Time-Sensitive Question-Answering Beyond LLMs’ Memorization

UnSeenTimeQA

ACL 2025

HUMANE Lab 윤예준

25.09.25

Time-Sensitive Question-Answering (TSQA)

- A temporal-sensitive question is a question whose answer varies depending on specific temporal information (e.g., dates, durations, events)

- Examples:
    - Question 1: “Which football club did Leo Messi play for in 2010?” Answer: FC Barcelona
    - Question 2: “Which football club did Leo Messi play for in 2023?” Answer: Inter Miami CF
    - Question 3: “Which football club did Leo Messi play for after FC Barcelona?” Answer: Paris Saint-Germain

- In general, answering time-sensitive questions
    - Requires knowledge about entities at specific points in time
    - Involves reasoning over multiple events

Issues Existing TSQA Benchmarks

- Benchmarks developed before the LLMs' knowledge cut-off date
  - Data Contamination
    - Include questions based on real-world facts contained in the LLMs' pre-training data (e.g., from Wikipedia)
    - LLMs can answer correctly by relying on memorized facts acquired during the pre-training

Issues Existing TSQA Benchmarks

- Benchmarks developed after the LLMs' knowledge cut-off date
    - Periodic manual updates
        - Benchmarks developed after the LLMs' knowledge cut-off date must be continuously curated to remain "unseen" by newer LLMs
        - Once updates cases, these benchmarks also risk becoming contaminated

Do Existing TSQA Benchmarks Address Temporal Reasoning?

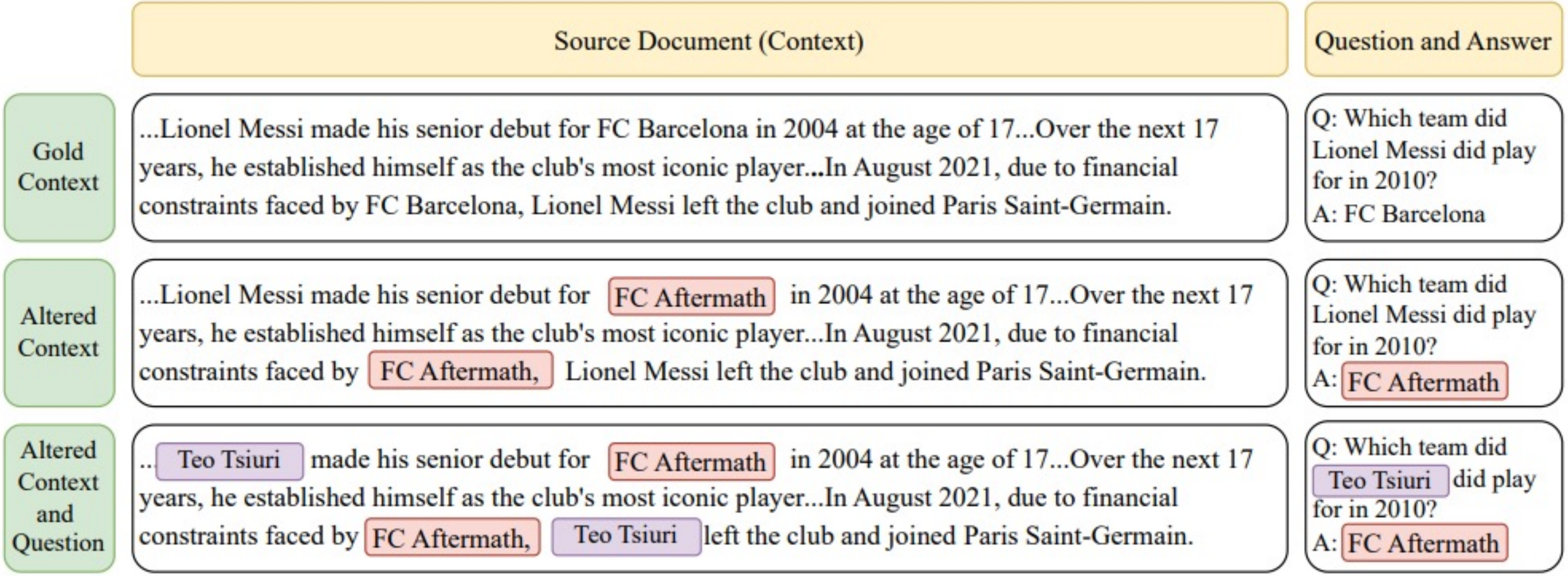
- Benchmarks developed before the LLM knowledge cut-off date
  - TimeQA, TempReason, and MenatQA derive questions from real-world facts using Wikidata
  - Vulnerable to data contamination — gold contexts were public during LLM pre-training

- Benchmarks developed after the LLM knowledge cut-off date
  - FreshQA, RealTimeQA, and TAQA target events after LLM Knowledge cut-off to avoid overlap with pre-training data
  - Depend on manual periodic updates

