NAACL 2025

Transient Tables : Evaluating LLMs' Reasoning on Temporally Evolving Semi-structured Tables

Abhilash Shankarampeta*, Harsh Mahajan*, Tushar Kataria, Dan Roth, Vivek Gupta

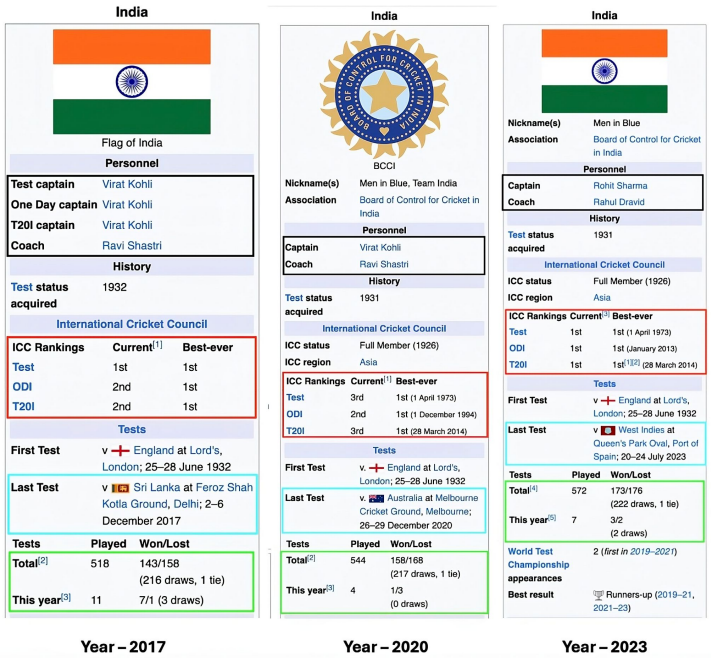
Affiliations: UC San Diego, University of Utah,

University of Pennsylvania, Arizona state university

* Equal Contribution

Motivation

- **Information is inherently transient and constantly updated**
    - Examples: company profits, political figures, sports rankings, etc.
- LLMs are typically trained on static datasets
- ***Research question:*** *Can LLMs effectively reason over temporal changes in information through in-context learning?*



Sample question - How many Test matches did the Indian Cricket Team play between 2020 and 2023?

Why is it Challenging?

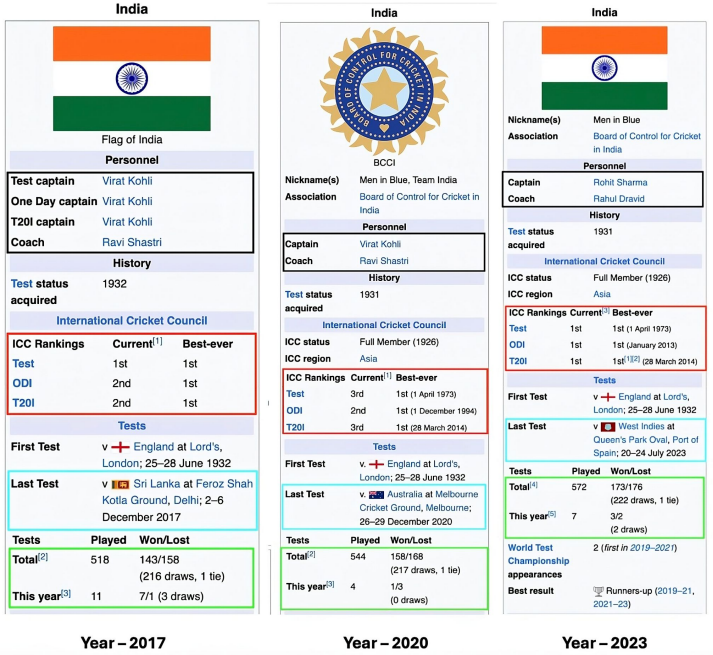
- Temporal Context Understanding
    - Models must ground questions in the right time period.
    - *"Who was captain before Rohit Sharma?"* requires identifying when Rohit became captain first

- Multi-Table Integration
    - Reasoning across 8-12 tables per entity timeline.
    - Tracking changes over time while maintaining consistency.

Why is it Challenging?

- Subtle Changes Detection
  - Key values change incrementally (e.g., Tests played: 518 → 544 → 572)
  - Easy to confuse values from different time periods.

- Token Length Limitations
  - Full entity timelines push context window limits.
  - Requires efficient attention to relevant information.



