1. COLLECT INITIAL DATA

We have secured two pivotal datasets, TORQUE and TempReason, that are essential for our project aimed at enhancing temporal reasoning in Large Language Models (LLMs). These datasets will be used in their entirety, as every attribute they contain is crucial for a comprehensive understanding and evaluation of the LLMs' capabilities in temporal reasoning.

1.1 Data Sources

- TORQUE dataset: A time-event relation dataset sourced from GitHub.
- TempReason Dataset: A temporal reasoning dataset accessible at Hugging Face Datasets.

Fields Necessity:
The datasets provide extensive coverage of temporal relations, with TORQUE focusing on time-event relationships, and TempReason providing time-time and event-event relationships. The fields include passages, questions, events, and temporal facts that are critical for our analysis.

1.2 Temporal Reasoning Analysis

Temporal reasoning in LLMs pertains to various event types and their interrelations, notably:

- Time-Time Relation (L1): Understanding relationships between different time points.
- Time-Event Relation (L2): Associating events with specific times or intervals.
- Event-Event Relation (L3): Inferring the sequence and causality of events.

1.3 Selection Criterion

We will employ all attributes in the TORQUE and TempReason datasets as they are necessary for our data mining goals. The TORQUE dataset will assist the analysis of L2 (Time-Event Relation), while TempReason will aid in the examination of L1 (Time-Time Relation) and L3 (Event-Event Relation).

Data Insertion and Querying:
The datasets will be queried and inserted using Python and libraries such as pandas for handling CSV data. This approach enables efficient extraction and manipulation of data, preparing it for analysis against the LLM outputs.

1.4 Large Language Model Capabilities Analysis

LLMs will be assessed on their ability to interpret and reason with temporal information. This includes their precision in understanding event sequences, temporal intervals, and causality.

Evaluation of LLM Outputs:
Outputs from LLMs will be evaluated using methods like fuzzy matching, alongside other metrics such as precision, recall, and F1 score, to quantitatively and qualitatively assess the models' temporal reasoning abilities.

2. DESCRIBE DATA

Data sets we plan to use:

1) **TORQUE:**
    a) Background
        i) TORQUE is a time-event relation dataset intended to be used to test the reading comprehension and temporal reasoning abilities of LLMs
    b) Basic Statistics

Type	Subtype	Example	%
Standard		*"What happened before Bush gave four key speeches?"*	53%
Fuzzy	begin only	*"What started before Mr. Fournier was prohibited from organizing his own defense?"*	15%
	overlap only	*"What events were occurring during the competition?"*	10%
	end only	*"What will end after he is elected?"*	1%
Modality	uncertain	*"What might happen after the FTSE 100 index was quoted 9.6 points lower?"*	10%
	negation	*"What has not taken place before the official figures show something?"*	5%
	hypothetical	*"What event will happen if the scheme is broadened?"*	2%
	repetitive	*"What usually happens after common shares are acquired?"*	1%
Misc.	participant	*"What did Hass do before he went to work as a spy?"*	4%
	opinion	*"What should happen in the future according to Obama's opinion?"*	3%
	intention	*"What did Morales want to happen after Washington had a program to eradicate coca?"*	1%

Table 2: Temporal phenomena in TORQUE. "Standard" are those that can be directly captured by the previous single-interval-based label set, while other types cannot. Percentages are based on manual inspection of a random sample of 200 questions from TORQUE; some questions can have multiple types.

        i) 3.2k passage annotations: Each passage contains approximately 50 tokens.
        ii) 24.9k events: This averages to about 7.9 events per passage.
        iii) 21.2k user-provided questions: Roughly half of these questions were labeled by crowd workers as modifications of existing ones.
    c) Structure:
        i) Data Types
            (1) Both the passage and QA are both strings
        ii) Features
            (1) Passage - The body of text fed into the LLM as context for a question
            (2) QA blocks - A question and set of answers paired together for the LLM to answer
    d) Misc
        i) This data set was recommended to us by the project owner's team
        ii) Rebalancing is not applicable for this set