Experiments with Benchmarks developed before LLM knowledge cut-off date

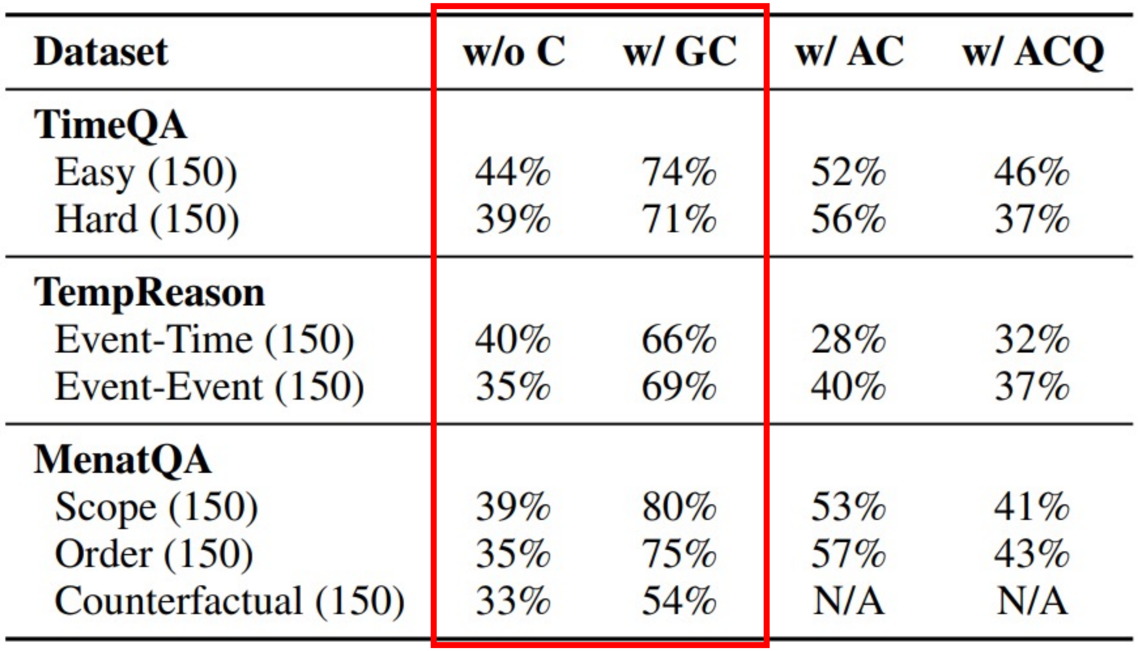
- Benchmarks: TimeQA, TempReason, and MenatQA
- Experimental Setup:
  - They randomly sampled 150 questions per split, for a total of 1,050 questions
  - Evaluated a GPT-4 model under four prompting conditions:
    - No Context: Model is given only the question, with no additional context
    - Gold Context: Model is given the original context containing all facts needed to answer
    - Altered Context: Correct answers in the context swapped with plausible dummy answers
    - Altered Context + Altered Question: Both the context and the question's key entities are replaced with entirely novel entities

An Example of Altering Entity in a Document



	Source Document (Context)	Question and Answer
Gold Context	...Lionel Messi made his senior debut for FC Barcelona in 2004 at the age of 17...Over the next 17 years, he established himself as the club's most iconic player...In August 2021, due to financial constraints faced by FC Barcelona, Lionel Messi left the club and joined Paris Saint-Germain.	Q: Which team did Lionel Messi did play for in 2010? A: FC Barcelona
Altered Context	...Lionel Messi made his senior debut for FC Aftermath in 2004 at the age of 17...Over the next 17 years, he established himself as the club's most iconic player...In August 2021, due to financial constraints faced by FC Aftermath, Lionel Messi left the club and joined Paris Saint-Germain.	Q: Which team did Lionel Messi did play for in 2010? A: FC Aftermath
Altered Context and Question	...Teo Tsiuri made his senior debut for FC Aftermath in 2004 at the age of 17...Over the next 17 years, he established himself as the club's most iconic player...In August 2021, due to financial constraints faced by FC Aftermath, Teo Tsiuri left the club and joined Paris Saint-Germain.	Q: Which team did Teo Tsiuri did play for in 2010? A: FC Aftermath