India

Flag of India

Personnel	
Test captain	Virat Kohli
One Day captain	Virat Kohli
T20I captain	Virat Kohli
Coach	Ravi Shastri
History	
Test status acquired	1932

International Cricket Council		
ICC Rankings	Current[1]	Best-ever
Test	1st	1st
ODI	2nd	1st
T20I	2nd	1st

Tests	
First Test	v England at Lord's, London; 25–28 June 1932
Last Test	v Sri Lanka at Feroz Shah Kotla Ground, Delhi; 2–6 December 2017

Tests	Played	Won/Lost
Total[2]	518	143/158 (216 draws, 1 tie)
This year[3]	11	7/1 (3 draws)

Year – 2017

India

BCCI

Nickname(s)	Men in Blue, Team India
Association	Board of Control for Cricket in India
Personnel	
Captain	Virat Kohli
Coach	Ravi Shastri
History	
Test status acquired	1931

International Cricket Council		
ICC status	Full Member (1926)	
ICC region	Asia	
ICC Rankings	Current[1]	Best-ever
Test	3rd	1st (1 April 1973)
ODI	2nd	1st (1 December 1994)
T20I	3rd	1st (28 March 2014)

Tests	
First Test	v. England at Lord's, London; 25–28 June 1932
Last Test	v. Australia at Melbourne Cricket Ground, Melbourne; 26–29 December 2020

Tests	Played	Won/Lost
Total[2]	544	158/168 (217 draws, 1 tie)
This year[3]	4	1/3 (0 draws)

Year – 2020

India

Nickname(s)	Men in Blue
Association	Board of Control for Cricket in India
Personnel	
Captain	Rohit Sharma
Coach	Rahul Dravid
History	
Test status acquired	1931

International Cricket Council		
ICC status	Full Member (1926)	
ICC region	Asia	
ICC Rankings	Current[3]	Best-ever
Test	1st	1st (1 April 1973)
ODI	1st	1st (January 2013)
T20I	1st	1st[1][2] (28 March 2014)

Tests	
First Test	v England at Lord's, London; 25–28 June 1932
Last Test	v West Indies at Queen's Park Oval, Port of Spain; 20–24 July 2023

Tests	Played	Won/Lost
Total[4]	572	173/176 (222 draws, 1 tie)
This year[5]	7	3/2 (2 draws)
World Test Championship appearances	2 (*first in 2019–2021*)	
Best result	Runners-up (2019–21, 2021–23)	

Year – 2023

Key Contributions

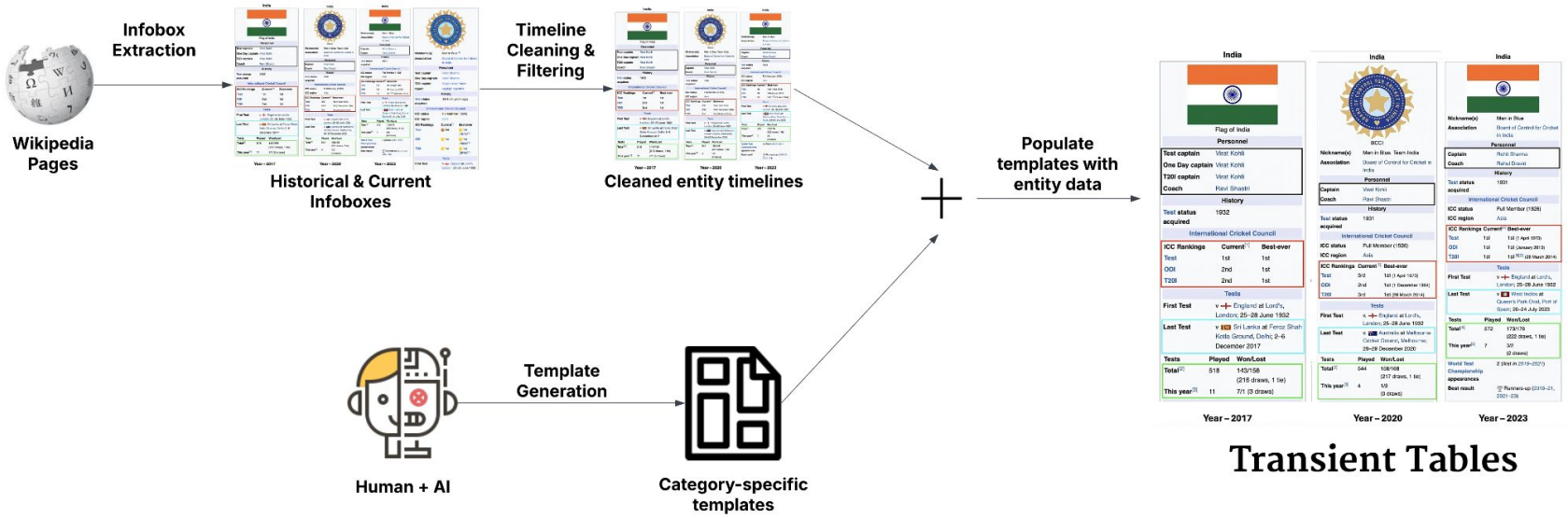
- Transient Tables Dataset
  - A novel QA dataset with **3,971 questions** from **14,000+ tables** spanning **1,238 entities** with **11 tables per entity** (on avg)
  - **Template-based question** generation pipeline using LLMs'

