

1 Introduction for Supplementary Materials

Here are some supplementary materials for the submitted paper *TIME: A Multi-level Benchmark for Temporal Reasoning of LLMs in Real-World Scenarios*.

Note that these supplementary materials is not another version of appendix for this paper. For the convenience of reviewing, we provide reviewers with 3 parts of supplementary materials: (1) the details of temporal knowledge graph construction for TIME-WIKI. (2) the prompts for QA synthesis. (3) More details for dataset statistics. (4) QA examples. (5) some supplementary performance for some LLMs (and human annotators) on the TIME-LITE.

For code, please refer to <https://github.com/sylvain-wei/TIME>. For dataset, please refer to <https://huggingface.co/datasets/SylvainWei/TIME>. And TIME’s GitHub Webpage URL is <https://omni-time.github.io>.

2 Supp 1: Benchmark Construction

2.1 Temporal Knowledge Graph Construction

2.1.1 Collect Context for Evaluation for TIME-WIKI

To provide comprehensive world knowledge contexts for evaluation, we synthesize evaluation contexts for LLMs by leveraging the previously constructed Temporal Knowledge Graph. Specifically, we first build timelines for link entities from the temporal knowledge graph as structured raw data. For each link entity, we chronologically concatenate temporal facts and paraphrase them into coherent stories. Finally, we prompt LLMs to generate contexts centered around three interconnected link entities. Below we present our few-shot prompts for paraphrasing facts into stories and concatenating stories for TIME-WIKI, followed by an example of the final concatenated context.

Few-shot Prompt for Paraphrasing Facts into Stories for TIME-WIKI

[Instruction:] You are a very skilled storyteller. Below are the facts I've extracted from WikiData. Please craft a complete story centered around {link_entity}.

[Requirements:]

1. Generate content that accurately conveys all the facts I have provided without omitting any of them.
2. Particular attention must be paid to the accuracy of time expressions, and the temporal relationships between sentences must be coherent and natural.
3. The story should be coherent, natural, and smooth, consistently revolving around the core facts, with concise and precise language, avoiding any unnecessary digressions.

[Example:] [Core entity:] Mauro Morelli

[Facts:] Mauro Morelli holds the position of diocesan bishop from May 25, 1981 to March 30, 2005.. Mauro Morelli holds the position of titular bishop in December 12, 1974.. Mauro Morelli holds the position of auxiliary bishop in December 12, 1974.. Mauro Morelli works as a Catholic priest in April 28, 1965.. Mauro Morelli works as a Catholic bishop in January 25, 1975.

[Generated Story:] Mauro Morelli's journey in the Catholic Church is a testament to his unwavering dedication and service. Born into a world of faith, his path was set early on, leading him to become a significant figure in the Church's hierarchy.

On April 28, 1965, Mauro Morelli took his first major step in his religious career by being ordained as a Catholic priest. This ordination marked the beginning of his lifelong commitment to serving the Church and its followers. For nearly a decade, he devoted himself to his priestly duties, gaining the respect and admiration of his parishioners.

[Example:] (Here is another example with the same format of the example above.)

[TODO:] Below are the facts and the link entity. Please generate a story in accordance with the aforementioned requirements and with reference to the examples provided.

[Link entity:]{link_entity}

[Facts:]{raw_text}

[Generated story:]

Prompt for Concatenating Stories for TIME-WIKI

[Instruction:] You are adept at weaving multiple complete stories together to form a coherent and comprehensive narrative. I will provide you with several stories and the connections between them. I hope you can retain all the facts from the stories without any omissions or errors in time or events.

[Requirements:]

1. The content you generate must not omit any time-related information from the texts being concatenated.
2. Particular attention must be paid to the accuracy of time expressions, and the temporal relationships between sentences must be coherent and natural.
3. When concatenating, pay attention to the inherent logical connections between the given multiple stories, and refer to the provided linking facts to generate transitional paragraphs that seamlessly connect the different stories.

[Link facts:] {linked_facts}

[Stories:] {stories}

[Generated Story:]

23

One Example for the Concatenated Story for TIME-WIKI

Mauro Morelli's life was a testament to unwavering dedication and service to the Catholic Church. His journey began with a profound commitment to his faith, which shaped his path toward becoming a respected leader within the Church's hierarchy. On April 28, 1965, Mauro Morelli was ordained as a Catholic priest, marking the beginning of his spiritual and pastoral vocation. For nearly a decade, he devoted himself to his priestly duties, guiding and nurturing the communities he served. His compassion, wisdom, and deep connection to his faith earned him the trust of his parishioners and the attention of his superiors.

Recognizing his leadership qualities, the Church appointed Mauro Morelli as both a titular bishop and an auxiliary bishop on December 12, 1974. These dual roles signified a new chapter in his journey, as he took on greater responsibilities in assisting the diocesan bishop and overseeing pastoral and administrative tasks. His dedication to these roles showcased his ability to balance spiritual guidance with effective governance. Just over a month later, on January 25, 1975, Mauro Morelli was consecrated as a Catholic bishop. This formal consecration solidified his position as a key figure within the Church, empowering him to lead with authority and grace. His work as a bishop further deepened his impact on the communities he served, as he continued to champion the values of the Church.

Skiping 3 paragraphs here...

As we reflect on Mauro Morelli's life and the broader history of auxiliary and titular bishops, we are reminded of the importance of service, humility, and faith. The roles of these bishops, though sometimes overlooked, are a testament to the Church's enduring commitment to its mission of love, guidance, and spiritual care. Their contributions, woven together, form a testament to the enduring legacy of leadership within the Catholic tradition.

24

25 2.2 QA Synthesis

26 2.2.1 Level 1: Basic Temporal Understanding and Retrieval

27 At this level, we primarily evaluate models' capabilities in temporal information retrieval and compre-
28 hension. We design tasks focusing on temporal information extraction, including directly extracting
29 time expressions from context (Extract) and determining event occurrence times given specific
30 events (Localization). Additionally, we assess fundamental temporal understanding through tasks
31 such as duration calculation (Computation), duration comparison (Duration Compare), and event
32 ordering (Order Compare).

33 **Extract** For the Extract task, we first provide a set of authentic time points extracted from the
 34 source data. We then instruct the LLM to generate five novel time expressions that are distinct from
 35 all existing ones. Subsequently, we randomly select 0-4 authentic time points to form multiple-
 36 choice options, ensuring their randomness. This approach combines rule-based selection with
 37 LLM-generated distractors, thereby maintaining high QA quality. The task is formulated using a
 38 fixed question template: *Which of the following are time expressions mentioned in the context? (Note:*
 39 *There may be one or more correct options. If you think NONE of the time expressions in options*
 40 *A/B/C/D are mentioned, then you can choose E. Do not choose E together with other options.)*

Prompt for generating fake time expressions (Extract) (TIME-WIKI)

[Rules:] Given a list of time expressions, please generate FIVE new time expressions randomly. You should follow the instructions below:

1. The 5 new time expressions should totally different from all of the the given time expressions.
2. Each new time expression should be in the format of <Month> <Day>, <Year>. For example: "May 4, 1998", "April 1992", "1934".
3. The 5 new time expressions should closely resemble the given time expression in format and structure, creating a high level of confusion, yet they must represent entirely different times. For instance, you could alter the year while keeping the month and day unchanged, such as changing May 2, 1922 to May 2, 1923; or you could modify the month and day while retaining the same year, for example, transforming May 2, 1922 into July 2, 1922. Additionally, you might consider changing the day while keeping the month and year consistent, like adjusting May 2, 1922 to May 3, 1922. Another approach could involve altering the month and year but keeping the day the same, such as changing May 2, 1922 to May 2, 1921. Lastly, you could shift the entire date by a consistent interval, for example, moving May 2, 1922 to June 3, 1923, ensuring that each change introduces a subtle yet distinct variation from the original time expression.