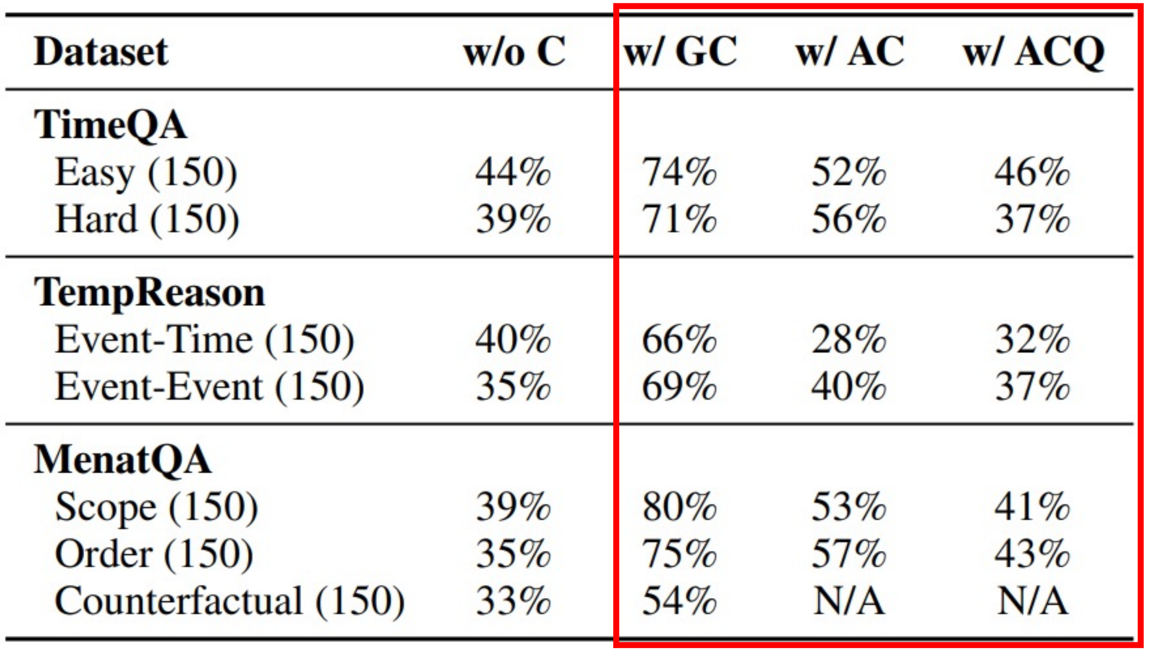
Key Findings



Dataset	w/o C	w/ GC	w/ AC	w/ ACQ
TimeQA				
Easy (150)	44%	74%	52%	46%
Hard (150)	39%	71%	56%	37%
TempReason				
Event-Time (150)	40%	66%	28%	32%
Event-Event (150)	35%	69%	40%	37%
MenatQA				
Scope (150)	39%	80%	53%	41%
Order (150)	35%	75%	57%	43%
Counterfactual (150)	33%	54%	N/A	N/A

- This performance drop across all three benchmarks hints at possible memorization of facts when answering time-sensitive questions
- Human reviewers found the correct answers via simple web searches in 88~98.8% of cases, confirming high contamination risk

Key Findings



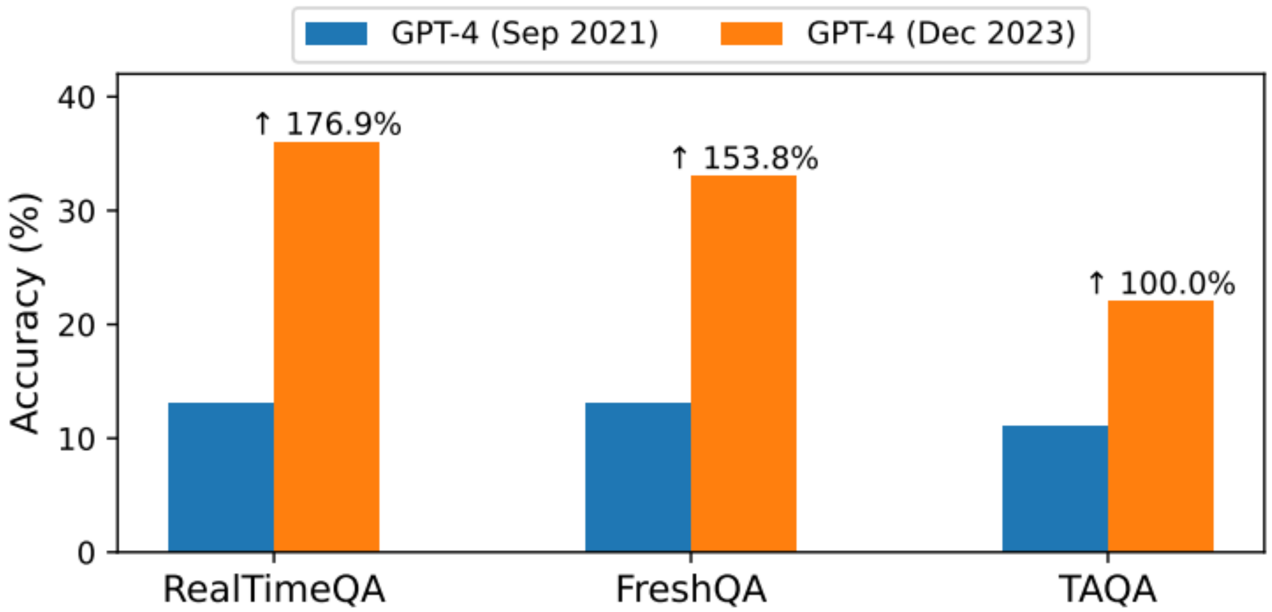
Dataset	w/o C	w/ GC	w/ AC	w/ ACQ
TimeQA				
Easy (150)	44%	74%	52%	46%
Hard (150)	39%	71%	56%	37%
TempReason				
Event-Time (150)	40%	66%	28%	32%
Event-Event (150)	35%	69%	40%	37%
MenatQA				
Scope (150)	39%	80%	53%	41%
Order (150)	35%	75%	57%	43%
Counterfactual (150)	33%	54%	N/A	N/A