- Baseline results with **state-of-the-art models**
  - **GPT-4o, Llama3-70B, Gemini 1.5, GPT-4o-mini, Llama3-8B, Mixtral**

- Novel modeling strategies using **task decomposition** to enhance performance

Dataset Creation

- Entity timeline selection from Wikipedia infoboxes
- Timeline cleaning and filtering (8-12 tables per entity)
- Query-answer generation through templates



Question Categorization

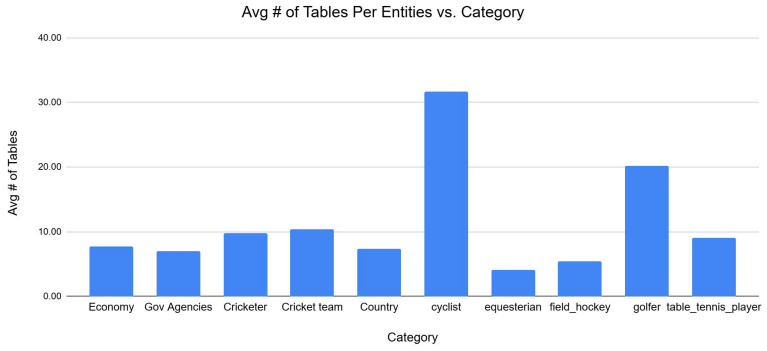
- Time information:
    - 2,985 implicit vs. 986 explicit questions

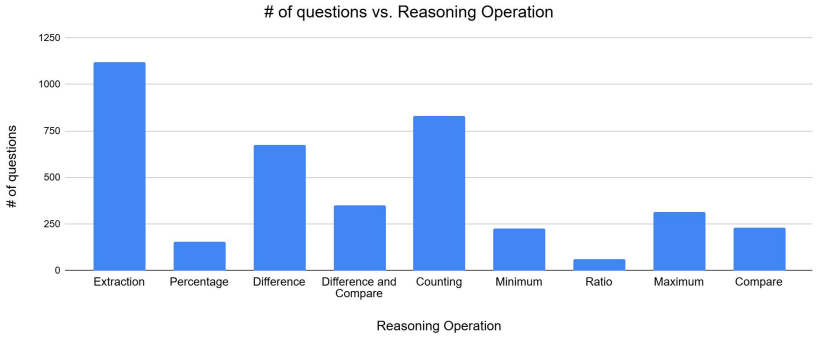
Example

- 'Who was the captain when India won the World Cup?' - **Implicit**
- 'Who was the captain in 2020?' - **Explicit**

Question Categorization

- Reasoning types:
  - extraction, counting, comparison, etc.
- Complexity:
  - 2,113 single key questions
  - 1,858 multiple key questions