[Example:] Here are some examples showing the writing style. NOTE that the content of the examples are irrelevant to the question you will generate.

[Given time expressions:] ['1656', '1763', '1782', '1784', '1787', '1815', 'September 20, 1817', 'January 20, 1848', ..., 'March 30, 2005', 'June 27, 1965', '1973', '2009', 'May 25, 1963', '1980', '2008', 'June 11, 1949', '1983', '1984', '2007', 'August 14, 2004', 'October 7, 2014', 'November 11, 2021', 'October 14, 2024', 'May 31, 1971', 'October 28, 1975']

[New 5 time expressions:] ['1682', 'June 12, 1974', 'May 8, 1907', 'May 11, 1949', 'May 25, 2011']

Skipping 2 examples here.....

[Output:] Now please write 5 new time expressions following the instructions and examples above. You should output the 5 new time expressions along with its answer, in the format of ["YY", "MM, YY", "MM DD, YY"]. NOTE that the time expressions should be in chronological order. Now the given time expressions are:

[Given time expressions:] {Given_time_expressions}

[New 5 time expressions:]

41

42 **Localization** We select three facts from the temporal knowledge graph and employ LLMs to
 43 generate corresponding QA pairs.

Prompt for generating QA. (Localization) (TIME-WIKI)

[Rules:] Given 3 facts, please generate one question along with its gold answer for each given fact. You should follow the instructions below:

1. The answer **MUST** be short and concise, avoiding using redundant words or repeating the information in the question.
2. You should output the question and its answer without any other explanation, such as "Question: xxx? Answer: xxx."
3. I will give you 3 facts in each line, and then you should output the question and its answer each line in the same sequence. So your output should be 3 lines of "Question: xxx? Answer: xxx.".
4. The question can be phrased in different ways, such as 'When is...?', 'What is the time for...?', 'What time did...?', and so on.
5. The answer should be "From xxx to xxx." or directly "xxx."

[Example:] Here are some examples showing the writing style. NOTE that the content of the examples are irrelevant to the question you will generate.

* **[Given facts:]**

Debra Hamel worked at Wesleyan University from 1998 to 2001.

Oliver Marcy attended Wesleyan University in 1846.

Debra Hamel completed her studies at Yale University in 1993.

[Generated QA:]

Question: When did Debra Hamel work at Wesleyan University? Answer: From 1998 to 2001.

Question: What time did Oliver Marcy attend Wesleyan University? Answer: 1846.

Question: What is the time for Debra Hamel to complete her studies at Yale University? Answer: 1993.

Skipping 1 example here.....

[Output:] Now please write a question following the instructions and examples above.

You should output the question along with its answer, in the format of "Question: xxx? Answer: xxx.". NOTE that the answer should be "From xxx to xxx." or directly "xxx."

[Given facts:]

{given_facts

[Generated QA:]

44

45 **Computation** We provide multiple pairs of temporal facts and utilize LLMs to generate corre-
46 sponding questions. The temporal computations are systematically derived through script-based rules
47 to ensure accuracy and consistency.

Prompt for generating QA. (Computation) (TIME-WIKI)

[Rules:] Each time, I will provide you with several pairs of text snippets, with each pair occupying one line. For each line containing a pair of text snippets, you need to generate a question. You should follow the instructions below:

1. The question should be based on the snippet pair.
2. Each text snippet pair includes two snippets. Each snippet is composed of a fact and a 'Happen/Begin/End' time. You should translate the snippets into a format like:
3. Please refer to the following examples and learn the patterns well.

[Example:] Here are one example showing the writing style. NOTE that the content of the examples are irrelevant to the question you will generate.

[Snippet pairs:]

José Manuel Pasquel served as an auxiliary bishop. Happen time: January 20, 1848. José Manuel Pasquel became a priest. Happen time: September 20, 1817.

Antonius Grech Delicata Testaferrata was ordained. Happen time: October 19, 1845. Diego Fabbrini played for Watford F.C.. Begin time: 2013.

Máximo Alcócer played for the Bolivia men's national football team. Begin time: 1957. Diego Fabbrini played for Udinese Calcio. 1944.

[Questions:]

What was the duration from the time José Manuel Pasquel became a priest until he served as an auxiliary bishop?

How long was it between Antonius Grech Delicata Testaferrata was ordained and Diego Fabbrini began to play for Watford F.C.?

How much time passed from Máximo Alcócer began to play for the Bolivia men's national football team to Diego Fabbrini end playing for Udinese Calcio?

[Instruction:]

Now please write a question following the instructions and examples above. You should output the question only for each line. NOTE that there is NO any prefix like "Question:", just output the question string.

[Snippet pairs:]

{snippet_pairs}

[Questions:]

48

49 **Duration Compare** The Duration Compare task is designed to evaluate models' capability in
50 comparing the lengths of two time intervals. Specifically, we represent each time interval by the
51 span between two distinct event timestamps. To construct questions for this task, we extract pairs of
52 non-overlapping events from the timeline to form two comparable time intervals. A sample question
53 template is: "Which of the following two durations is longer? *Duration 1:* Between {fact1_1}
54 and {fact1_2} *Duration 2:* Between {fact2_1} and {fact2_2}" For the TIME-WIKI benchmark,
55 we have developed alternative question templates based on temporal facts, including: "Which fact
56 lasted longer, Fact 1: {fact1} or Fact 2: {fact2}?", "Which of the two events, Fact 1: {fact1} or Fact
57 2: {fact2}, had a longer duration?", "Compare the duration of Fact 1: {fact1} and Fact 2: {fact2}.
58 Which one was longer?"

59 **Order Compare** The Order Compare task evaluates models' capability to comprehend temporal
60 ordering between two time points. To avoid direct comparison of timestamps, we utilize event
61 occurrence times as the comparison points. By leveraging a collection of temporal facts, we can
62 directly select two facts and embed them into predefined question templates, as illustrated below.

Question Templates. (Order Compare)

"For Fact1: {fact1} and Fact2: {fact2}, which one happened earlier?"

"Which started earlier, Fact1: {fact1} or Fact2: {fact2}?"

"Which ended earlier, Fact1: {fact1} or Fact2: {fact2}?"

"Did Fact1: {fact1} start before Fact2: {fact2} ended?"

"Did Fact1: {fact1} end before Fact2: {fact2} started?"

"Did Fact1: {fact1} start before Fact2: {fact2} happened?"