- This performance drop across all three benchmarks hints at possible memorization of facts when answering time-sensitive questions
- Human reviewers found the correct answers via simple web searches in 88~98.8% of cases, confirming high contamination risk

Experiments with Benchmarks developed after LLM knowledge cut-off date

- Benchmarks: FreshQA, RealTimeQA, and TAQA
- Experimental Setup:
  - Randomly sampled 150 questions per benchmark, total 450 questions
  - Evaluated two GPT-4 versions (Sep 2021 cutoff vs. Dec 2023 cutoff) without any context

Key Findings

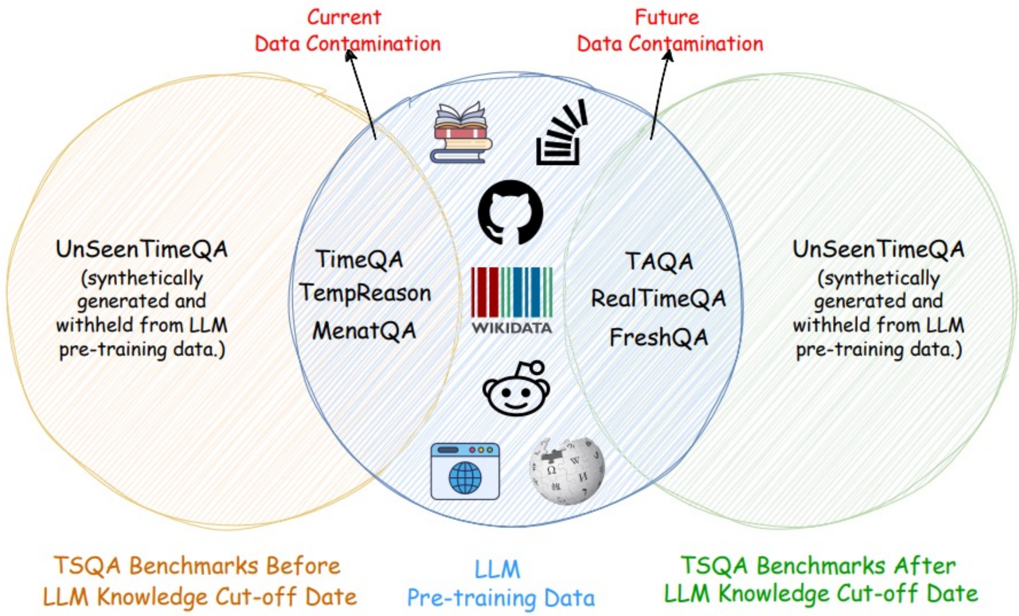


- Once new LLMs ingest updated corpora, these benchmarks become ‘contaminated’ again

A more reliable TSQA benchmark is needed to avoid data contamination and the need for frequent manual updates for time-sensitive questions

UnSeenTimeQA

- A contamination-free TSQA benchmark
  - Decoupling memorization from temporal reasoning
  - Created using synthetically generated facts



Current Data Contamination

Future Data Contamination

UnSeenTimeQA (synthetically generated and withheld from LLM pre-training data.)

TimeQA
TempReason
MenatQA

TAQA
RealTimeQA
FreshQA

UnSeenTimeQA (synthetically generated and withheld from LLM pre-training data.)

TSQA Benchmarks Before LLM Knowledge Cut-off Date

LLM Pre-training Data

TSQA Benchmarks After LLM Knowledge Cut-off Date

Data Source

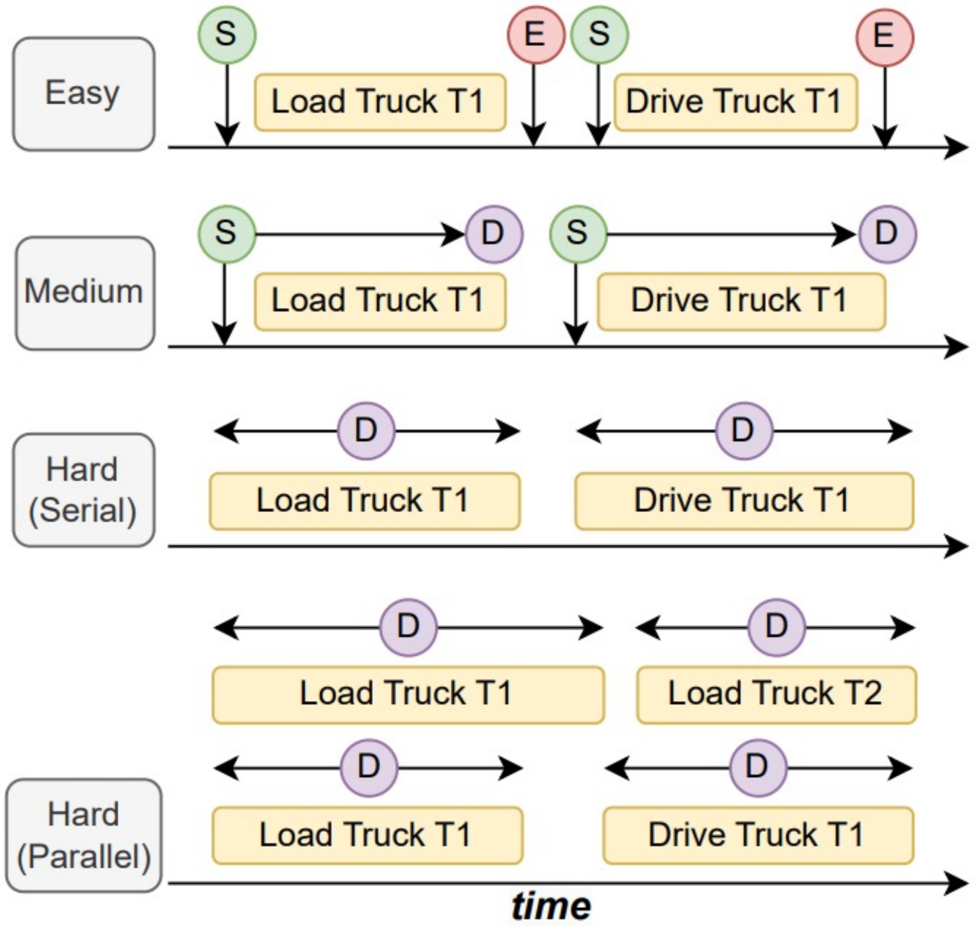
- Domain Selection
  - Logistics domain from International Planning Competitions (IPC)
  - Event Types: Six event define package movement

