MulTiple: A Multi-answer Time-sensitive Complex Question Answering Dataset

Anonymous ACL submission

Abstract

Time-sensitive question answering is to answer questions with specific timestamps from the given long document. Existing works mostly focus on only one of the high-quality answers, but it is common that multiple answers simultaneously satisfy the constraints of a specific timestamp in the time-sensitive question. For example, an individual may play two different roles during a specific timestamp. In this paper, we construct a Multi-answer Time-sensitive question answering dataset, MulTiple, consisting of 17,580 multi-answer instances. Each contains a question, a corresponding long document and multiple answers. To ensure that the generated questions have multiple answers, we propose a global iteration method to obtain time-evolving events with multiple objects for the same subject and relation. Moreover, the baseline model IterBird is proposed to progressively gain multiple answers by integrating an iterative mechanism with the single-answer model. We construct extensive experiments on MulTiple and results show that IterBird significantly outperforms other baselines with SEM scores of 25.65% and 22.69%. It demonstrates that existing models struggle to obtain the full answers, even as clue words are provided in the time-sensitive questions. The dataset and code are released in http://github.com/multipledata/MTQA.

1 Introduction

Time is universally acknowledged as a pivotal factor affecting people's work, daily routines, and social activities in the real world. According to statistics (Chia et al., 2022), time-related qualifiers account for 48% in the widely used knowledge base WiKidata (Vrandecic and Krötzsch, 2014). Recently, several time-sensitive question answering datasets and models have been proposed, as it has drawn increasing interest over the past few years (Jia et al., 2018, 2021; Chen et al., 2021, 2022; Tan et al., 2023). They have in common that only a high-quality answer is provided for each time-sensitive question. However, is a single answer enough? As shown in Figure 1, it is notably insufficient to answer "*member of the European Parliament*" or "*rapporteur on Corporate Governance*" for the time-sensitive question "*What was the position of Ashley Fox from 2011 to 2012?*". Therefore, it is valuable to encourage models to provide an appropriate number of answers.

Figure 1: An Example of multi-answer time-sensitive question and answers pair on MulTiple, where question consists of two categories: question-dependent (including club words) and document-dependent.

Indeed, a few datasets contain multi-answer questions, such as TimeQA (Chen et al., 2021) and TempReason (Tan et al., 2023). However, models tend to focus on only one of the answers (Raffel et al., 2020; Izacard and Grave, 2021; Zhu et al.,

2023; Li et al., 2023) and overlook research on multi-answer questions. It could be attributed to the fact that (1) the number of multi-answer questions is not large enough; (2) models on existing datasets are still designed for single-answer time-sensitive questions with imbalanced proportions. For example, the amount (proportion) of multi-answer questions are 2667 (7%) and 373 (2.3%) in TimeQA (Chen et al., 2021) and TempReason (Tan et al., 2023) datasets, respectively. Therefore, it is imperative to construct a dataset for studying multi-answer time-sensitive questions.

In this paper, we introduce MulTiple, a multi-answer time-sensitive question answering dataset, including 17,580 multi-answer instances. And each contains a question, a corresponding long document and multiple answers, as shown in Figure 1. Specifically, MulTiple is constructed by four steps: data mining, data preprocessing, quality control and question generation. To ensure that the generated questions have multiple answers, we mine and reconstruct time-evolving events by considering temporal relations between events with the same subject and relation. Also, we create two visions of questions based on whether the generated question contained clue words, prompting the number of answers. Document-dependent questions are more challenging, as they have an uncertain number of answers. To build a more realistic and challenging variant of the dataset, we add a comparable number of single–answer and unanswerable to the MulTiple by the same preprocessing. The total size of the expanded dataset is 49046 instances, about 2.79 times the basic version.

Moreover, we also propose a novel baseline method, IterBird, to extract multiple answers iteratively based on the single-answer model BigBird. Experimental results demonstrate IterBird achieves the best performance in almost all baselines.

In a nutshell, our contributions are as follows:

- We construct a multi-answer dataset of time-sensitive question answering named MulTiple, consisting of 17,580 high-quality multi-answer instances. In doing so, we design a global iteration method to construct time-evolving events with multiple objects for the same subjects and relations.
- We propose a novel baseline method, IterBird, to obtain multiple answers by integrating an iterative mechanism with the basic single-answer model.
- We conduct extensive experiments on MulTiple and results show that existing models, including the large language models, struggle to obtain the all answers, even as clue words are provided in the time-sensitive questions.

2 Task Formulation

Multi-answer Time-sensitive Question Answering (MTQA) is defined to generate a set of answers $A = \{A_1, A_2, \ldots, A_k\}(k \geq 0)$ for the given time-sensitive question Q based on the given long document D, where k is the number of answers and each answer A_i often originates from the document D. Time-sensitive questions typically consist of the subject, relation, a certain timestamp and club words (optional), such as, "*What were the **three*** [Club word] ***positions of*** [Relation] ***Ashley Fox*** [subject] ***from 2011 to 2012*** [Timestamp]*?* ". The long document D describes the corresponding subject, which comprises various related relations and timestamps except for question mentions, as shown in the middle of Figure 1. Note that multiple answers are typically scattered across a long document and appear in diverse styles. For example, answers "*member of the European Parliament*", "*Chief Whip of the European Conservative & Reformists Group (ECR)*" and "*rapporteur on Corporate Governance*" are found in different sentences and even paragraphs, as shown in Figure 1.

3 Dataset Construction

The multi-answer time-sensitive question answering dataset is constructed by four steps: **Data Collection**, **Data Preprocessing**, **Quality Control** and **Question Generation**, as shown in Figure 2.

3.1 Data Collection

To obtain question-document pairs, we excavate temporal events and their corresponding documents from the widely used knowledge base Wikidata (Vrandecic and Krötzsch, 2014) and Wikipedia.