63

64 2.2.2 Level 2: Temporal Expression Reasoning

65 This level comprises three distinct subtasks that collectively assess the model’s capability in perform-
 66 ing multi-hop reasoning over temporal expressions. Each question can only be correctly answered if
 67 the LLM successfully conducts accurate multi-hop reasoning on the temporal expressions themselves.

68 **Explicit Reasoning** In this task, we employ temporal expressions composed of explicit time points
 69 in the questions. Notably, these explicit time points do not exist in the original context. We first
 70 randomly select temporal facts as ground truth answers, then transform their time points to generate
 71 non-existent temporal references. Finally, we utilize these modified time points to formulate questions.
 72 Below demonstrates the prompt template for generating Explicit Reasoning task questions in the
 73 TIME-WIKI dataset.

Generate Questions by Modifying Time Expressions (Explicit Reasoning) (TIME-WIKI)

[Rules:] Given the original questions and their corresponding updated time expressions, generate new questions by replacing the time expressions in the original questions. Follow these guidelines:

1. Modify only the time expressions; leave all other parts of the questions unchanged.
2. Output only the questions; do not include any answers.
3. Present each question on a separate line.
4. If the temporal expression in the given original question is a time point, such as "on September 9, 2002," then each time I will provide you with a time period expressed with "from ... to ...", for example, "from April 6, 1999 to June 2003." What you need to pay attention to is that your revised question should carry a tone of uncertainty. For instance, you should change "Which team did Ted play for on September 9, 2002?" to "Which team might Ted have played for from April 6, 1999 to June 2003?"

[Example:] Here are some examples showing the writing style. NOTE that the content of the examples are irrelevant to the question you will generate.

[Original Question:]

What position did Alexandre da Sagrada Família hold on August 8, 1782?

[New Time Expression:]

from April 1772, to June 1784

[New Question:]

What position have Alexandre da Sagrada Família held from April 1772, to June 1784?

Skipping some examples...

[Output:] Now please write a question following the instructions and examples above.

[Original Question:]

{original_question}

[New Time Expression:]

{new_time_expression}

[New Question:]

74

75 **Order Reasoning** This task evaluates the model’s comprehension of ordinal temporal expressions,
 76 such as "the second time serving as a professional basketball player" or "the last time attending a
 77 ballet performance." Correctly interpreting these expressions requires the LLM to fully understand
 78 the timeline to accurately identify the specific time point referenced. For task construction, we first
 79 establish a timeline for the same entity. Then, we identify sub-timelines sharing the same factual
 80 relationship and select the k-th fact in chronological order to generate questions. Below are two
 81 prompt templates for question generation in TIME-WIKI, given specific facts and their corresponding
 82 timeline orders.

Generate Question based on Subject-oriented Fact Timeline (Order Reasoning) (TIME-WIKI)

[Rules:] Given a sentence describing a simple fact and an order number, please generate one question along with its answer. You should follow these instructions:

1. The question **MUST** target the subject in the factual statement. For example, given "Bruno Aguiar plays for Portugal national under-21 football team from 2001 to 2004." with order number "3", generate "Who is the third person affiliated with Portugal national under-21 football team?" with answer "Bruno Aguiar".
2. Formulate questions exclusively based on the provided factual content and numerical order.
3. Exclude temporal expressions from generated questions while maintaining factual integrity.
4. Craft unambiguous questions using diverse interrogative structures (e.g., "Which individual...", "What entity...") that require contextual analysis rather than lexical matching.
5. Ensure answers contain only the factual subject without explanatory content or repetition.
6. Present results strictly as: "Question: xxx? Answer: xxx." with continuous formatting(e.g. with no line breaks).

[Example:] Here are some examples showing the writing style. NOTE that the content of the examples are irrelevant to the question you will generate.

[Given fact:]

Julius Babatunde Adedokun holds the position of diocesan bishop from April 13, 1973 to November 4, 2009.

[Order number:]

2

[Generated QA:]

Question: Who is the second person to hold the position of diocesan bishop? Answer: Julius Babatunde Adedokun.

Skipping some examples here...

[Output:] Now please write a question and its answer following the instructions and examples above. You should output the question along with its answer, in the format of "Question: xxx? Answer: xxx.". NOTE that the answer should be as short as possible.

[Given fact:]

{ given_fact }

[Order number:]

{ order_number }

[Generated QA:]

Generate Question based on Object-oriented Fact Timeline (Order Reasoning) (TIME-WIKI)

[Rules:] Given a sentence describing a simple fact and an order number, please generate one question along with its answer. You should follow the instructions below:

1. This question **MUST** be directed at the object in the fact. For example, given a fact "Bruno Aguiar plays for Portugal national under-21 football team from 2001 to 2004.", and an order number "3", your question should be "What was the third team Bruno Aguiar was affiliated with during her professional career?" and the answer should be "Portugal national under-21 football team".
2. The question should be derived directly from the factual content.
3. The question must exclude time expressions present in the original fact.
4. Phrase questions unambiguously using varied interrogative patterns (e.g., "Which team...", "What position...", "What organization...") while avoiding simple string matching.
5. The answer **MUST** contain only the factual object without explanations or repetitions.
6. Output strictly in the format: "Question: xxx? Answer: xxx." with no line breaks.

[Example:] Here are some examples showing the writing style. NOTE that the content of the examples are irrelevant to the question you will generate.

[Given fact:]

Julius Babatunde Adelakun holds the position of diocesan bishop from April 13, 1973 to November 4, 2009.

[Order number:]

2 **[Generated QA:]**

Question: What was the second position Julius Babatunde Adelakun held in his role?

Answer: diocesan bishop.

[Output:] Now please write a question and its answer following the instructions and examples above. You should output the question along with its answer, in the format of "Question: xxx? Answer: xxx.". NOTE that the answer should be as short as possible.

[Given fact:]

{given_fact}

[Order number:]

{order_number}

[Generated QA:]

84

85 **Relative Reasoning** This task evaluates models' reasoning capabilities with relative temporal
86 expressions. For instance, "within the last two weeks before Trump's second official election as
87 US President" and "within 7 months and 2 days after Xiao Ming's official graduation" demonstrate
88 temporal reasoning based on specific event anchors. Such expressions pose significant challenges
89 to LLMs' temporal reasoning abilities. To construct QA pairs for this task, we first provide ground
90 truth temporal fact statements along with their corresponding relative temporal expressions. Notably,
91 these relative temporal expressions are pre-extracted from all temporal facts, each consisting of a
92 factual statement and a relative temporal expression. The QA generation methodology is detailed in
93 the following prompt examples.

Generate Question based on Subject-oriented Fact Timeline (Relative Reasoning) (TIME-WIKI)

[Rules:] Given a sentence describing a simple fact and a time expression, please generate one question. You should follow the instructions below:

1. The question **MUST** be directed at the subject in the fact. For example, given a fact "Bruno Aguiar plays for Portugal national under-21 football team.", and the time expression "Assuming today is January, 2002 | before today", your question should be "Assuming today is January, 2002, who was the most recently player that played for Portugal national under-21 football team before today?".