Load Truck
Drive Truck
Unload Truck
Load Airplane
Fly Airplane
Unload Airplane

  - Environment: multiple cities and locations; vehicles: trucks and airplanes; packages
  - Goal: Moving the packages from initial location to the goal location

Difficulty Levels

- Easy: explicit start/end times
  - (e.g., 08:11 AM to 08:54 PM)
- Medium: start time + duration
  - (e.g., 08:11 AM, 43 minutes)
- Hard-Serial: durations only (sequential)
- Hard-Parallel: durations only (overlapping)
  - (e.g., 43 minutes)



Question and Answers Generation

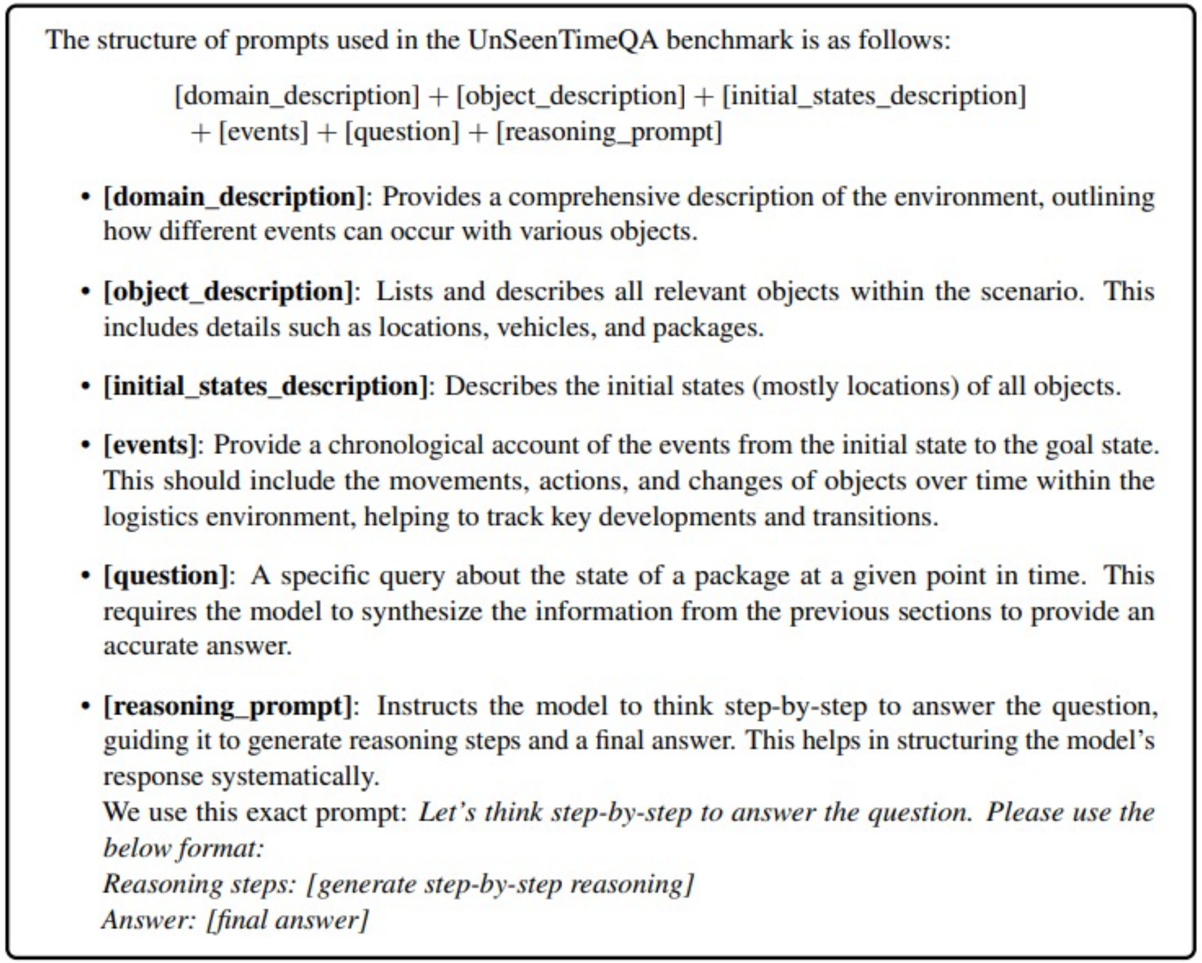
- Static-Time: Questions ask for a package's location at a specific absolute timestamp
  - Example Question: Where is package p0 at 10:53 PM?
- Relative-Time: Questions ask for a package's location at a time offset from a given timestamp
  - Example Question: Where is package p0 2 hours after 8:13 PM?
- Hypothetical-Time: Questions ask for a package's location after altering an event's duration
  - Example Question: If driving truck t1 from location l1_1 to location l1_0 is delayed by 66 minutes, where is the product p0 at 10:18 PM?