Mining temporal facts from Wikidata. We first utilize existing annotations to identify events over time and mine them by resorting to Wikidata. Followed by (Chen et al., 2021), we first mine time-evolving events with time quantifiers P580 (start time), P582 (end time) and P585 (point in time) and structure them in the form of quadruples {*subject, relation, object, timestamp*}, where timestamp includes two types of time point and time interval. Then, time-evolving events with the same subject

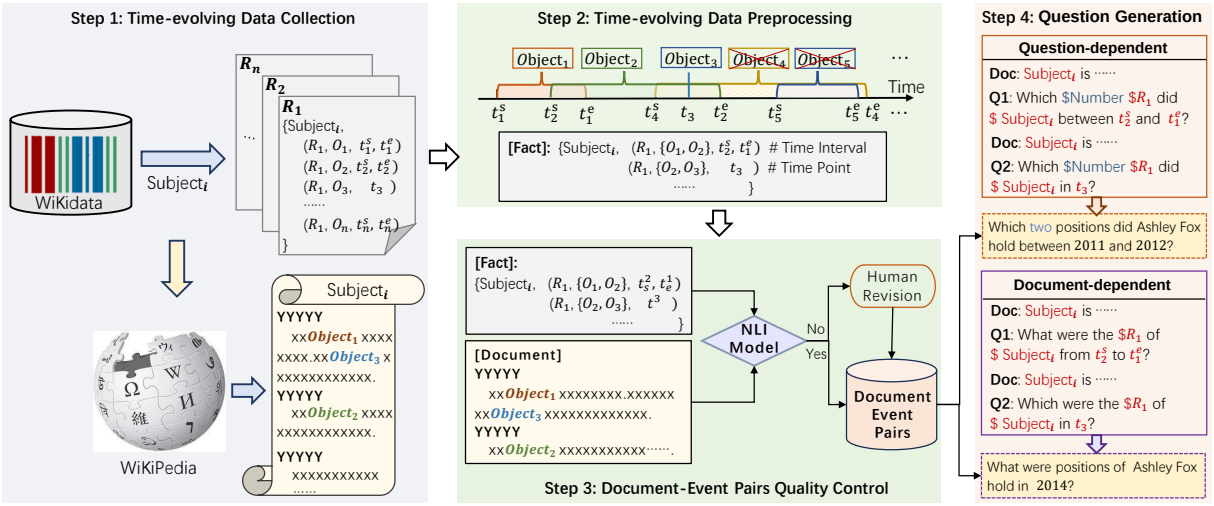


Figure 2: The overall framework for constructing dataset, MulTiple. It consists of four steps: Data Collection, Data Preprocessing, Quality Control and Question Generation.

and relation are arranged in chronological order and merged as temporal facts, formulated as:

$$(Sub, \{R_i, O_1, t_1\}, \{R_i, O_2, t_2\}, \ldots, \{R_i, O_n, t_n\})$$

where sub, R_i, O_j and t_j denote the subject and its i-th relation, object and timestamp.

We discarded some temporal facts and events that exist: 1) events with numeric objects, as these numerical events are less likely to appear in Wikipedia texts. 2) The timing of time-evolving events does not overlap at all, since the multi-answer phenomenon is less likely to occur in relations with non-overlapping times. 3) There is only one time-evolving event because a single event is unable to generate tuples with multiple objects. We have successfully mined roughly 180K time-evolving events and 36K temporal facts.

Mining long documents from Wikipedia. After getting those temporal facts, we require tracing back to the corresponding Wikipedia pages of subjects as their context. However, the directly mined documents often contain a lot of noise, since the two large knowledge bases are not a perfect match. We employ a distance supervision (Mintz et al., 2009) approach to determine whether the long document D contains relation R_i and objects $O_1, O_2, \ldots, O_n$ within the temporal fact $(Sub, R_i, \{O_1, t_1\}, \{O_2, t_2\}, \ldots, \{O_n, t_n\})$. We discard the instance if the relation R_i is not present in the document or there are fewer than two objects. Finally, we have successfully mined about 10K long documents.

3.2 Data Preprocessing

To ensure the generated time-sensitive questions with multiple answers, we construct multi-object time-evolving events based on the above coarse screening temporal facts. During this process, we merge multiple time-evolving events within the same temporal fact for creating multi-object quadruples. Similar to the original ones, they are categorized into two types: time intervals (t^s, t^e) and time points t (i.e., $t^s = t^e$), where t^s and t^e denote the start and end time, respectively. The following takes two time-evolving events $\{Sub, R_i, O_1, t_1\}$ and $\{Sub, R_i, O_2, t_2\}$ $(t_1^s \leq t_2^s)$ as an example to explain.

For multi-object time-evolving events with time points, it is constructed in two manners:

- If two events are $\{Sub, R_i, O_1, (t_1^s, t_1^e)\}$ and $(Sub, R_i, O_2, t_2\}$, we first determine the chronological relation between their timestamps. And when $t_1^s \leq t_2 \leq t_1^e$, a new quadruple $\{Sub, R_i, \{O_1, O_2\}, t_2\}$ is constructed.

- If two events are $\{Sub, R_i, O_1, (t_1^s, t_1^e)\}$ and $\{Sub, R_i, O_2, (t_2^s, t_2^e)\}$, the chronological relation is determined and the new quadruple $\{Sub, R_i, \{O_1, O_2\}, t_1^e\}$ is constructed when $t_1^e = t_2^s$.

For multi-object time-evolving events with time intervals, it also created in two manners:

- If two events are $\{Sub, R_i, O_1, (t_1^s, t_1^e)\}$ and $(Sub, R_i, O_2, t_2\}$, we start by reordering them according to the chronological relation between their timestamps. The new quadruple

[Document] Joseph Armitage Robinson (9 January 1858 - 7 May 1933) was a priest in the Church of England and scholar... Mary, Cambridge until 1886, then a Cambridge Whitehall preacher from 1886 to 1888. That year he was appointed examining chaplain to the Bishop of Bath and Wells and vicar of All Saints Church, Cambridge where he stayed from 1888 until 1892. He was also a dean of Christs College, Cambridge, from 1884 to 1890. In 1893, he was appointed Norrisian professor of Divinity at Cambridge University, serving as such until 1899, during which he was also a prebendary of Wells Cathedral. He served as rector of St Margarets, Westminster 1899-1900, and was appointed a canon of Westminster in 1899, serving until his appointment as dean. In 1902...		
Question-dependent	Times Point	(1) Which **four** positions did Armitage Robinson take *in 1888*? (2) Which **two** positions did Armitage Robinson take *in 1896*?
	Times Interval	(1) Which **two** positions were occupied by Armitage Robinson *between 1886 and 1888*? (2) *Before 1892*, which **four** positions had Armitage Robinson held?
Document-dependent	Times Point	(1) Which roles did Armitage Robinson occupy *in 1895*? (2) *In 1899*, what positions did Armitage Robinson hold?
	Times Interval	(1) During the period *from 1893 to 1899*, which positions did Armitage Robinson occupy? (2) *After 1899*, which positions did Armitage Robinson assume?

Table 1: Examples of multi-answer time-sensitive questions on MulTiple. It consists of two separate visions of the question based on the presence of clue words: Question-dependent and Document-dependent.

$\{Sub, R_i, \{O_1, O_2\}, \{t_1^s, t_1^e\}\}$ is constructed when it satisfies that $t_1^s \leq t_2 \leq t_1^e$.