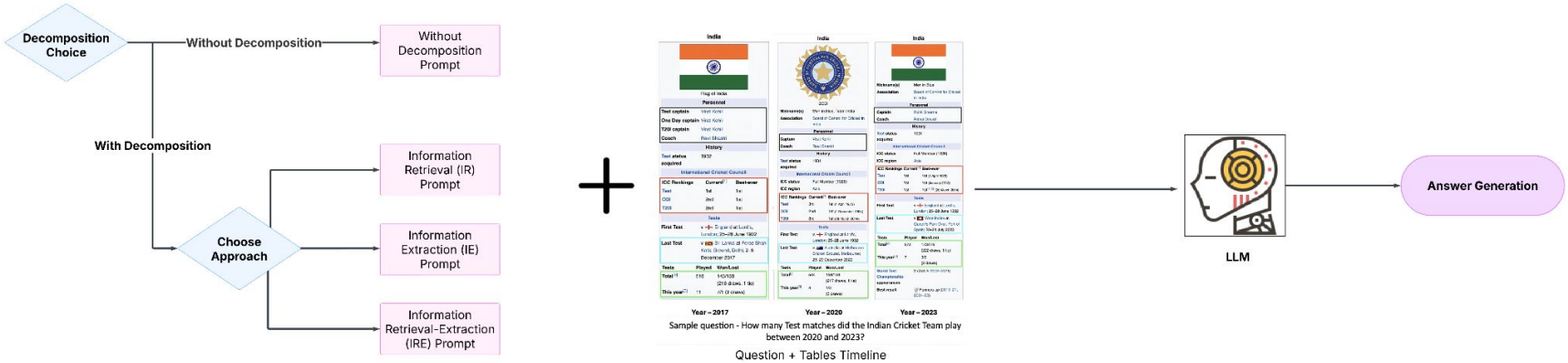


Modeling Techniques

- Information granularity variations:
    - Closed book
        - No context
    - Single table
        - Random or latest table only
    - Full timeline
        - All tables for the entity
    - Oracle timeline
        - Only the most relevant tables

Modeling Techniques

- Task decomposition approaches:
  - Without decomposition
  - Information retrieval
  - Information extraction
  - Information retrieval-extraction



Task Decomposition: Why It Matters

Question: "Who was the coach of the Indian Cricket Team when Virat Kohli was captain and the team had its highest ICC Test ranking?"

Approach 1: Without Decomposition

Task Decomposition: Why It Matters

Question: "Who was the coach of the Indian Cricket Team when Virat Kohli was captain and the team had its highest ICC Test ranking?"

Approach 1: Without Decomposition

- Process full timeline of tables simultaneously

Task Decomposition: Why It Matters

Question: "Who was the coach of the Indian Cricket Team when Virat Kohli was captain and the team had its highest ICC Test ranking?"

Approach 1: Without Decomposition

- Process full timeline of tables simultaneously
- Must identify Kohli's captaincy period, coaches during this time, and ICC rankings

Task Decomposition: Why It Matters

Question: "Who was the coach of the Indian Cricket Team when Virat Kohli was captain and the team had its highest ICC Test ranking?"

Approach 1: Without Decomposition

- Process full timeline of tables simultaneously
- Must identify Kohli's captaincy period, coaches during this time, and ICC rankings
- Perform temporal correlation across multiple attributes

Task Decomposition: Why It Matters

Question: "Who was the coach of the Indian Cricket Team when Virat Kohli was captain and the team had its highest ICC Test ranking?"

Approach 2: With Decomposition

Task Decomposition: Why It Matters

Question: "Who was the coach of the Indian Cricket Team when Virat Kohli was captain and the team had its highest ICC Test ranking?"

Approach 2: With Decomposition

✦

01

Table Retrieval

Identify tables from periods when Kohli was captain

Task Decomposition: Why It Matters

Question: "Who was the coach of the Indian Cricket Team when Virat Kohli was captain and the team had its highest ICC Test ranking?"

Approach 2: With Decomposition

01

Table Retrieval

Identify tables from periods when Kohli was captain

02

Key Extraction

Extract "*Captain*," "*Coach*," and "*ICC Test ranking*" keys from these tables

Task Decomposition: Why It Matters

Question: "Who was the coach of the Indian Cricket Team when Virat Kohli was captain and the team had its highest ICC Test ranking?"

Approach 2: With Decomposition

01

Table Retrieval

Identify tables from periods when Kohli was captain

02

Key Extraction

Extract "*Captain*," "*Coach*," and "*ICC Test ranking*" keys from these tables

03

Answer Generation

Identify when ranking was highest during Kohli's captaincy and determine coach during that specific period

Task Decomposition : Levels

Information Retrieval (IR)

Stage 1: "Table Retrieval" (identify relevant tables from timeline)

Stage 2: "Answer Generation" (reason over retrieved tables)

Information Extraction (IE)

Stage 1: "Key Extraction" (extract relevant attributes from tables)

Stage 2: "Answer Generation" (reason over extracted keys)

Information Retrieval Extraction (IRE)

Stage 1: "Table Retrieval" (identify relevant tables)

Stage 2: "Key Extraction" (extract relevant attributes from retrieved tables)

Stage 3: "Answer Generation" (reason over extracted keys)