2. The question should come from the facts.

3. The question should be unambiguous and challenging, avoiding simple string matching. NO sub-questions allowed.

4. You should output the question without any other explanation. You should output the question directly.

[Example:] Here are some examples showing the writing style. NOTE that the content of the examples are irrelevant to the question you will generate.

Example 1:

[Given fact:]

Julius Babatunde Adhlakun holds the position of diocesan bishop.

[Time Expression:]

Assuming today is January, 2002 | before today

[Generated Question:]

Assuming today is January 2002, who was the most recent person that held the position of diocesan bishop before today?

Example 2:

[Given fact:]

Gerolamo Castaldi holds the position of diocesan bishop.

[Time Expression:]

before Francesco Marmaggi works as a Catholic priest

[Generated Question:]

Who was the most recent person to hold the position of diocesan bishop before Francesco Marmaggi works as a Catholic priest?

Example 3:

[Given fact:]

Włodzimierz Roman Juszczyk holds the position of diocesan bishop.

[Time Expression:]

before September 9, 2009 | within the span of 3 years 2 months 28 days

[Generated Question:]

Who was the most recent person to hold the position of diocesan bishop before September 9, 2009, within the span of 3 years 2 months 28 days?

[Output:] Now please write a question following the instructions and examples above. You should directly output the question.

[Given fact:]

{given_fact}

[Time Expression:]

{time_expression}

[Generated Question:]

Generate Question based on Object-oriented Fact Timeline (Relative Reasoning) (TIME-WIKI)

[Rules:] Given a sentence describing a simple fact and a time expression, please generate one question. You should follow the instructions below:

1. The question **MUST** be directed at the object in the fact. For example, given a fact "Bruno Aguiar plays for Portugal national under-21 football team.", and the time expression "Assuming today is January, 2002 | before today", your question should be "Assuming today is January, 2002, which team did Bruno Aguiar play for?".
2. The question should come from the facts.
3. The question should be unambiguous and challenging, avoiding simple string matching. NO sub-questions allowed.
4. You should output the question without any other explanation. You should output the question directly.

[Example:] Here are some examples showing the writing style. NOTE that the content of the examples are irrelevant to the question you will generate.

Example 1:

[Given fact:]

Julius Babatunde Adelakun holds the position of diocesan bishop.

[Time Expression:]

Assuming today is January, 2002 | before today

[Generated Question:]

Assuming today is January, 2002, what position did Julius Babatunde Adelakun most recently hold before today?

Example 2:

[Given fact:]

Gerolamo Castaldi holds the position of diocesan bishop.

[Time Expression:]

before Francesco Marmaggi works as a Catholic priest

[Generated Question:]

What position did Gerolamo Castaldi most recently hold before Francesco Marmaggi worked as a Catholic priest?

Example 3:

[Given fact:]

Wlodzimierz Roman Juszcak holds the position of diocesan bishop.

[Time Expression:]

before September 9, 2009 | within the span of 3 years 2 months 28 days

[Generated Question:]

What position did Wlodzimierz Roman Juszcak most recently hold within the 3 years, 2 months, and 28 days prior to September 9, 2009?

[Output:] Now please write a question following the instructions and examples above. You should directly output the question.

[Given fact:]

{given_fact}

[Time Expression:]

{time_expression}

[Generated Question:]

95

96 **2.2.3 Level 3: Complex Temporal Relationship Reasoning**

97 At this level, we focus on evaluating the model's capability to comprehend both implicit and explicit
98 temporal relationships between events. Specifically, we assess the model from three perspectives:
99 (1) temporal co-occurrence between events (Co-temporality), (2) complete reordering of multiple
100 distinct events (Timeline), and (3) the model's simultaneous understanding of both the original

101 context and the question when temporal expressions in the question contradict the source text
102 (Counterfactual).

103 **Co-temporality** This task evaluates the model’s ability to comprehend temporal co-occurrence
104 between two events. For instance, the question "When Sam Altman co-founded OpenAI, what
105 positions did Elon Musk hold?" implicitly assumes the temporal overlap between "Sam Altman
106 founding OpenAI" and "Elon Musk serving as co-founder of OpenAI and CEO of Tesla and SpaceX".
107 To construct questions for this task, we provide the LLM with two key elements: a condition fact that
108 serves as the temporal reference, and a query fact that forms the basis for question generation. The
109 following demonstrates the prompt template.

Generate Question based on Subject-oriented Fact Timeline (Co-temporality) (TIME-WIKI)

[Rules:] Given a condition fact and a query fact, please generate one question. You should follow these instructions:

1. The question **MUST** target the subject in the factual statement. For example, given the condition fact "Mauro Morelli holds the position of diocesan bishop." with query fact "William Weigand holds the position of diocesan bishop.", generate "When Mauro Morelli holds the position of diocesan bishop, who held the position of diocesan bishop?"
2. The question should come from the facts.
3. The question should be unambiguous and challenging, avoiding simple string matching. NO sub-questions allowed.
4. You should output the question without any other explanation. You should output the question directly.

[Example:] Here are some examples showing the writing style. NOTE that the content of the examples are irrelevant to the question you will generate.

Example 1:

[Given fact:]

Julius Babatunde Adedokun holds the position of diocesan bishop from April 13, 1973 to November 4, 2009.

[Generated QA:]

Question: Who served as diocesan bishop from April 13, 1973 to November 4, 2009?

Answer: Julius Babatunde Adedokun.

Skipping the rest of the examples.

[Output:] Now please write a question following the instructions and examples above. You should directly output the question.

[Condition fact:]

{condition_fact}

[Query fact:]

{query_fact}

[Generated Question:]

110

Generate Question based on Object-oriented Fact Timeline (Co-temporality) (TIME-WIKI)

[Rules:] Given a condition fact and a query fact, please generate one question. You should follow these instructions:

1. The question **MUST** target the object in the factual statement. For example, given the condition fact "Mauro Morelli holds the position of diocesan bishop." with query fact "William Weigand holds the position of diocesan bishop.", generate "When Mauro Morelli held the position of diocesan bishop, what position did William Weigand hold?"
2. The question should come from the facts.
3. The question should be unambiguous and challenging, avoiding simple string matching. NO sub-questions allowed.
4. You should output the question without any other explanation. You should output the question directly.

[Example:] Here are some examples showing the writing style. NOTE that the content of the examples are irrelevant to the question you will generate.

Example 1:

[Condition fact:]

Mauro Morelli holds the position of diocesan bishop.

[Query fact:]

William Weigand holds the position of diocesan bishop.

[Generated Question:]

When Mauro Morelli held the position of diocesan bishop, what position did William Weigand hold?

Skipping the rest of the examples.

[Output:] Now please write a question following the instructions and examples above. You should directly output the question.

[Condition fact:]

{condition_fact}

[Query fact:]

{query_fact}

[Generated Question:]

111

112 **Timeline** The Timeline task is designed to evaluate a model’s ability to chronologically reorder
113 multiple facts within a given context. In TIME-WIKI and TIME-DIAL, we assess the model’s
114 capability to determine temporal relationships among eight distinct facts, generating only one
115 Timeline task question per timeline and context. In contrast, for TIME, we directly utilize multiple
116 reordering questions from TCELongBench[1].