- If two events are $\{Sub, R_i, O_1, (t_1^s, t_1^e)\}$ and $\{Sub, R_i, O_2, (t_2^s, t_2^e)\}$, it further divides into two distinct scenarios after reordering: (1) When $t_1^s \leq t_2^s < t_2^e \leq t_1^e$, a new quadruple $\{Sub, R_i, \{O_1, O_2\}, (t_2^s, t_2^e)\}$ is constructed; (2) When $t_1^s < t_2^s < t_1^e < t_2^e$, a new quadruple $\{Sub, R_i, \{O_1, O_2\}, (t_2^s, t_1^e)\}$ is constructed.

Regarding the constructed quadruples with two objects as new time-evolving events, and repeat the above step until there are no more objects to add. In addition, we again correct these objects within time-evolving events based on the corresponding document. We have successfully obtained roughly 4K multi-object events.

3.3 Quality Control

With the above steps, there is hardly an issue that objects in multi-object time-evolving events do not exist in the corresponding documents. However, it still cannot guarantee whether the document entails the multi-object time-evolving event. In this paper, we conceptualize it as the Natural Language Inference (NLI) task, followed by (Yue et al., 2023). To do it, we adopt XLM-RoBERTa (Conneau et al., 2020) fine-tuned on large-scale corpora as the NLI model, which is good at handling the long-short text NLI task (Cabot and Navigli, 2021). Specifically, we regard the document as the premise and multi-object event as a hypothesis. If given the premise, assuming the hypothesis is true, we consider that the document contains the multi-object event; otherwise, it does not hold. Further, we manually annotate these false instances by employing workers who have fully understood the annotation principles and passed the preliminary examination.

During this process, about 37% of instances are manually revised and more details are in Appendix A.3. Among these adjustments, 76.4% of objects are corrected by boundary retuning and expression rewriting, 12.8% are removed due to semantic mismatches and 9.2% of documents are replenished. After filtering, we obtain 3000 golden document-event pairs as the final release. The rectified multi-object time-evolving events involve 46 different relations, such as 'position', 'play for', etc.

3.4 Question Generation

The procedure is to transform the multi-object time-evolving event into the time-sensitive question Q and a set of answers $A = \{A_1, A_2, \ldots, A_n\}$. Specifically, we regarded the subject, relation and timestamps in the time-evolving event as the source for generating questions. Objects and the document are viewed as the set of answers and context.

Followed by previous works (Chen et al., 2021; Tan et al., 2023), we initially defined several common time quantifiers, such as 'in', 'between', 'before' and "after", and create 4-7 different templates for each relation, as shown in the right of Figure 2.

- For multi-object events with time points $\{Sub, R_i, \{O_1, O_2, \ldots, O_m\}, t\}$, we directly insert the subject, relation, timestamp and time quantifier into the placeholder of question templates, such as "Which roles did Armitage Robinson occupy [in] 1895?".

Split	Question Type	#Questions	#Documents	$\frac{\#\text{Questions}}{\#\text{Documents}}$	$\frac{\#\text{Answers}}{\#\text{Questions}}$	Distance	#Doc-Token	# Que-Token
Train	Question-dependent	12703	2142	5.93	2.12	203.4	1559.9	11.9
	Document-dependent	12807	2142	5.98	2.12	205.4	1564.3	10.7
Dev	Question-dependent	2405	425	5.66	2.14	171.9	1513.8	11.8
	Document-dependent	2406	425	5.66	2.14	168.9	1527.2	10.7
Test	Question-dependent	2346	424	5.53	2.16	294.9	1449.6	11.8
	Document-dependent	2367	424	5.58	2.16	307.3	1457.8	10.8
All	Question-dependent	17454	2991	5.64	2.13	220.9	1540.8	11.9
	Document-dependent	17580	2991	5.88	2.13	227.5	1543.9	10.7

Table 2: Statistics of MulTiple. 'Distance' denotes the average distance between adjacent answers. #Doc-Token and #Que-Token are the average number of tokens in long documents and questions, respectively.

- For multi-object events with time intervals $\{Sub, R_i, \{O_1, O_2, \ldots, O_m\}, (t^s, t^e)\}$, we first randomly select one of the time quantifiers and then determine one or two times following the guidelines below:
  (1) If the time quantifier is 'in', we randomly select the time point $\bar{t}$ within the interval (t^s, t^e) and synthesize timestamp as "in $\bar{t}$".
  (2) If the time quantifier is 'before (after)', we opt for the end time t^e (start time t^s) and generate timestamp as "before t^e (after t^s)".
  (3) If the time quantifier is 'between', we take both start and end times to compose the timestamp as "between t^s and t^e".

We observed that there is a significant disparity in the distance between the start and end times. Consequently, we introduce the time quantifier "between-subset". In this case, we randomly select a starting time t_q^s within the time interval (t^s, t^e) and then randomly choose an end time t_q^e from the resulting time interval (t_q^s, t^e) to generate the timestamp as "between t_q^s and t_q^e".

Further, we create two versions with different levels of difficulty to explore the effect of uncertainty in the number of answers for multi-answer time-sensitive questions. They differ in whether the generated questions contain clue words, prompting the number of answers. Hence, we write question templates of question-dependent questions by adding clue words, obtained by the number of objects in the multi-object events.

3.5 Dataset Analysis

To understand the properties of MulTiple, we analyze it from basic statistics and extension.

Document-dependent and question-dependent datasets contain 17,454 and 17,580 samples, where the average token lengths of questions and contexts are 11.9 and 1,540.8, 10.7 and 1,543.9, respectively. More statistics are reported in Table 2.

We added a substantial number of single-answer question-answer pairs and unanswerable instances (the answer is not in the document) to the Multiple, since they also exist in real-world QA scenarios. The extended dataset comprises a total of 97,956 instances and is divided into document-dependent and question-dependent versions, following a distribution similar to the base version. For more dataset details, please refer to the Appendix A.4.

4 Proposed Model

In this section, we propose a stronger baseline method, IterBird, to obtain multiple answers utilizing the iterative mechanism.

Specifically, we choose BigBird (Zaheer et al., 2020) as the basic model, which extracts answers by predicting start and end positions from the given long document. Firstly, the input sequence $X = (q_1, q_2, \cdots, q_M, [SEP], d_1, d_2, \cdots, d_N)$ is the concatenation of the question Q and document D. Since the given document is long, the input sequence easily exceeds 4K tokens. Therefore, a more generalized attention mechanism is used to obtain the top-level representation $R_X \in \mathbb{R}^{(N+M)\times D}$, where D denotes the hidden dimension. We also project R_X to $p_s \in \mathbb{R}^{N+M}$ and $p_e \in \mathbb{R}^{N+M}$, calculated as follows:

$$\begin{aligned} p_s &= \text{softmax}(squeeze(R_X \cdot W_s)); \\ p_e &= \text{softmax}(squeeze(R_X \cdot W_e)) \end{aligned} \quad (1)$$

where $W_s, W_e \in \mathbb{R}^{D\times 1}$ are learnable matrices.