- Answer Generation Rules
  - In transit → vehicle id
  - Loading/unloading → both location & vehicle id

Experimental Setup

- Models:
  - Open-weight: Gemma-2-9B, Gemma-2-27B, Llama-3.1-8B, Llama-3.1-70B
  - Closed-weight: GPT-4o Metric: Accuracy (correct final answer)
- Prompting:
  - Zero-shot chain-of-thought and few-shot

Prompt for Evaluating UnSeenTimeQA



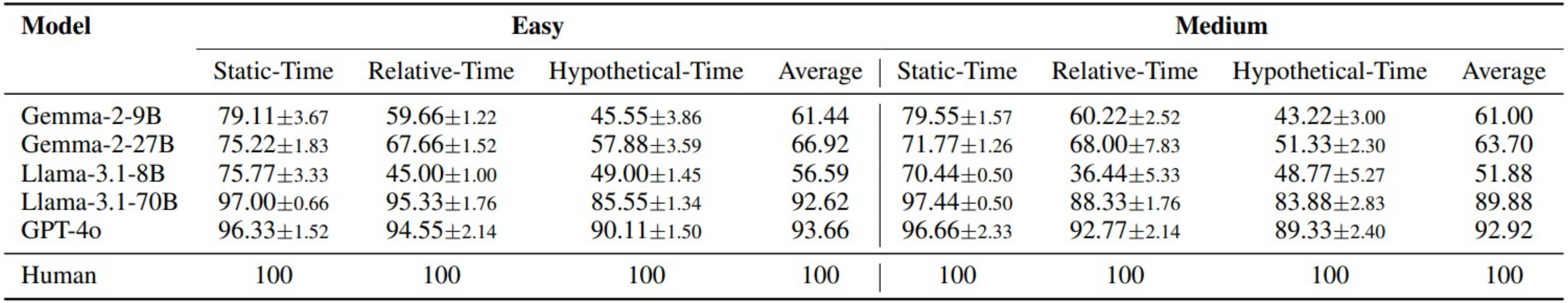
The structure of prompts used in the UnSeenTimeQA benchmark is as follows:

[domain_description] + [object_description] + [initial_states_description]
+ [events] + [question] + [reasoning_prompt]

- **[domain_description]**: Provides a comprehensive description of the environment, outlining how different events can occur with various objects.
- **[object_description]**: Lists and describes all relevant objects within the scenario. This includes details such as locations, vehicles, and packages.
- **[initial_states_description]**: Describes the initial states (mostly locations) of all objects.
- **[events]**: Provide a chronological account of the events from the initial state to the goal state. This should include the movements, actions, and changes of objects over time within the logistics environment, helping to track key developments and transitions.
- **[question]**: A specific query about the state of a package at a given point in time. This requires the model to synthesize the information from the previous sections to provide an accurate answer.
- **[reasoning_prompt]**: Instructs the model to think step-by-step to answer the question, guiding it to generate reasoning steps and a final answer. This helps in structuring the model's response systematically.
  We use this exact prompt: *Let's think step-by-step to answer the question. Please use the below format:*
  Reasoning steps: [generate step-by-step reasoning]
  Answer: [final answer]

Easy and Medium Difficulty Results

- Easy (start and end timestamps)
- Medium (start timestamps and the duration)

