Ice Breakers, Polar Storm, Arctic Foxes, Blizzard, Snow Leopards
towns	126	Fluoriteville, Galenaville, Heliodorhill, Iolitetown, Jadebrook, Kyaniteville, Labradoritehill
people	3,000	Angelina Romito, Matthew Thompson, Clifford Jump, Barbara Martinez, Martin Dudley, Joseph Parker, Harry Hatch, Richard Driskell, Catherine Scianna
schools	224	Yellowwood College, Azura University, Bluebell College, Cactus University, Daybreak College
awards	86	Masterful Memoirist Medal, Stellar Science Fiction Story Award

Table 13: Examples of fictional entities used for pseudo-instruction tuning (PIT). The fictional names are obtained by conversation with the free version of ChatGPT.

ReasonQA Context	Type	Example Question	Answers
Mary Bartlebaugh studied at Quartz College from May 1872 to May 1874.	L3 M-hop	Which employers did Mary Bartlebaugh work for when he/she was studying at Yam University?	Synergy Dynamics
Mary Bartlebaugh worked for Synergy Dynamics from May 1869 to May 1872.	L2 M-hop	Which employer did work for Mary Bartlebaugh from Oct 1888 to June 1897	Solaris Solutions
Mary Bartlebaugh studied at Yam University from May 1867 to May 1871.	L3 M-hop	Where was Mary Bartlebaugh educated when he/she was working for Synergy Dynamics?	Yam University
Mary Bartlebaugh worked for Solaris Solutions from May 1887 to May 1899.	L2 1-hop	Where was Mary Bartlebaugh educated at in June 1873?	Quartz College

Table 14: A pseudo-data example, where the subject is a fictional name, **Mary Bartlebaugh**.



id L2_Q58345040_P108_0

Question: Which employer did Robyn Denholm work for in August 1988?
Answers: Arthur Andersen
Context 1
Title: Robyn Denholm
URL: https://en.wikipedia.org/wiki/Robyn_Denholm
Content:
Robyn M. Denholm (born 27 May 1963) is an Australian business executive. In November 2018, Denholm succeeded Elon Musk as chair of Tesla, Inc.Denholm was born on 27 May 1963 in Milperra, New South Wales. She was raised there, where her parents owned a service station. Working at her parents' service station, Denholm handled the financial accounts, repaired cars, pumped petrol and became interested in cars.Denholm graduated from the University of Sydney with a bachelor's degree in economics, and from the University of New South Wales in 1999 with a master's degree in commerce. Denholm is a member of Institute of Chartered Accountants in Australia and New Zealand.After graduating, Denholm worked in accountancy for Arthur Andersen in Sydney. This was followed by seven years at Toyota Australia. Denholm worked at the IT companies Sun Microsystems, and then Juniper Networks for nine years in finance and operations roles, rising to chief financial officer of Juniper. In 2014, Denholm became a non-executive

Figure 3: Annotation interface for the human verification process. Annotators are only asked to give True or False labels to the QA pairs and their contexts.