During the inference process, we select $i, j = \text{argmax}_{i,j}(p_s(i) \times p_e(j))$ as the start and end position of the prediction span. In addition, we adopt an iterative strategy to predict multiple positions of starting and ending for multiple answers, inspired by (Zhang et al., 2023). In each iteration, we append the previously extracted answers to the ques-

Methods	Question-dependent MTQA						Document-dependent MTQA					
	Dev			Test			Dev			Test		
	SEM	EM	PM	SEM	EM	PM	SEM	EM	PM	SEM	EM	PM
Large Language Models												
Llama2-7b (Touvron et al., 2023)	13.50	33.45	44.65	12.62	30.48	43.76	10.37	30.34	41.01	8.33	26.13	35.51
Llama2-13b (Touvron et al., 2023)	17.64	40.21	52.39	16.38	38.25	50.06	14.07	39.26	52.84	12.86	40.64	50.63
ChatGPT (Ouyang et al., 2022)	21.50	42.45	58.61	20.00	40.74	55.98	14.88	45.33	60.81	12.50	43.32	56.99
QaAp (Zhu et al., 2023)	24.50	47.87	65.55	22.50	47.18	66.87	17.50	46.33	64.33	14.00	45.91	64.13
ReAct (Yao et al., 2023)	22.00	44.81	62.04	20.50	45.98	65.01	16.87	46.91	60.24	13.83	44.89	64.53
GPT-4 (OpenAI, 2023)	*28.50*	*52.63*	*70.01*	*26.00*	*49.83*	*72.19*	*22.44*	*52.70*	*67.77*	*17.00*	*52.72*	*66.70*
Pre-trained Language Models												
Li et al.[BERT](Li et al., 2022)	14.59	42.85	66.22	10.70	38.31	58.04	11.89	39.96	57.28	10.35	36.60	55.15
MTMSN[BERT] (Hu et al., 2019)	15.22	37.90	57.91	10.53	34.34	55.24	14.96	38.61	57.11	9.51	34.25	54.13
TASE[BERT$_{LARGE}$] (Segal et al., 2020)	17.22	49.15	64.10	14.95	46.36	63.62	16.69	46.24	61.31	11.81	42.44	59.23
ITERATIVE[RoBERTa] (Zhang et al., 2023)	23.87	50.49	66.87	21.27	47.95	64.93	19.29	46.69	63.83	16.34	43.10	61.18
T5 (Raffel et al., 2020)	25.91	50.04	67.21	21.06	45.17	63.84	**25.69**	50.73	66.81	21.63	45.97	64.10
FiD (Izacard and Grave, 2021)	23.78	50.60	67.86	19.44	45.87	64.21	23.94	50.58	67.51	20.79	45.88	64.59
REMEMO (Yang et al., 2023)	14.92	31.78	53.23	11.44	30.33	50.67	13.31	33.25	52.67	10.34	29.21	52.22
IterBird (Ours)	**29.56**	**54.79**	**70.26**	**25.65**	**52.41**	**68.34**	25.64	**52.91**	**68.41**	**22.69**	**49.36**	**66.50**
Human	-	-	-	85.14	89.71	94.03	-	-	-	81.43	84.64	90.05

Table 3: Results on MulTiple, including Question-dependent and Document-dependent MTQA. The **Best** results of fine-turn PLMs are highlighted in bold, and the *Best* results of LLMs are labelled underlined.

tion with the word 'except' in the middle and then feed the updated question into the single-answer MTQA model. The iterative process terminates when the model predicts no more answers.

5 Experiments

In this section, we construct and analyze extensive experiments on our proposed dataset, MulTiple.

5.1 Baselines

We implement multiple baselines to provide benchmark performances, which can be divided into two categories: For the Pre-trained Models, we selected seven models for targeted adaptation to multiple answers, including four multi-span question-answering models, Li et.(Li et al., 2022), MTMSN(Hu et al., 2019), TASE(Segal et al., 2020), and ITERATIVE(Zhang et al., 2023), and three temporal question-answering models, T5(Raffel et al., 2020), FiD(Izacard and Grave, 2021), and REMEMO(Yang et al., 2023); We also selected various popular Large Language Models as base models to obtain multiple answers by Prompts, including the Llama2(Touvron et al., 2023) and GPT(Ouyang et al., 2022; OpenAI, 2023) families. More details about the baseline are given in the Appendix B. In addition, a manual evaluation was conducted to observe the best human performance, and the manual evaluation approach is described in the AppendixC.

5.2 Main Results

Table 3 shows the results for all baselines in our two versions of MulTiple and Table 4 illustrates further experiments on MulTiple(expand) with several baselines that worked better. IterBird achieves SEM, EM and PM of 25.65, 52.41, and 68.34 on question-dependent MTQA and 22.69, 49.36 and 66.50 on document-dependent MTQA, which exceeds almost all baselines. It demonstrates that IterBird is effective by iterative mechanism, especially for question-dependent mode. But it still has a long way from human evaluation. Then, GPT-4 achieves the best performance among large language models, achieving competitive performance. For almost all baselines the value of EM is much larger than SEM whether on question-dependent or document-dependent questions. It demonstrates that existing methods, including IterBird and GPT-4, struggle to obtain the all answers, even as clue words are provided in the time-sensitive questions.

In addition, there is essentially the same trend in the MulTiple (expand). The difference is that the IterBird does not perform best for all evaluation metrics, as shown in Table 4. It demonstrates that IterBird is not always optimized to answer multi-answer questions when they are unbalanced.

5.3 Analysis

To take a deep look into the proposed datasets, we further analyze the performance of models from three different perspectives.