117 For constructing Timeline task questions, we employ existing timelines containing eight temporal
118 facts (without explicit temporal expressions or timestamps). Using a Python program, we first
119 compute the chronological order of events, then randomly shuffle the order of the eight facts, with
120 the correct temporal sequence serving as the ground truth. In this question construction process, we
121 solely rely on question templates to generate the final output.

122 The question template is as follows:

Question Template. (Timeline)

"Below are 8 facts. You need to sort these facts in chronological order. Requirements:
Your output format must be numbers enclosed in parentheses without any other symbols or
whitespace. For example: (1)(5)(2)(7)(3)(8)(6)(4){eight facts list}"

123

124 **Counterfactual** To thoroughly evaluate models’ understanding of temporal relationships in context,
125 we modify the Explicit Reasoning task by counterfactually altering temporal expressions in
126 questions, making them contradict the temporal information in the provided real context. During

127 evaluation, we instruct models to strictly adhere to the new temporal expressions in their responses.
 128 This approach eliminates direct reliance on contextual information (i.e., surface-level event-event
 129 correlations), enabling a fair assessment of models' genuine comprehension of temporal sequences.

130 Specifically, we construct task questions by directly modifying temporal expressions in Explicit
 131 Reasoning questions while preserving all other event details. After altering the temporal conditions,
 132 we employ a Python program with rule-based matching to determine whether the original answer
 133 remains valid under the new conditions. If the original answer no longer satisfies the new conditions,
 134 the correct answer becomes "There is no answer"; otherwise, the original gold answer remains
 135 unchanged.

136 We employ two distinct prompt strategies for question construction. The first prompt requires the
 137 model to generate a new temporal expression that replaces the original one while maintaining the
 138 same answer as before. The second prompt instructs the model to generate a new temporal expression
 139 that results in a different answer from the original.

Generate Question based on False Fact Premise and the new answer is the same as the original answer (Counterfactual Question) (TIME-WIKI)

[Scenario:] You are an annotator who is exceptionally skilled at generating false temporal facts and premises. First, you carefully comprehend the question I provided you and its corresponding correct answer. Based on the question and answer, you cleverly imagine an "if" clause that represents a hypothesis. This hypothesis must contradict the facts in the given context but must simultaneously ensure that the provided answer remains correct. Your hypothesis only needs to modify the temporal elements to differ from the original, such as altering the year, month, or day. You need to add the imagined "if" clause to the original question and only output the question itself with the added "if" clause.

[Example:] Here are some examples showing the writing style. NOTE that the content of the examples are irrelevant to the question you will generate.

[Given info:]

Question: Who might have worked as a Catholic priest before April 25, 1830?

Answer: Alexandre da Sagrada Família

[Generated new question:]

Who might have worked as a Catholic priest before April 25, 1830, if Alexandre da Sagrada Família worked as a Catholic priest in 1815?

Skiping the rest of the examples...

[NOTE:]

Now please write a question following the instructions and examples above. You should ONLY output the question, and there should be no other output.

[Given info:]

Context: {story}

Question: {question}

Answer: {answer}

[Generated new question:]

Table 1: Average context token counts for each dataset and task category. Token counts are derived using the ‘cl100k_base’ encoder from the ‘tiktoken’ library. Task abbreviations are: Ext. (Extract), Loc. (Localization), Comp. (Computation), D.C. (Duration Compare), O.C. (Order Compare); E.R. (Explicit Reasoning), O.R. (Order Reasoning), R.R. (Relative Reasoning); C.T. (Co-temporality), T.L. (Timeline), C.F. (Counterfactual). A dash (—) indicates that data was not available for the corresponding combination. Note that for TIME-NEWS and TIME-LITE-NEWS, the context token counts represent the average token counts of the top-3 chunks retrieved by three distinct retrievers.

Dataset	Level 1					Level 2			Level 3		
	Ext.	Loc.	Comp.	D.C.	O.C.	E.R.	O.R.	R.R.	C.T.	T.L.	C.F.
TIME											
TIME-WIKI	1157.39	1156.37	1159.06	1188.26	1156.37	1155.05	1150.47	1150.58	1158.69	1156.00	1155.05
TIME-NEWS	—	1473.98	1568.53	1527.80	1499.17	1502.02	1511.58	1496.59	1441.80	1561.75	1515.70
TIME-DIAL	20862.69	20709.06	20667.67	20667.67	20667.67	20667.67	20667.67	20667.67	20667.67	20667.67	20667.67
TIME-LITE											
TIME-LITE-WIKI	1332.37	1332.37	1284.88	1328.86	1332.37	1332.37	1332.37	1332.37	1332.37	1332.37	1332.37
TIME-LITE-NEWS	—	1498.64	1572.68	1500.36	1425.72	1361.89	1494.90	1458.50	1490.09	1523.03	1423.47
TIME-LITE-DIAL	21665.93	21665.93	21665.93	21665.93	21665.93	21665.93	21665.93	21665.93	21665.93	21665.93	21665.93

2.3 Dataset Statistics

3 Supp 2: Benchmark Details

3.1 QA Examples

3.1.1 TIME-WIKI

QA Example. (Extract) (TIME-WIKI) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

Which of the following are time expressions mentioned in the context? (Note: There may be one or more correct options. If you think NONE of the time expressions in options A/B/C/D are mentioned, then you can choose E. Do not choose E together with other options.)

- A. 1930
- B. 2011
- C. 1929
- D. 1932
- E. None of the above.

Answer:

A B

QA Example. (Localization) (TIME-WIKI) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

When did Nicola Agnozzi become an auxiliary bishop?

Answer:

April 2, 1962

QA Example. (Computation) (TIME-WIKI) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

What was the duration from the time George Omaira began to serve as an auxiliary bishop until Nicola Agnozzi became an auxiliary bishop? (Hint: Please answer in the form of Month Day, Year. e.g. 1 year 2 months 3days, or 2 days, or 9 months, or 3 months 15 days.)

Answer:

362 years

QA Example. (Duration Compare) (TIME-WIKI) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

Which of the two events, Fact 1: George Omaira served as an auxiliary bishop. or Fact 2: Maxim Hermaniuk served as an auxiliary bishop., had a longer duration?

- A. Fact 1 lasts longer.
- B. Fact 2 lasts longer.
- C. They last almost the same amount of time.

Answer:

A

148

QA Example. (Order Compare) (TIME-WIKI) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

For Fact1: Mauro Morelli was appointed as both a titular bishop and an auxiliary bishop. and Fact2: Pedro Bantigue y Natividad was appointed as an auxiliary bishop., which one happened earlier?

- A. Fact 1 happened earlier.
- B. Fact 2 happened earlier.
- C. They happen at almost the same time.

Answer:

B

149

QA Example. (Explicit Reasoning) (TIME-WIKI) (Level 2: Temporal Expression Reasoning)

Question:

Who might have held the position of auxiliary bishop from 1902 to February 31, 1916?

Answer:

Edward Joseph Hanna

150

QA Example. (Order Reasoning) (TIME-WIKI) (Level 2: Temporal Expression Reasoning)

Question:

What was the third position Mauro Morelli held in his role?