Experimental Setup

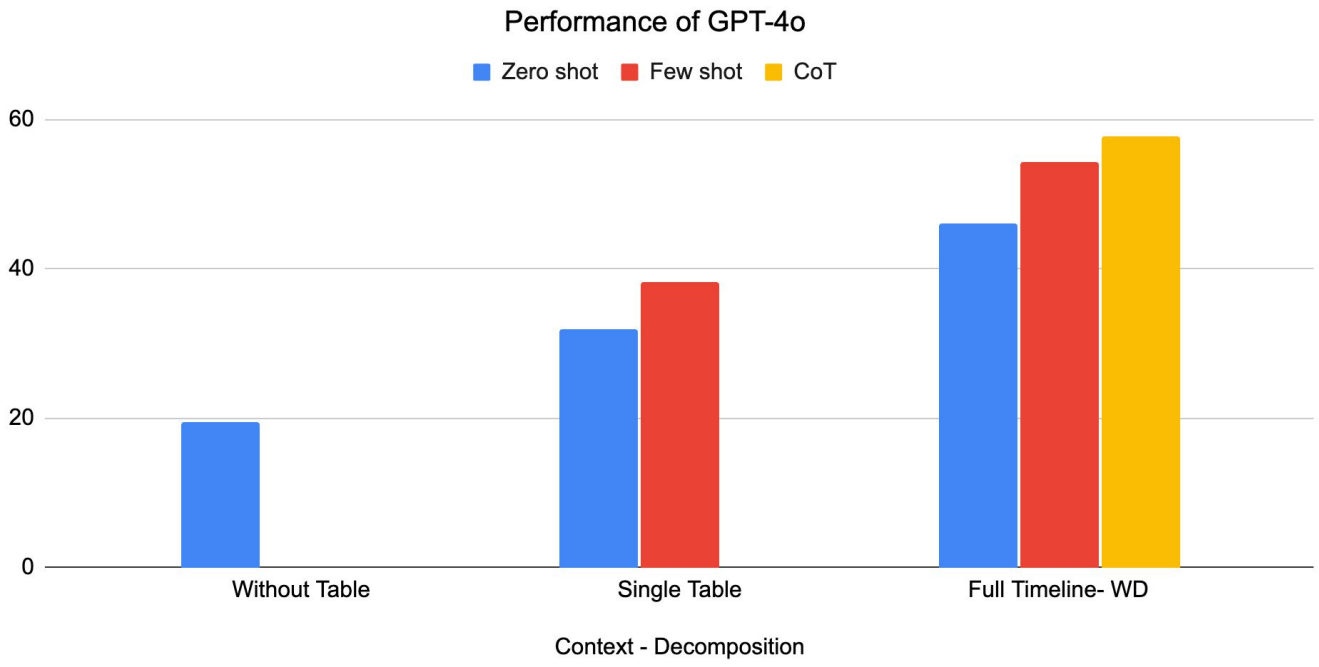
- Models evaluated
    - GPT-4o, Gemini-1.5-flash, GPT-4o-mini (proprietary)
    - Llama3-70B, Llama3-8B, Mixtral-7x8B (open source)

- Prompting Techniques
    - Zero shot, Few Shot, Chain of Thought

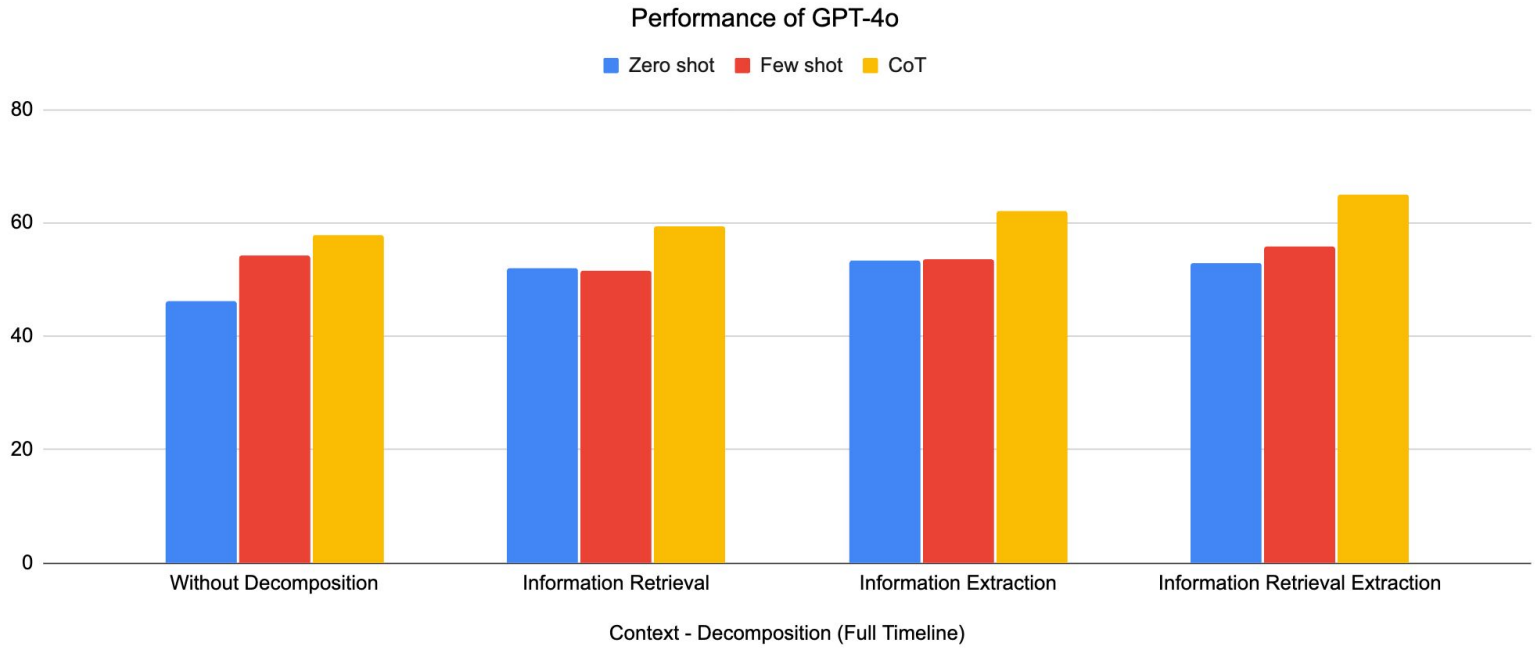
- Evaluation metrics
    - F1, Exact Match (EM), Rouge-1, Rouge-L