Question-dependent & Document-dependent MTQA. Question-dependent and document-dependent questions are distinguished based on whether the question contains clue words or not, as shown in Table 1. Most models perform slightly lower on document-dependent questions compared

Methods	Overall			Single			Multiple		
	SEM	EM	PM	SEM	EM	PM	SEM	EM	PM
Question-dependent MTQA									
Llam2-13b (Touvron et al., 2023)	30.11	35.56	55.93	32.98	33.87	56.67	15.76	39.96	50.16
ChatGPT (Ouyang et al., 2022)	34.96	44.76	58.50	36.60	36.16	59.28	20.58	43.93	56.02
GPT-4 (OpenAI, 2023)	42.54	51.26	68.06	51.81	51.74	68.95	25.98	50.06	**71.54**
ITERATIVE[RoBERTa] (Zhang et al., 2023)	40.27	48.59	*65.99*	50.34	50.34	67.89	20.71	46.90	64.17
T5 (Raffel et al., 2020)	40.68	47.97	62.24	47.34	47.30	58.37	25.36	48.72	66.49
FiD (Izacard and Grave, 2021)	*41.68*	*48.84*	60.37	50.16	50.41	55.64	22.17	47.38	65.57
IterBird (Ours)	**44.65**	**53.28**	**69.43**	**54.24**	**54.24**	**70.54**	**26.04**	**52.37**	68.38
Document-dependent MTQA									
Llama2-13b (Touvron et al., 2023)	28.38	32.80	52.90	31.02	31.79	53.31	12.25	38.40	50.29
ChatGPT (Ouyang et al., 2022)	30.03	43.41	56.57	34.91	35.16	57.60	13.87	43.36	56.59
GPT-4 (OpenAI, 2023)	38.15	49.16	65.15	49.01	51.16	68.22	17.19	**48.17**	63.73
ITERATIVE[RoBERTa] (Zhang et al., 2023)	37.63	*47.46*	*64.56*	48.85	50.85	68.94	16.74	45.81	62.25
T5 (Raffel et al., 2020)	35.96	46.15	60.05	43.66	45.71	57.57	**18.22**	46.72	63.47
FiD (Izacard and Grave, 2021)	38.10	46.82	*55.38*	**49.67**	50.13	53.79	11.39	42.42	57.83
IterBird (Ours)	**38.51**	**50.26**	**66.96**	49.65	**52.45**	**69.14**	16.78	47.83	**64.89**

Table 4: Performance of Baselines on MulTiple (expand), consisting of overall, single- and multi-answer questions.

to question-independent questions, especially for MTMSN and TASE. It suggests that these models are more sensitive to the number of answers and better suited for answering questions with a specified number of answers. From the experiments on MulTiple (expand), it can be observed that the variations between the two types of questions are not significant for single-answer questions, especially in terms of EM scores. It indicates that baseline models tend to excel in considering the number of answers as one. The reason could be the lack of multi-answer datasets available so far.

Analysis of MulTiple & MulTiple (expand). MulTiple (expand) is an expanded dataset based on MulTiple, with proportionally in the number of single-answer and unanswered questions, as described in Section 3.5. As shown in Figure 4, almost all baseline models exhibit a notable improvement in terms of SEM compared to the MulTiple, ranging from 13.73% to 22.24%. It suggests that existing models are better at learning single-answer questions and struggle to strictly match multi-answer questions. However, nearly all baselines show a decrease in terms of PM, measuring the overlap between the predictions and ground truth answers at the token level. In addition, EM scores are relatively comparable on both datasets compared to the other two metrics. It implies that the models face similar learning difficulties across both datasets. Overall, the three levels of indicators exhibit different trends after the inclusion of single-answer questions.

Single-answer & Multi-answer Questions. To further evaluate the performance of models on single- and multi-answer questions, we have conducted fine-grained experiments on MulTiple (expand), as shown in Table 4. It is observed that the performance of SEM degrades rapidly compared with EM for multi-answer questions, while the performance of single-answer questions is quite distributed across two matrices. It confirms the consistency with our metric calculation in Appendix D, where SEM and EM are equal for single-answer questions. In terms of PM, the performance of single-answer questions is generally lower than that of multi-answer questions. We conjecture that labels with multiple answers have more tokens, and a greater number of those tokens are correctly predicted. In addition, the performance of EM varies depending on the specific model.

5.4 Error Analysis

In this section, we analyze error cases predicted by the best model, IterBird, in the test set and analyze the challenges of MTQA. As shown in Figure 3, there are three examples incorrectly predicted by the model, including (1) Answer missing. It denotes that predicted answers are correct, but its number is less than the number of ground-truths. (2) Partial error. It means that there is a partial intersection between the set of predicted answers and the set of ground-truths, even though their predictions are imperfectly correct. (3) Complete error. It represents that the model does not get any correct answer. The percentage of these cases in the error samples is shown in Figure 4.

Through the above analysis, we further summarize several specific challenges for MTQA. 1) **Un-**

Answer Missing

Context: she played ... the United States womens national soccer team from 1987 to 2004 ... she played for the Washington Freedom from 2001 to 2003.

Question: Which team did the player Mia Hamm belong to in 2001?

Prediction: Washington Freedom **Ground Truth:** United States womens national soccer team; Washington Freedom

Partial Errors

Context: Robinson was Lord High Almoner from 1906 to 1933…He served at Westminster until he was appointed Dean of Wells in 1911 …

Question: Which position did Armitage Robinson hold from 1911 to 1916?

Prediction: Dean of Westminster; Lord High Almoner **Ground Truth:** Lord High Almoner ; Dean of Wells

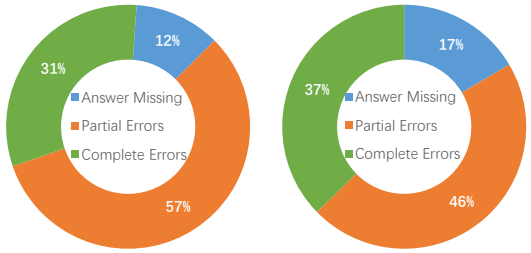
Complete Errors

Context: Between 1999 and 2000 he was … in Maastricht … In 2001 he participated … in Amsterdam … finishing in 2002. In 2004 , he was made Cocheme Fellow at Byam Shaw School of Art, London…

Question: Ryan Gander worked in which location before 2004?

Prediction: London **Ground Truth:** Maastricht; Amsterdam

Figure 3: Error cases. The orange and blue boxes denote wrong and correct answers, respectively.



(a) Question-dependent (b) Document-dependent

Figure 4: Proportion of error types on both datasets.

certainty in the number of answers. The number of answers is variable for multi-answer tasks, but existing multi-answer models almost struggle to obtain full answers. 2) **Long distance between answers**. Answers are typically scattered throughout a long document in diverse ways.

6 Related Works

6.1 Time-sensitive Question Answering

Time-sensitive Question Answering aims to answer questions with time qualifiers based on the given document(Chen et al., 2021), which is crucial for language models to be successful in real-world applications (Tan et al., 2023). Recently, several time-sensitive question answering datasets have been proposed to focus on the temporal inference of time-based QA tasks, such as TimeQA (Chen et al., 2021), MenatQA (Wei et al., 2023), and TempReason (Tan et al., 2023). And some recent works (Mathur et al., 2022; Yang et al., 2023; Su et al., 2023) achieve state-of-the-art levels by employing graph structures to capture temporal relationships between contexts. In addition, LLMs have been used for answering temporal questions by utilizing their advanced event extraction capabilities (Li et al., 2023; Zhu et al., 2023).

Different from all the above works, we construct MulTiple, a new dataset that focuses on time-sensitive questions with multiple answers.