Answer:

auxiliary bishop

151

QA Example. (Relative Reasoning) (TIME-WIKI) (Level 2: Temporal Expression Reasoning)

Question:

Who was the most recent person to hold the position of titular bishop before Mauro Morelli stepped down as diocesan bishop?

Answer:

José Antonio Eguren

152

QA Example. (Co-temporality) (TIME-WIKI) (Level 3: Complex Temporal Relationship Reasoning)

Question:

When José Antonio Eguren held the position of auxiliary bishop, what position did Mauro Morelli hold?

Answer:

diocesan bishop

153

QA Example. (Timeline) (TIME-WIKI) (Level 3: Complex Temporal Relationship Reasoning)

Question:

Below are 8 facts. You need to sort these facts in chronological order. Requirements: Your output format must be numbers enclosed in parentheses without any other symbols or whitespace. For example: (1)(5)(2)(7)(3)(8)(6)(4)

- (1) Mauro Morelli holds the position of auxiliary bishop.
- (2) Mauro Morelli holds the position of titular bishop.
- (3) Jean-Claude Miche holds the position of titular bishop.
- (4) Belchior Carneiro Leitão holds the position of titular bishop.
- (5) Timothy Norton holds the position of titular bishop.
- (6) Nicola Agnozzi holds the position of titular bishop.
- (7) Edward Joseph Hanna holds the position of titular bishop.
- (8) John Joseph Swint holds the position of auxiliary bishop.

Answer:

(4)(3)(7)(8)(6)(2)(1)(5)

154

QA Example. (Counterfactual) (TIME-WIKI) (Level 3: Complex Temporal Relationship Reasoning)

Question:

Which institutions might Victorina Durán have attended from 1915 to 1934, if she graduated from the Madrid Royal Conservatory in 1915?

Answer:

Royal Academy of Fine Arts of San Fernando

155

156 **3.1.2 TIME-NEWS**

QA Example. (Localization) (TIME-NEWS) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

When is Israeli Prime Minister Benjamin Netanyahu scheduled to address the US Congress?

Answer:

March 3, 2015

157

QA Example. (Computation) (TIME-NEWS) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

How many days passed between the initial opposition to Netanyahu's speech on February 19, 2015, and the announcement of his address to Congress on February 27, 2015? (Hint: Please answer in the form of Month Day, Year. e.g. 1 year 2 months 3days, or 2 days, or 9 months, or 3 months 15 days.)

Answer:

8 days

158

QA Example. (Duration Compare) (TIME-NEWS) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

Which of the following two durations is longer? *Duration 1:* The time span between when political tensions escalated over Netanyahu's planned address to Congress and when the White House announced measures to counter Netanyahu's speeches. *Duration 2:* The time span between when the White House considered limiting communication with Israel and when the White House began strategizing a response to Netanyahu's upcoming speech.

A. Duration 1 is longer.

B. Duration 2 is longer.

C. The two durations are approximately the same length. **Answer:**

B

159

QA Example. (Order Compare) (TIME-NEWS) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

For Fact1: Democratic Party representatives echoed calls to postpone Netanyahu's address. and Fact2: The White House explored strategies to counter Netanyahu's speeches during his trip to Washington., which one happened earlier?

A. Fact 1 happened earlier.

B. Fact 2 happened earlier.

C. They happen at almost the same time.

Answer:

A

160

QA Example. (Explicit Reasoning) (TIME-NEWS) (Level 2: Temporal Expression Reasoning)

Question:

What was the focus of the political controversy surrounding Israeli Prime Minister Benjamin Netanyahu's planned address to Congress in April 2015?

A. The controversy focused on Netanyahu's support for the US-led nuclear negotiations with Iran and his efforts to align with the White House.

B. The controversy focused on Netanyahu's advocacy for increased military aid to Israel and his strained relationship with Congress.

C. The controversy focused on Netanyahu's opposition to the US-led nuclear negotiations with Iran and the perceived lack of coordination with the White House.

D. The controversy focused on Netanyahu's criticism of the Democrats' handling of the Iran nuclear issue and his collaboration with Republican lawmakers **Answer:**

C

161

QA Example. (Order Reasoning) (TIME-NEWS) (Level 2: Temporal Expression Reasoning)

Question:

What was the first action taken by liberal Democrats in response to Netanyahu's planned address to Congress?

- A. They issued a public statement criticizing the speech.
- B. They held a press conference to announce their support for the speech.
- C. They signed a letter requesting the delay of the speech.
- D. They organized a protest against the speech.

Answer:

C

162

QA Example. (Relative Reasoning) (TIME-NEWS) (Level 2: Temporal Expression Reasoning)

Question:

Who publicly reaffirmed their endorsement of Netanyahu on February 21, 2015, within the context of his upcoming speech to Congress?

- A. Susan Rice, Samantha Power.
- B. Hillary Clinton, Bernie Sanders.
- C. Barack Obama, Joe Biden.
- D. Jeb Bush, Ted Cruz

Answer:

D

163

QA Example. (Co-temporality) (TIME-NEWS) (Level 3: Complex Temporal Relationship Reasoning)

Question:

At the same time as Israeli Prime Minister Benjamin Netanyahu's planned address to a joint meeting of Congress, what did liberal Democrats formally request?

- A. the endorsement of Netanyahu's address
- B. the rescheduling of Netanyahu's address
- C. the cancellation of Netanyahu's address
- D. the delay of Netanyahu's address

Answer:

D

164

QA Example. (Timeline) (TIME-NEWS) (Level 3: Complex Temporal Relationship Reasoning)

Question:

Below are 3 facts. You need to sort these facts in chronological order. Requirements: You must output a sequence of uppercase letters separated by commas, such as 'A,B,C', without any other characters.

- A. Democratic leaders expressed concern about the announcement of the Israeli Prime Minister's speech to the joint meeting of the House of Representatives and the Senate without consulting neither the White House nor the Democrats.
- B. A number of Democrats have expressed opposition to Israeli Prime Minister Benjamin Netanyahu's planned address to a joint meeting of Congress in March.
- C. Some Democrats criticize Benjamin Netanyahu, who is mistaken to agree to it and playing politics with the critical issue of Israel's security.

Answer:

B,A,C

165

QA Example. (Counterfactual) (TIME-NEWS) (Level 3: Complex Temporal Relationship Reasoning)

Question:

If no liberal Democrats signed the letter requesting the postponement of Netanyahu's address to Congress from January 2015 to March 2015, who signed the letter?

- A. The National Iranian American Council drafted the letter.
- B. A bipartisan group of lawmakers initiated the letter.
- C. Conservative Democrats signed the letter.
- D. There is no answer

Answer:

D

166

167 **3.1.3 TIME-DIAL**

QA Example. (Extract) (TIME-DIAL) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

Which of the following are time expressions mentioned in the context? (Note: There may be one or more correct options. And the time expressions are mentioned directly or indirectly in the context.)

- A. April 17, 2021
- B. 2018
- C. March 16, 2020
- D. March 14, 2019

Answer:

C

168

QA Example. (Localization) (TIME-DIAL) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

When is Debra Ryan working on starting her own business?