- Human evaluation baseline for comparison

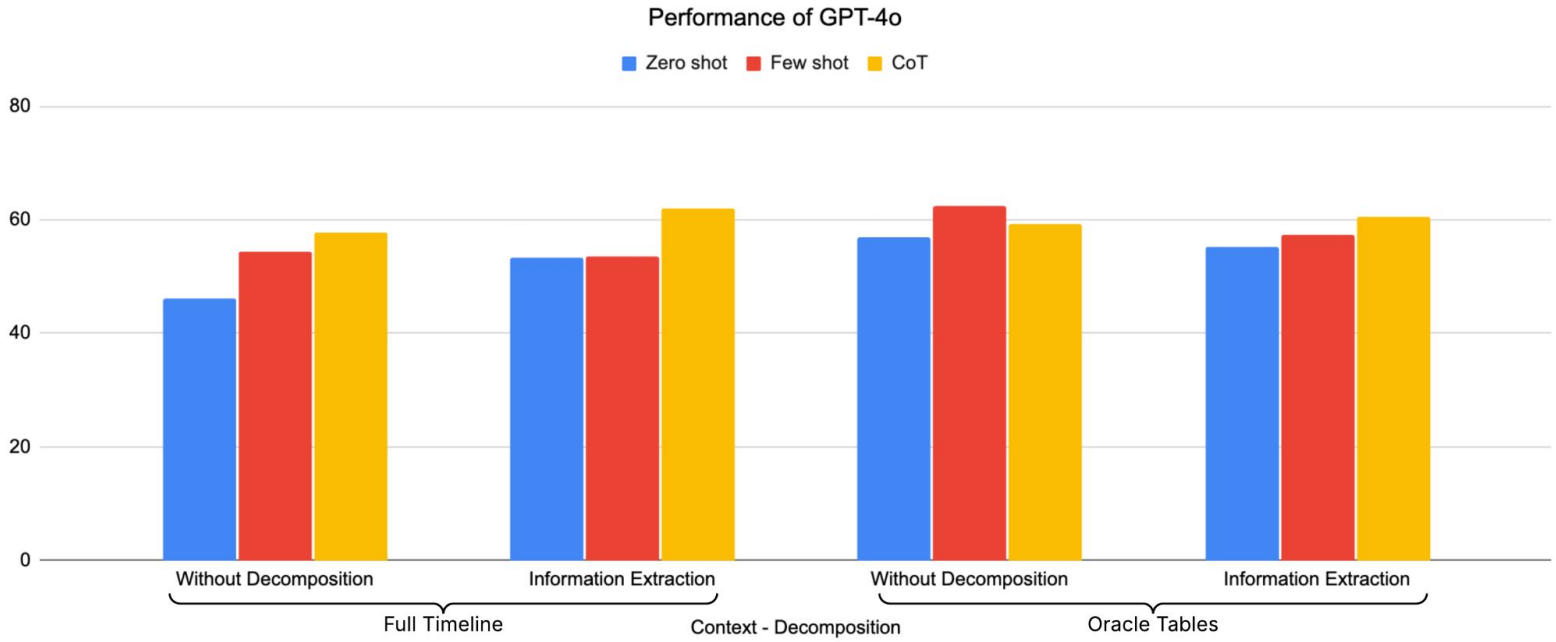
Results: Context Decomposition



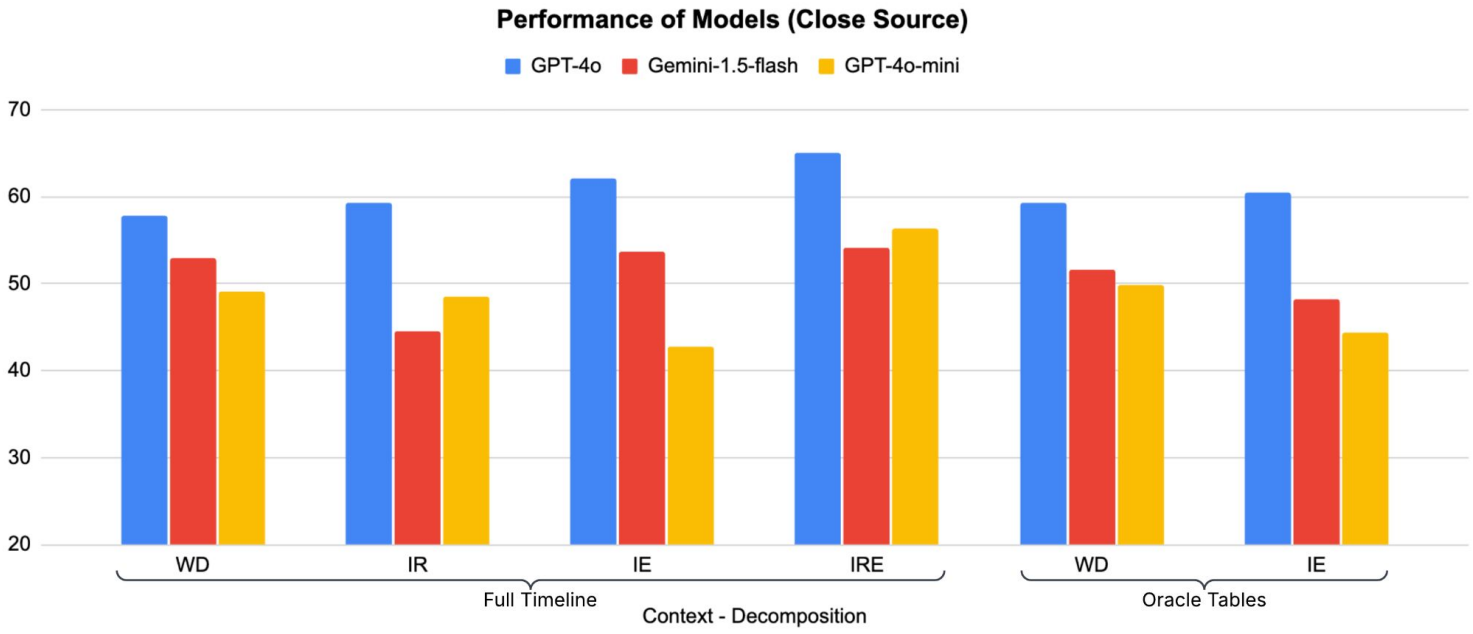