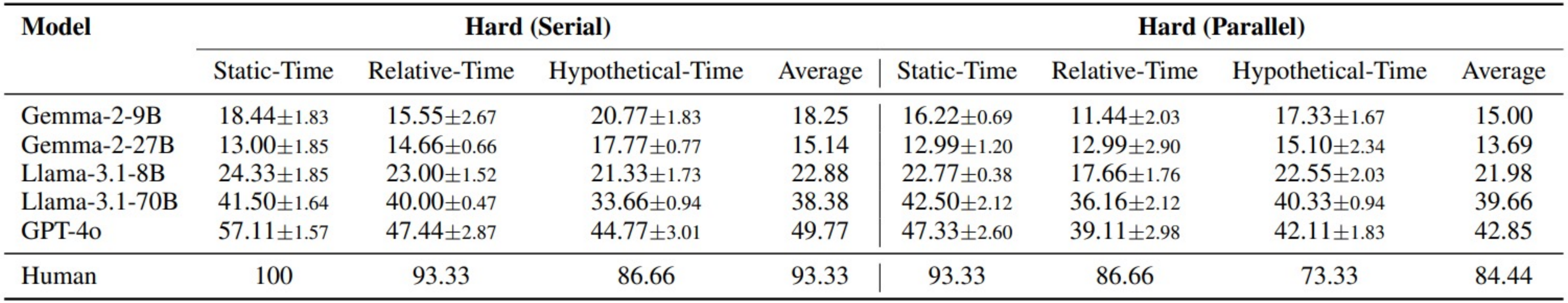


Model	Easy				Medium			
	Static-Time	Relative-Time	Hypothetical-Time	Average	Static-Time	Relative-Time	Hypothetical-Time	Average
Gemma-2-9B	79.11±3.67	59.66±1.22	45.55±3.86	61.44	79.55±1.57	60.22±2.52	43.22±3.00	61.00
Gemma-2-27B	75.22±1.83	67.66±1.52	57.88±3.59	66.92	71.77±1.26	68.00±7.83	51.33±2.30	63.70
Llama-3.1-8B	75.77±3.33	45.00±1.00	49.00±1.45	56.59	70.44±0.50	36.44±5.33	48.77±5.27	51.88
Llama-3.1-70B	97.00±0.66	95.33±1.76	85.55±1.34	92.62	97.44±0.50	88.33±1.76	83.88±2.83	89.88
GPT-4o	96.33±1.52	94.55±2.14	90.11±1.50	93.66	96.66±2.33	92.77±2.14	89.33±2.40	92.92
Human	100	100	100	100	100	100	100	100

All models handle explicit temporal data well

Hard Difficulty Results

- Hard-Serial (durations only, sequential)
- Hard-Parallel (durations only, parallel)