Answer:

8:35 pm, February 21, 2020

169

QA Example. (Computation) (TIME-DIAL) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

How long was it between Debra Ryan going skydiving and India Brown attending a street art fest in Brazil? (Please answer using natural time expressions that combine appropriate units based on duration length, e.g. 2 months 4 days for 64 days or 64 days for shorter spans)

Answer:

19 days

170

QA Example. (Duration Compare) (TIME-DIAL) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

Which of the following two durations is longer? *Duration 1:* Between Debra Ryan is learning to play the guitar. and Debra Ryan visited Adventure Land during the weekend trip. *Duration 2:* Between Debra Ryan rode a roller coaster called The Wild Ride at Adventure Land. and India Brown found flowers by a lake in the park.

- A. Duration 1 is longer.
- B. Duration 2 is longer.
- C. The two durations are approximately the same length.

Answer:

A

171

QA Example. (Order Compare) (TIME-DIAL) (Level 1: Basic Temporal Understanding and Retrieval)

Question:

For Fact1: India Brown became a Queen fan. and Fact2: India Brown found flowers by a lake in the park., which one happened earlier?

- A. Fact 1 happened earlier.
- B. Fact 2 happened earlier.
- C. They happen at almost the same time.

Answer:

A

172

QA Example. (Explicit Reasoning) (TIME-DIAL) (Level 2: Temporal Expression Reasoning)

Question:

What notable artistic or outdoor activities did India Brown participate in between April 1, 2020, and April 9, 2020?

- A. India Brown attended a street art fest in Brazil.
- B. India Brown took a photo of a feather and shells on a beach.
- C. India Brown went hiking and sketching at a nearby national park.
- D. India Brown received positive feedback on her artwork

Answer:

B

173

QA Example. (Order Reasoning) (TIME-DIAL) (Level 2: Temporal Expression Reasoning)

Question:

What was India Brown's third teaching engagement in 2020?

- A. Running a painting workshop for kids.
- B. Teaching art at an orphanage in Cambodia.
- C. Conducting a live demonstration for her college art club.
- D. Instructing a pottery class at a local studio

Answer:

A

174

QA Example. (Relative Reasoning) (TIME-DIAL) (Level 2: Temporal Expression Reasoning)

Question:

What was India Brown's most recent job before 12:00 am, March 09, 2020?

- A. India Brown is working on a new series of abstract artworks based on her trip.
- B. India Brown is working as a travel guide based on her trip experiences.
- C. India Brown is working on a new painting technique learned at a street art festival.
- D. India Brown is testing watercolors for her new series of abstract artworks

Answer:

A

175

QA Example. (Co-temporality) (TIME-DIAL) (Level 3: Complex Temporal Relationship Reasoning)

Question:

At the same time as Debra Ryan is learning to play the guitar, what collection does India Brown have?

- A. India Brown has a collection of soap sculptures.
- B. India Brown has a collection of watercolor paintings.
- C. India Brown has a collection of CDs.
- D. India Brown has a collection of vinyl records

Answer:

C

176

QA Example. (Timeline) (TIME-DIAL) (Level 3: Complex Temporal Relationship Reasoning)

Question:

Below are 8 facts. You need to sort these facts in chronological order. Requirements: Your output format must be numbers enclosed in parentheses without any other symbols or whitespace. For example: (1)(5)(2)(7)(3)(8)(6)(4)

- (1) India Brown is working on a new painting technique learned at a street art festival.
- (2) India Brown shared an image of a mural made by kids.
- (3) India Brown had her first art show at a local gallery.
- (4) India Brown became a Queen fan.
- (5) India Brown got invited to exhibit at a local gallery.
- (6) India Brown took a photo of a feather and shells on a beach.
- (7) India Brown sketched a waterfall during a hike.
- (8) India Brown received positive feedback on her artwork.

Answer:

(4)(5)(1)(7)(6)(2)(8)(3)

177

QA Example. (Counterfactual) (TIME-DIAL) (Level 3: Complex Temporal Relationship Reasoning)

Question:

What notable artistic or outdoor activities did India Brown participate in between April 1, 2020, and April 9, 2020, if she visited the Louvre in Paris in March 2020?

- A. India Brown carved a mini sculpture from a soap bar.
- B. India Brown took a photo of a feather and shells on a beach.
- C. India Brown took a photograph in Santorini, Greece.
- D. India Brown sketched a waterfall during a hike

Answer:

B

178

179 **4 Supp 3: Experimental Details**

180 **4.1 Prompt Templates for Evaluation**

Evaluation Prompt Template for **free-form** tasks (excluding Counterfactual and Computation)

Context: {context}

You need to answer the following question based on the given context. If you can infer the answer from the context, please output your answer directly, keeping it concise and accurate, without any explanatory text. **If you are certain there are multiple answers in the context that satisfy the question, please output all answers, one per line (i.e., separate each answer with a line break).** And you will never refuse to answer any question.

Question: {question}

Therefore, the answer is

181

Evaluation Prompt Template for **multi-choice** tasks

Context: {context}

Instruction: You're an expert in answering multiple-choice questions. You should choose the options that you think is most likely to be correct in the following question. And you will never refuse to answer any question.

Rules:

1. You need to answer the following multiple-choice question based on the given context.
2. You should output the answer in the format of "[[X Y ...]]", WITHOUT anything else, where 'X', 'Y', etc. are the uppercase letters of the correct options. Do not include any other explanatory text in your answer.
3. Example Outputs: (NOTE: The following are only examples, which are NOT relevant to the question and your answer. Your output should be formatted exactly like this.)

Answer: [[A C]]

Answer: [[B]]

Answer: [[B D]]

Question: {question}

Therefore, the answer is

182

Evaluation Prompt Template for **single-choice** tasks (excluding Counterfactual)

Context: {context}

Instruction: You're an expert in answering single-choice questions. You should choose the option that you think is most likely to be correct in the following question. And you will never refuse to answer any question.

Rules:

1. You need to answer the following single-choice question based on the given context.
2. You should output the answer in the format of "[X]", WITHOUT anything else, where 'X' is the choice's uppercase letter. Do not include any other explanatory text in your answer.
3. Example Outputs: (NOTE: The following are only examples, which are NOT relevant to the question. Your output should be formatted exactly like this.)

Answer: [[A]]

Answer: [[B]]

Answer: [[C]]

Answer: [[D]]

Question: {question}

Therefore, the answer is

183

Evaluation Prompt Template for **free-form** Counterfactual

Context: {context}

You need to answer the following question based on the given context. If you can infer the answer from the context, please output your answer directly, keeping it concise and accurate, without any explanatory text. **If you are certain there are multiple answers in the context that satisfy the question, please output all answers, one per line (i.e., separate each answer with a line break).** Otherwise, if there is no answer, simply output "There is no answer."

Hint: The following question is a free-form question. This question is based on a premise that contradicts the temporal information in the original text. You need to fully understand the temporal information in the original text and, while satisfying the false premise in the question, answer the question as accurately as possible. You should not include any explanatory text in your answer, just output the answer directly.