6.2 Multi-answer MRC

The multi-answer phenomenon refers to the fact that a question may have multiple answers scattered throughout a document(Bai et al., 2023). It has been focused primarily on Machine Reading Comprehension (MRC) task (Zhang et al., 2023; Li et al., 2022). These efforts are approached from two perspectives: datasets and models. Specifically, (Yue et al., 2023) proposes MA-MRC, a high-quality multiple-answer MRC dataset, in which each sample contains a question, corresponding document and multiple answers. In terms of models, (Segal et al., 2020) casts question answering as a sequence tagging task to predict whether each token is part of the answer. MTMSN (Hu et al., 2019) combines a multi-type answer predictor with a multi-span extraction method for dynamically producing one or multiple text spans.

Different from multi-answer MRC, multiple-answer time-sensitive question answering requires inferring objects corresponding to a specific timestamp for the mentioned subject and relation within the given question. These objects are often scattered throughout the given long document and appear in diverse styles, posing a greater challenge.

7 Conclusion

In this paper, we construct the first multi-answer dataset of time-sensitive question answering, MulTiple, which is critical for evaluating whether the model thoroughly understands temporal concepts. We also propose a baseline model, IterBird, to extract multiple answers by integrating an iterative mechanism with the single-answer model. In addition, a series of models are implemented on MulTiple through targeted adaptation of existing methods. Experiments demonstrate that IterBird significantly outperforms other baselines, and existing models still struggle to obtain the full multiple answers, even as clue words are provided in the time-sensitive question. Therefore, we believe MulTiple could serve as a valuable benchmark in studying Multi-answer questions. Though our dataset is high-quality and large-scale, the type of answers and relations are not sufficiently diverse and balanced due to the limitations of the data source. In the future, we would like to sample more diverse data and control its quality through LLMs.

References

Yang Bai, Anthony Colas, and Daisy Zhe Wang. 2023. Mythqa: Query-based large-scale check-worthy claim detection through multi-answer open-domain question answering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 3017–3026, Taipei, Taiwan. ACM.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, (NeurIPS Datasets and Benchmarks)*.

Ziyang Chen, Xiang Zhao, Jinzhi Liao, Xinyi Li, and Evangelos Kanoulas. 2022. Temporal knowledge graph question answering via subgraph reasoning. *Knowledge-Based Systems*, 251:109134.

Yew Ken Chia, Lidong Bing, Sharifah Mahani Aljunied, Luo Si, and Soujanya Poria. 2022. A dataset for hyper-relational extraction and a cube-filling approach. In *Proceedings of the 24th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10114–10133.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, Online. Association for Computational Linguistics.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1596–1606, Hong Kong, China. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL)*, pages 874–880, Online. Association for Computational Linguistics.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Proceedings of the 37th International World Wide Web Conferences (WWW)*, pages 1057–1062, Lyon, France. ACM.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 792–802, Queensland, Australia. ACM.

Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022. Multispanqa: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 1250–1260, Seattle, WA, United States. Association for Computational Linguistics.

Xingxuan Li, Liying Cheng, Qingyu Tan, Hwee Tou Ng, Shafiq Joty, and Lidong Bing. 2023. Unlocking temporal question answering for large language models using code execution. *CoRR*, abs/2305.15014.

Puneet Mathur, Vlad I. Morariu, Verena Kaynig-Fittkau, Jiuxiang Gu, Franck Dernoncourt, Quan Hung Tran, Ani Nenkova, Dinesh Manocha, and Rajiv Jain. 2022. Doctime: A document-level temporal dependency graph parser. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 993–1009. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Singapore. The Association for Computer Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2020. A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3074–3080, Online. Association for Computational Linguistics.

Xin Su, Phillip Howard, Nagib Hakim, and Steven Bethard. 2023. Fusing temporal graphs into transformers for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 948–966.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, and Peter Albert. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. *CoRR*, abs/2310.05157.

Sen Yang, Xin Li, Lidong Bing, and Wai Lam. 2023. Once upon a time in graph: Relative-time pretraining for complex temporal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11879–11895. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, Kigali, Rwanda. OpenReview.net.

Zhiang Yue, Jingping Liu, Cong Zhang, Chao Wang, Haiyun Jiang, Yue Zhang, Xianyang Tian, Zhedong Cen, Yanghua Xiao, and Tong Ruan. 2023. MA-MRC: A multi-answer machine reading comprehension dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2144–2148, Taipei, Taiwan. ACM.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, Baltimore, Maryland, USA. PMLR.

Chen Zhang, Jiuheng Lin, Xiao Liu, Yuxuan Lai, Yansong Feng, and Dongyan Zhao. 2023. How many answers should I give? an empirical study of multi-answer reading comprehension. In *Findings of the Association for Computational Linguistics: ACL*, pages 5811–5827, Toronto, Canada. Association for Computational Linguistics.

Xinyu Zhu, Cheng Yang, Bei Chen, Siheng Li, Jian-Guang Lou, and Yujiu Yang. 2023. Question answering as programming for solving time-sensitive questions. pages 12775–12790.

A Dataset Creation Details

We report more statistics of the created dataset, consisting of relation covered, question template and the extended dataset.

A.1 Relation Covered

There are a total of 46 relations in the MulTiple dataset. As shown in Figure 5, it illustrates relations frequency distribution in this dataset. "Member of sports team" and "position held" are the most frequent relations, accounting for 42% and 17%. "Employer" and "work location" are the next most common, both at 7%.

To ensure consistency with the real world, the selection of relations in the dataset leans towards using more common relations, such as "member of sports team" and "position held". To maintain diversity, the dataset includes relations from a wide range of fields, including sports, geography, and biography. This indicates that our dataset strikes a balance between consistency and diversity. In Table 5, we analyze the high-frequency relations and their descriptions within the dataset.

A.2 Question Template

To ensure the diversity of the dataset, we utilized a lot of templates to generate questions. In Table 6, we present the generation templates corresponding to high-frequency relations. In these templates, "[Subject]" denotes important entities in the question, and "[Timestamp]" denotes timestamps in the questions. For the Question-dependent Template, a special label "[Number]" is used to indicate the number of answers. In our question templates, each question is generated by 2-5 different templates. This template diversity enhances the variety of questions in the dataset.

A.3 Manual Revision

During quality control, it could be verified that all objects within the multi-object time-evolving event appear in the corresponding document before manual revision. Therefore, we first automatically retrieved and located the target objects in the document. Then, we extracted and located all timestamps within the document by using the time extraction tool, Time-Extractor. Next, workers revise each instance by reading and understanding the snippet annotated with objects or timestamps. Finally, they can revise the document or quadruple until the semantics of quadruple are fully contained in the corresponding document.