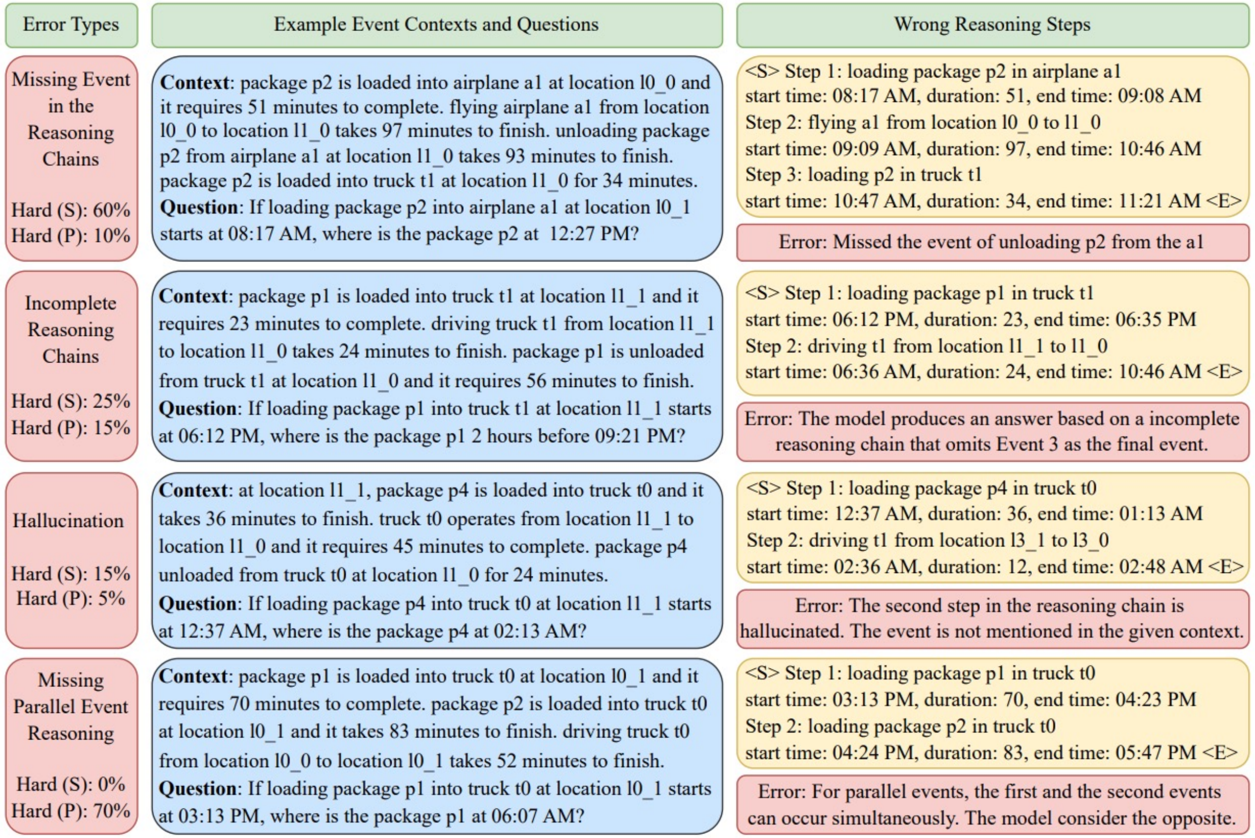


Model	Hard (Serial)				Hard (Parallel)			
	Static-Time	Relative-Time	Hypothetical-Time	Average	Static-Time	Relative-Time	Hypothetical-Time	Average
Gemma-2-9B	18.44±1.83	15.55±2.67	20.77±1.83	18.25	16.22±0.69	11.44±2.03	17.33±1.67	15.00
Gemma-2-27B	13.00±1.85	14.66±0.66	17.77±0.77	15.14	12.99±1.20	12.99±2.90	15.10±2.34	13.69
Llama-3.1-8B	24.33±1.85	23.00±1.52	21.33±1.73	22.88	22.77±0.38	17.66±1.76	22.55±2.03	21.98
Llama-3.1-70B	41.50±1.64	40.00±0.47	33.66±0.94	38.38	42.50±2.12	36.16±2.12	40.33±0.94	39.66
GPT-4o	57.11±1.57	47.44±2.87	44.77±3.01	49.77	47.33±2.60	39.11±2.98	42.11±1.83	42.85
Human	100	93.33	86.66	93.33	93.33	86.66	73.33	84.44

Models struggle when inferring timings purely from durations

Error Analysis on UnSeenTimeQA

- Manual Error Analysis on the best performing model (GPT-4o)



Error Types	Example Event Contexts and Questions	Wrong Reasoning Steps
Missing Event in the Reasoning Chains Hard (S): 60% Hard (P): 10%	**Context**: package p2 is loaded into airplane a1 at location l0_0 and it requires 51 minutes to complete. flying airplane a1 from location l0_0 to location l1_0 takes 97 minutes to finish. unloading package p2 from airplane a1 at location l1_0 takes 93 minutes to finish. package p2 is loaded into truck t1 at location l1_0 for 34 minutes. **Question**: If loading package p2 into airplane a1 at location l0_1 starts at 08:17 AM, where is the package p2 at 12:27 PM?	<S> Step 1: loading package p2 in airplane a1 start time: 08:17 AM, duration: 51, end time: 09:08 AM Step 2: flying a1 from location l0_0 to l1_0 start time: 09:09 AM, duration: 97, end time: 10:46 AM Step 3: loading p2 in truck t1 start time: 10:47 AM, duration: 34, end time: 11:21 AM <E> Error: Missed the event of unloading p2 from the a1
Incomplete Reasoning Chains Hard (S): 25% Hard (P): 15%	**Context**: package p1 is loaded into truck t1 at location l1_1 and it requires 23 minutes to complete. driving truck t1 from location l1_1 to location l1_0 takes 24 minutes to finish. package p1 is unloaded from truck t1 at location l1_0 and it requires 56 minutes to finish. **Question**: If loading package p1 into truck t1 at location l1_1 starts at 06:12 PM, where is the package p1 2 hours before 09:21 PM?	<S> Step 1: loading package p1 in truck t1 start time: 06:12 PM, duration: 23, end time: 06:35 PM Step 2: driving t1 from location l1_1 to l1_0 start time: 06:36 AM, duration: 24, end time: 10:46 AM <E> Error: The model produces an answer based on a incomplete reasoning chain that omits Event 3 as the final event.
Hallucination Hard (S): 15% Hard (P): 5%	**Context**: at location l1_1, package p4 is loaded into truck t0 and it takes 36 minutes to finish. truck t0 operates from location l1_1 to location l1_0 and it requires 45 minutes to complete. package p4 unloaded from truck t0 at location l1_0 for 24 minutes. **Question**: If loading package p4 into truck t0 at location l1_1 starts at 12:37 AM, where is the package p4 at 02:13 AM?	<S> Step 1: loading package p4 in truck t0 start time: 12:37 AM, duration: 36, end time: 01:13 AM Step 2: driving t1 from location l3_1 to l3_0 start time: 02:36 AM, duration: 12, end time: 02:48 AM <E> Error: The second step in the reasoning chain is hallucinated. The event is not mentioned in the given context.
Missing Parallel Event Reasoning Hard (S): 0% Hard (P): 70%	**Context**: package p1 is loaded into truck t0 at location l0_1 and it requires 70 minutes to complete. package p2 is loaded into truck t0 at location l0_1 and it takes 83 minutes to finish. driving truck t0 from location l0_0 to location l0_1 takes 52 minutes to finish. **Question**: If loading package p1 into truck t0 at location l0_1 starts at 03:13 PM, where is the package p1 at 06:07 AM?	<S> Step 1: loading package p1 in truck t0 start time: 03:13 PM, duration: 70, end time: 04:23 PM Step 2: loading package p2 in truck t0 start time: 04:24 PM, duration: 83, end time: 05:47 PM <E> Error: For parallel events, the first and the second events can occur simultaneously. The model consider the opposite.

Conclusion

- **Contamination-Free by Design:** Uses synthetically generated facts from a logistics planning domain — not found on the web, so LLMs cannot have memorized them
- **Focus on Temporal Reasoning:** Forces models to reason about event timelines from scratch
- **Complex Scenarios:** Introduces parallel events where multiple things happen at once