Question: {question}

Therefore, the answer is

184

Evaluation Prompt Template for single-choice Counterfactual

Context: {context}

Instruction: You're an expert in answering single-choice questions. You should choose the option that you think is most likely to be correct in the following question. And you will never refuse to answer any question.

Rules:

1. You need to answer the following single-choice question based on the given context.
2. You should output the answer in the format of "[X]", WITHOUT anything else, where 'X' is the choice's uppercase letter. Do not include any other explanatory text in your answer.
3. Example Outputs: (NOTE: The following are only examples, which are NOT relevant to the question. Your output should be formatted exactly like this.)

Answer: [[A]]

Answer: [[B]]

Answer: [[C]]

Answer: [[D]]

Hint: The following question is a single-choice question. This question is based on a premise that contradicts the temporal information in the original text. The correct option is the one that satisfies the premise (although it contradicts the temporal information in the original text) and satisfies the temporal information in the context. Choose only one option that best aligns with the temporal information in the original text.

Question: {question}

Therefore, the answer is

185

Evaluation Prompt Template for Computation

Context: {context}

You need to answer the following question based on the given context. If you can infer the answer from the context, please output your answer directly, keeping it concise and accurate, without any explanatory text. **If you are certain there are multiple answers in the context that satisfy the question, please output all answers, one per line (i.e., separate each answer with a line break).** And you will never refuse to answer any question.

Hint: The following question is a free-form question. Your answer needs to follow the correct format. Here are some examples of answer formats: 1. March 9, 2020 (format: year-month-day) 2. 1:44 pm, April 16, 2020 (format: hour:minute, year-month-day) 3. 1972 (format: year) 4. April, 2020 (format: month, year) The above examples are only for reference regarding the format of your answer, and are not related to the actual content of your answer. You only need to follow one of the above formats. In terms of content, you need to make your answer as consistent as possible with the original context.

Question: {question}

Therefore, the answer is

186

References

187

- 188 [1] Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. Analyzing
189 temporal complex events with large language models? A benchmark towards temporal, long
190 context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings*
191 *of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*
192 *Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 1588–1606. Association for
193 Computational Linguistics, 2024.

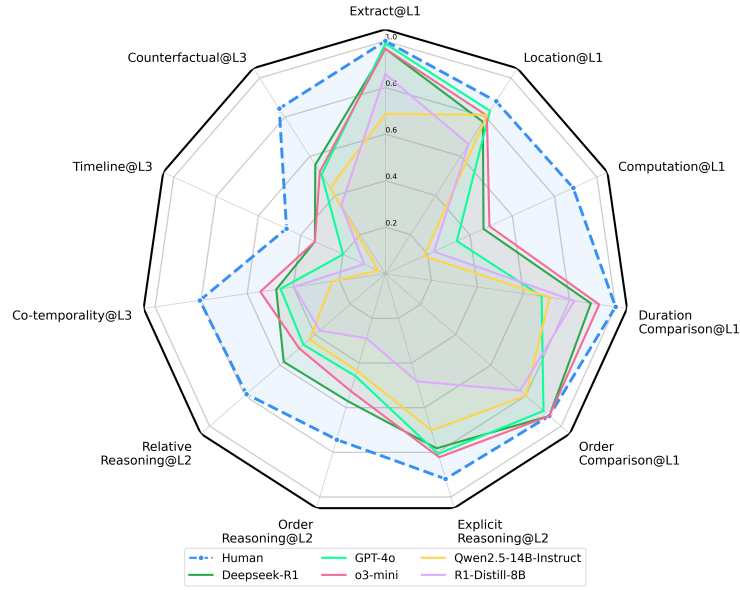


Figure 1: Performance of Human, GPT-4o, Qwen2.5-14B-Instruct, Deepseek-R1, o3-mini, and Deepseek-R1-Distill-Llama-8B on various subtasks of TIME-WIKI.

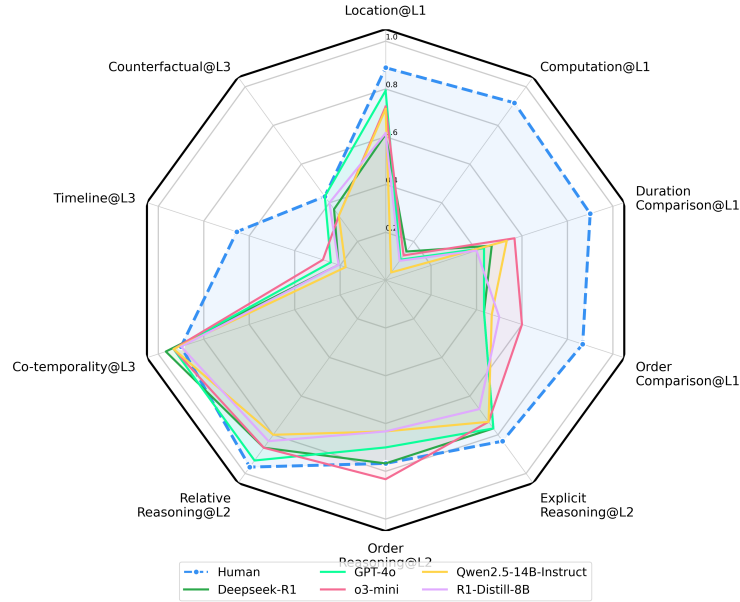


Figure 2: Performance of Human, GPT-4o, Qwen2.5-14B-Instruct, Deepseek-R1, o3-mini, and Deepseek-R1-Distill-Llama-8B on various subtasks of TIME-NEWS.

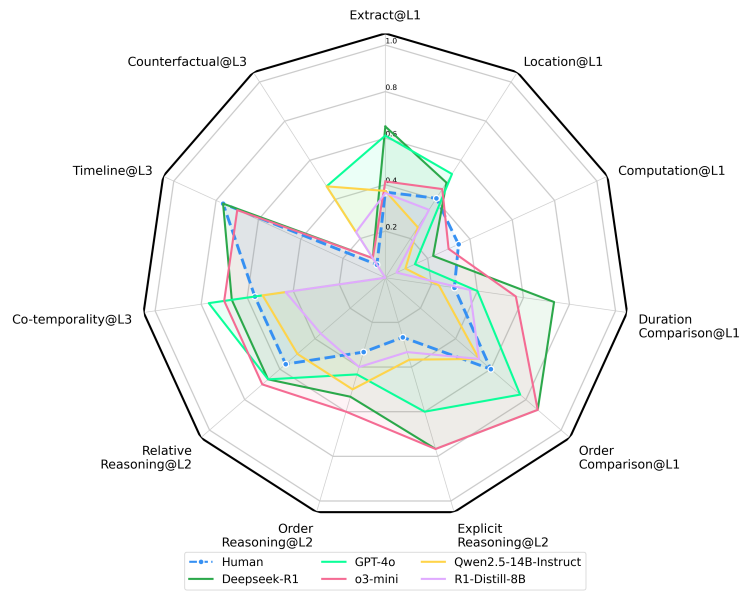


Figure 3: Performance of Human, GPT-4o, Qwen2.5-14B-Instruct, Deepseek-R1, o3-mini, and Deepseek-R1-Distill-Llama-8B on various subtasks of TIME-DIAL.