Specifically, there mainly exist the following three types of errors: 1) different boundary or expression: the object takes a different surface form in the quadruple and document (76.4%); 2) missing timestamps in documents: the object is mentioned in the text, but its corresponding timestamps are not mentioned and could not be inferred from available information (9.2%); 3) semantic mismatches: the object is mentioned in the text, but it does not express the same things as the quadruple (12.8%).

In addition, we make manual revisions for multi-object events, also considering the size of questions is large. For 37% of multi-object events (about 1480), 5 workers are selected to complete the annotation and check.

A.4 Extended Dataset

The extended dataset is an expansion of the basic dataset and includes single-answer and zero-answer questions. In the question-dependent extended dataset, there are 35,046 training examples, 6,930 validation examples, and 6,934 testing examples. In the document-dependent extended dataset, there are 35,164 training examples, 6,948 validation examples, and 6,934 testing examples. For detailed data analysis, please refer to Table 7.

B Baselines

In this section, we investigate how existing models could be adapted for multi-answer time-sensitive question answering.

B.1 Large Language Models

To further evaluate the MTQA task, we also conducted experiments using the existing latest large language models. Specifically,

- **Llama2** (Touvron et al., 2023) is a collection of opensourced LLMs trained on 2T tokens with efficient groupedquery attention, which outperform other models in most tasks, the dialog-fine-tuned Llama2-7b and Llama2-13b are used.

- **ChatGPT/GPT-4**(Ouyang et al., 2022; OpenAI, 2023) ChatGPT is a chat model aligned through SFT and RLHF based on GPT-3. GPT-4 is an upgraded version of ChatGPT with enhanced reasoning capabilities, making it the most powerful LLM. Unless otherwise

Relation	Name	Relation Description
P54	member of sports team	sports teams or clubs that the subject represents or represented
P39	position held	subject currently or formerly holds the object position or public office
P108	employer	person or organization for which the subject works or worked
P937	work location	location where persons or organizations were actively participating in employment, business or other work
P127	owned by	owner of the subject
P69	educated at	educational institution attended by subject
P97	noble title	titles held by the person

Table 5: Relations and their description in the MulTiple Dataset.

Relation	Question Template	
P54	Question-dependent	**[Subject]** played for which **[Number]** teams **[Timestamp]**? Which **[Number]** teams did the player **[Subject]** belong to **[Timestamp]**?
	Document-dependent	**[Subject]** played for which team **[Timestamp]**? Which team did the player **[Subject]** belong to **[Timestamp]**?
P39	Question-dependent	Which **[Number]** positions did **[Subject]** hold **[Timestamp]**? What were the **[Number]** positions of **[Subject]** **[Timestamp]**?
	Document-dependent	What position did **[Subject]** take **[Timestamp]**? Which position did **[Subject]** hold **[Timestamp]**?
P108	Question-dependent	What was the name of the **[Number]** employers **[Subject]** work for **[Timestamp]**? **[Timestamp]** was an employee for which **[Number]** employers **[Timestamp]**?
	Document-dependent	Who did **[Subject]** work for **[Timestamp]**? What was the name of the employer **[Subject]** work for **[Timestamp]**?
P937	Question-dependent	Which **[Number]** locations did **[Subject]** work **[Timestamp]**? What were the **[Number]** working locations for **[Subject]** **[Timestamp]**?
	Document-dependent	What was the working location for **[Subject]** **[Timestamp]**? **[Subject]** worked in which location **[Timestamp]**?
P127	Question-dependent	Who were the **[Number]** owners of **[Subject]** **[Timestamp]**? Which **[Number]** persons owned **[Subject]** **[Timestamp]**?
	Document-dependent	Who was the owner of **[Subject]** **[Timestamp]**? **[Subject]** was owned by whom **[Timestamp]**?
P69	Question-dependent	Where were **[Number]** places **[Subject]** educated **[Timestamp]**? **[Subject]** went to which **[Number]** schools **[Timestamp]**?
	Document-dependent	Where was **[Subject]** educated **[Timestamp]**? Which school did **[subject]** go to **[Timestamp]**?
P97	Question-dependent	What were the **[Number]** noble titles of **[Subject]** **[Timestamp]**? From **[Timestamp]**, what were the **[Number]** noble titles of **[Subject]**?
	Document-dependent	What was the noble title of **[Subject]** **[Timestamp]**? From **[Subject]**, what noble title did **[Timestamp]** hold?

Table 6: Templates of Question-dependent and Document-dependent question for frequent relations.

Split	Question Type	#Questions	#Documents	$\frac{\#\text{Questions}}{\#\text{Documents}}$	#Doc-Token	# Que-Token
Train	Question-dependent	35046	7394	4.74	1461.3	12.3
	Document-dependent	35164	7402	4.75	1461.9	11.2
Dev	Question-dependent	6930	1489	4.65	1488.8	12.8
	Document-dependent	6948	1491	4.66	1486.7	11.1
Test	Question-dependent	6934	1489	4.66	1467.2	12.2
	Document-dependent	6934	1489	4.66	1479.5	11.2
All	Question-dependent	48910	10372	4.72	1467.9	12.4
	Document-dependent	49046	10382	4.72	1467.9	11.2

Table 7: Statistics of MulTiple (extended). #Doc-Token and #Que-Token are the average number of tokens in long documents and questions, respectively.

stated, ChatGPT refers to gpt-3.5-turbo-0613 and GPT-4 refers to gpt-4-0613.

- **ReAct** (Yao et al., 2023) incorporates external knowledge through additional search and lookup actions.
- **QAaP** (Zhu et al., 2023) employs ChatGPT to extract structured facts and convert TSQA into program execution.

For all experiments, we employ GPT-3.5-Turbo as the model unless otherwise specified. In addition, for cost considerations, we randomly select 300 questions to evaluate for both question-dependent and document-dependent questions, respectively. The prompts are listed in Table 8.