2) TempReason:
    a) Background
        i) TempReason is temporal reasoning dataset structured as a QA set focused on probing a models ability to recall dates and determine the temporal position of events
    b) Basic Statistics

	Train	Dev	Test
Time Range	1014-2022	634-2023	998-2023
L1-Questions	400,000	4,000	4,000
L2-Questions	16,017	5,521	5,397
L3-Questions	13,014	4,437	4,426
Subjects	3,000	1,000	1,000
Facts	16,017	5,521	5,397
Facts/subjects	5.3	5.5	5.4

Table 3: Dataset statistics of TEMPREASON.

c) Structure:
   i) Data Types
      (1) The question is of type string, the answer is of type datetime
   ii) Features
      (1) Question - A question for the LLM to answer based on a date
      (2) Answer - The time which corrusponds to the answer of the question
      (3) Example
         (a) Q: What is 3 days after Feburary 12th, 2023
         (b) A: February 15th, 2023

d) Misc
   i) This data set was recommended to us by the project owner's team
   ii) Rebalancing is not applicable for this set

After conducting a preliminary analysis on the data sets we plan to use. There is no indication that we will require a change in our assumptions or a pivot of any kind.

3. EXPLORE DATA

Since the data we are using does not have any numeric properties and is just for the purpose of LLM evaluation, we will not be able to conduct a hypothesis. Instead, I will go more into depth on how these evaluation sets were created.

3.1 TempReason Dataset

The TempReason dataset for temporal reasoning was curated through a multi-step process that involved leveraging the structured knowledge base (KB) Wikidata.

Wikidata Knowledge Base Processing: The process started with extracting structured information from the Wikidata knowledge base. Wikidata is a collaborative project that provides a free and open knowledge base where structured data about entities can be stored and edited.

Data Representation: The extracted data from Wikidata was represented in a structured format denoted as (s, r, o, t_s, t_e), where:

s: Represents the subject entity.

r: Represents the relation between the subject and the object.

o: Represents the object entity.
t_s: Represents the start time of the temporal fact.
t_e: Represents the end time of the temporal fact.
Temporal Facts Grouping: The temporal facts extracted from Wikidata were then grouped based on common subject (s) and relation (r). This grouping allowed for organizing related temporal information together, facilitating the creation of coherent datasets for temporal reasoning tasks.

Dataset Creation: The grouped temporal facts were compiled and organized into the TempReason dataset. This dataset serves as a valuable resource for research and development in temporal reasoning, providing structured temporal information suitable for various computational tasks and analysis.

3.2 TORQUE Dataset

Source Material Selection: Utilize 3.2k news snippets as the basis for the dataset to ensure a diverse and realistic representation of events and their temporal relationships.

Annotation Approach Development: Shift from using a fixed set of relation labels to natural language annotations for describing temporal relations, to better capture the complexities of temporal relationships, including fuzzy relations, modalities, and repetitive events.

3.3 Data Collection Process

Initial Pool Creation: Gather a pool of 26k two-sentence passages from 2.8k articles used in the TempEval3 workshop, sufficient for capturing non-trivial temporal relations.
Amazon Mechanical Turk Utilization: Engage crowd workers through Amazon Mechanical Turk for the annotation process, ensuring a wide and diverse data collection effort.

Annotation Task Design:

- Event Labeling: Annotators label all events in the passages.
- Temporal Relation Questioning: Formulate questions regarding the temporal relations between labeled events.
- Question Modification: Create new questions by modifying the temporal relation to ensure comprehensive coverage and to penalize potential shortcuts by contrasting questions.

Quality Control Measures: Implement multiple strategies for quality assurance, including:

- Qualification: Design a separate qualification task to test annotators on their ability to label events, ask temporal relation questions, and answer them.
- Pilot Annotations: Conduct a pilot phase where initial annotations are manually checked and feedback is provided.
- Validation: Use validation steps to ensure the accuracy and reliability of the dataset annotations.