B.2 Pre-trained Language Models

We select 7 models for targeted adjustments, including multi-span question answering and popular generative models:

- **Li et.** (Li et al., 2022) propose a multi-span QA model to capture global information by combining a sequence tagger with a span number predictor. Considering the answer as a span, we employ the multi-task learning framework for predicting the answers and their counts.
- **MTMSN** (Hu et al., 2019) is presented to predict various types of answers and dynamically extract one or multiple spans based on the produced number of answers. We treat predicting the number of answers as an auxiliary task and extract non-overlapped answers with a specific amount.
- **TASE** (Segal et al., 2020) is proposed for multi-span question answering by casting it as a sequence tagging task, predicting whether each token is part of the answer. We consider multi-answer questions as multi-span questions and train our model initialized with BERT$_{LARGE}$.
- **ITERATIVE** (Zhang et al., 2023) is designed to extract multiple answers iteratively. During each iteration, it appends the previously extracted answers to the question with the word 'except' in the middle and then feeds the updated question into the single-answer TimeQA model. The iterative process terminates when the model predicts no more answers.
- **T5** (Raffel et al., 2020) aims to perform supervised fine-tuning of traditional T5 models on each setting of MulTiple. We consider the concatenated multiple answers separated by semicolons as a single ground truth during this process.
- **FiD** (Izacard and Grave, 2021) suggests splitting the long document into multiple short paragraphs and generates the answer token by token in an autoregressive fashion. Similar to **T5** (Raffel et al., 2020), it treats the concatenated multiple answers as the ground truth during training.
- **REMEMO** (Yang et al., 2023) devises a graph view to explicitly connect all temporally-scoped facts by modeling the time relations between any two sentences. Note that REMEMO would not be fair to compare on the datasets because it is trained on selected samples and truncated context.

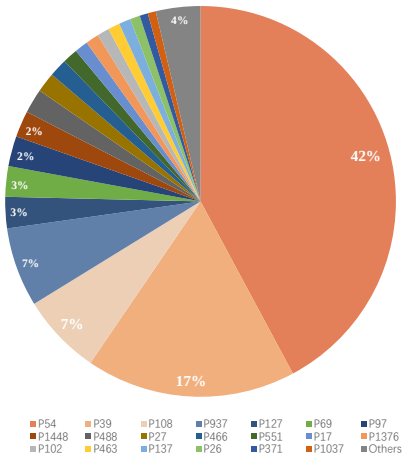


Figure 5: The proportion of relations in MulTiple.

C Human Evaluation

We also manually answer the selected 300 questions above for both question-dependent and document-dependent questions, respectively. Specifically, we hire three proficient English-speaking annotators, ensuring they can comprehend the questions and documents. Each annotator is required to independently complete the following tasks: 1) reading the document and comprehending its content; 2) extracting answers from the document for each question. As with the answers predicted by models, the results of three annotators

are further evaluated and averaged to calculate the final SEM, EM, and PM scores.

D Evaluation Metrics

We evaluate multi-answer time-sensitive question answering in three levels: token, individual answer and whole answer. They are computed by three metrics, including Exact Match, Strict Exact Match and Partial Match, followed by (Li et al., 2022) and (Li et al., 2023).

Exact Match (EM). An exact match occurs when the predicted answers exactly match one of the ground-truth answers. It is computed by treating the predicted and ground-truth answers as a set of answers based on the standard formulation of Precision (Pre), Recall (Rec) and F1 as follows:

$$Pre = \frac{TP}{TP + FP}; \quad Rec = \frac{TP}{TP + FN}; \quad (2)$$

$$F_1 = \frac{2Pre \cdot Rec}{Pre + Rec}. \quad (3)$$

where TP (True Positive) is the number of answers correctly predicted by the model, FP (False Positive) is the number of answers incorrectly predicted by the model, and FN (False Negative) is the number of answers predicted by the model but not exist in the ground-truth answers. **Strict Exact Match (SEM)** is counted as correct if and only if all ground-truth answers are matched exactly, that is, both FP and FN are 0.

Partial Match (PM). The partial match aims to measure the overlap between the predictions and ground truth answers. We compute it by treating the precision, recall, and F1 as a string. Specifically, for each pair of prediction p_i and ground truth answer a_j, Precision and Recall are defined as follows:

$$\begin{aligned} Pre &= \frac{\sum_{i=1}^{n} \max_{j \in [1,m]} s_{ij}^{ret}}{n} \\ Rec &= \frac{\sum_{j=1}^{m} \max_{i \in [1,n]} s_{ij}^{rel}}{n} \end{aligned} \quad (4)$$

where n and m are the number of generated and ground-truth answers, respectively. s_{ij}^{ret} and s_{ij}^{rel} are the partial retrieved score and partial relevant score, calculated as:

$$s_{ij}^{ret} = \frac{len(LCS(p_i, a_j))}{len(p_i)} \quad (5)$$

$$s_{ij}^{rel} = \frac{len(LCS(p_i, a_j))}{len(a_j)} \quad (6)$$

$LCS(p_i, a_j)$ denotes the length of the longest common substring between the prediction p_i and the ground truth answer a_j. $len(\cdot)$ represents the length of the string.

Modes	Prompt Template
Document-dependent	**Task description** Give a time-sensitive question and the corresponding long document, please read the long document and then answer the question. Note that: 1) each question has more than one answer; 2) please generate all answers and separate them with a semicolon. **Question**: What were the positions of Ashley Fox between 2011 and 2012? **Document**: Ashley Fox (born 15 November 1969) is a British Conservative Party politician. He was a Member of the European Parliament (MEP) for South West England and Gibraltar. He was leader of the Conservatives in the European Parliament from 2014 to 2019. He chairs the Independent... Fox was first elected to the member of the European Parliament in 2009 and was re-elected in 2014 before losing his seat in 2019. Fox served as Chief Whip of the European Conservative & Reformists Group (ECR) 2010-2014. In his first mandate... In 2011-12, Fox was rapporteur on Corporate Governance in Financial Institutions. In 2016,... **The answers are**: member of the European Parliament; Chief Whip of the European Conservative & Reformists Group (ECR); rapporteur on Corporate Governance Following the above example, please generate all answers to the following multi-answer time-sensitive questions and separate them with a semicolon. **Question**: [Question] **Document**: [Document] **The answers are**:
Question-dependent	**Task description** Give a time-sensitive question and the corresponding long document, please read the long document and then answer the question. Note that: 1) each question has more than one answer; 2) **the number of answers is mentioned in the question**; 3) please generate all answers and separate them with a semicolon. **Question**: What were the three positions of Ashley Fox between 2011 and 2012? **Document**: Ashley Fox (born 15 November 1969) is a British Conservative Party politician. He was a Member of the European Parliament (MEP) for South West England and Gibraltar. He was leader of the Conservatives in the European Parliament from 2014 to 2019. He chairs the Independent... Fox was first elected to the member of the European Parliament in 2009 and was re-elected in 2014 before losing his seat in 2019. Fox served as Chief Whip of the European Conservative & Reformists Group (ECR) 2010-2014. In his first mandate... In 2011-12, Fox was rapporteur on Corporate Governance in Financial Institutions. In 2016,... **The answers are**: member of the European Parliament; Chief Whip of the European Conservative & Reformists Group (ECR); rapporteur on Corporate Governance Following the above example, please generate **the specified number of answers** to the following multi-answer time-sensitive questions and separate them with a semicolon. **Question**: [Question] **Document**: [Document] **The answers are**:

Table 8: An illustration of instance formatting and two different methods for constructing the instruction-formatted instances.