Results: Context Decomposition



Results: Context Decomposition

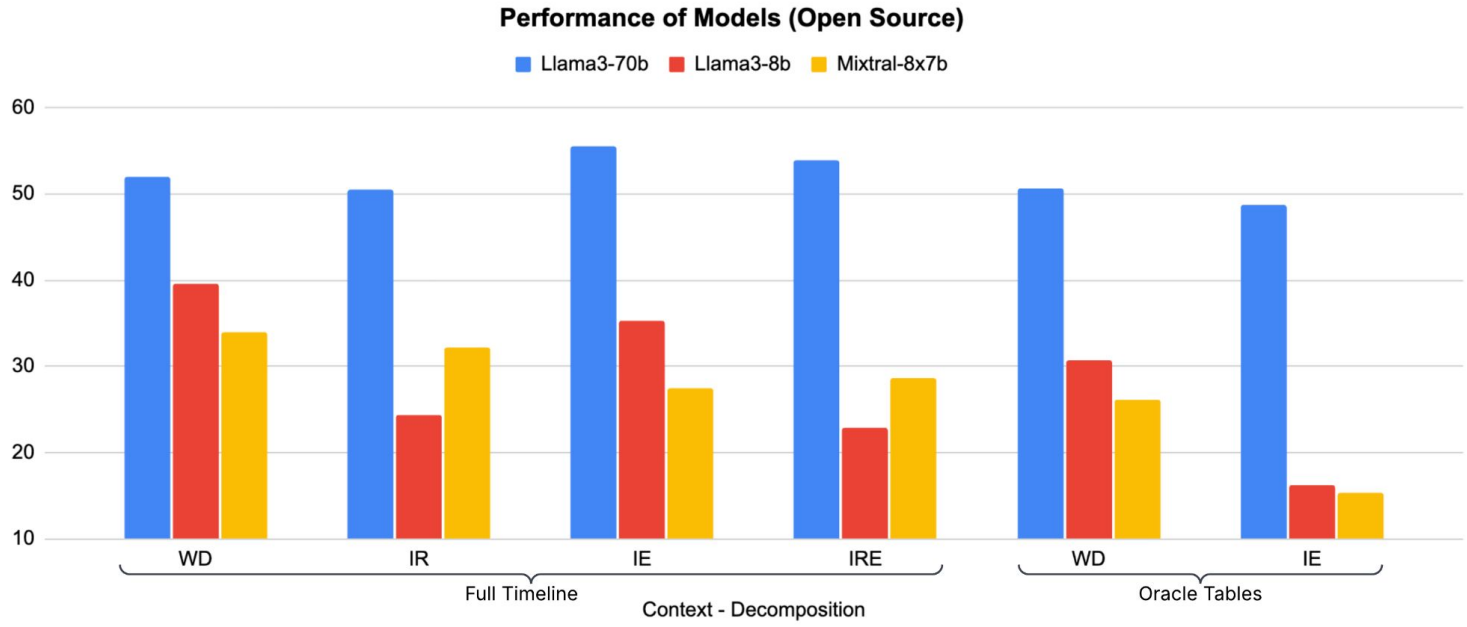


Results: Close Source Models



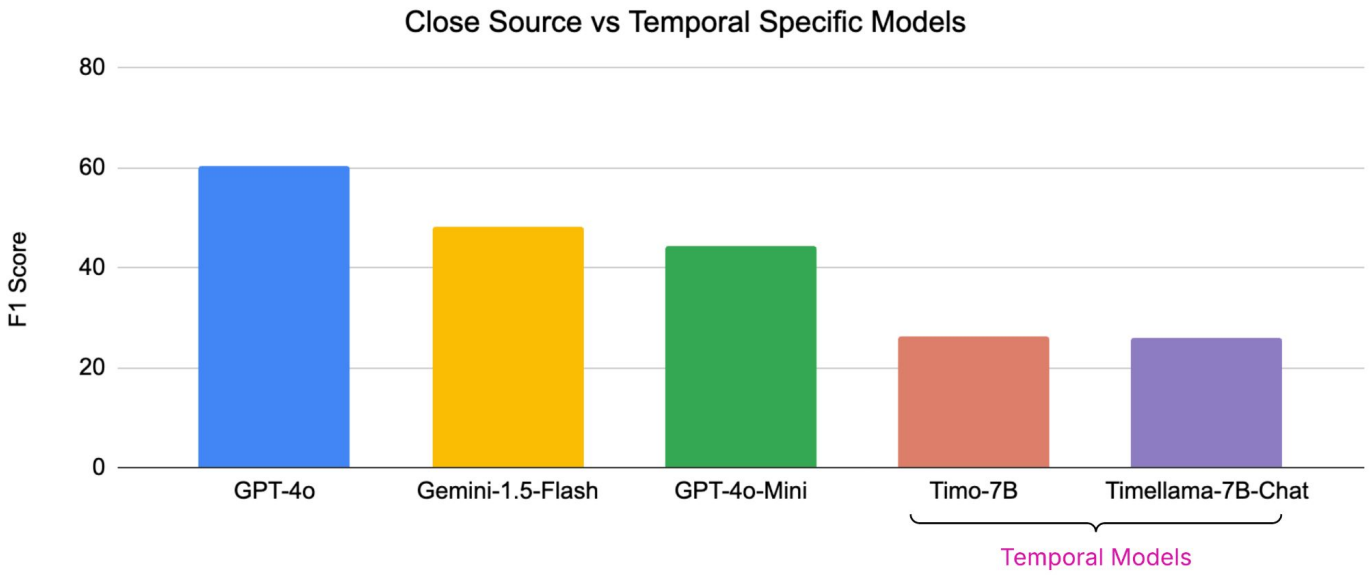
Results in different in-context variations and different intermediate task decompositions with various prompting methods.

Results: Open Source Models



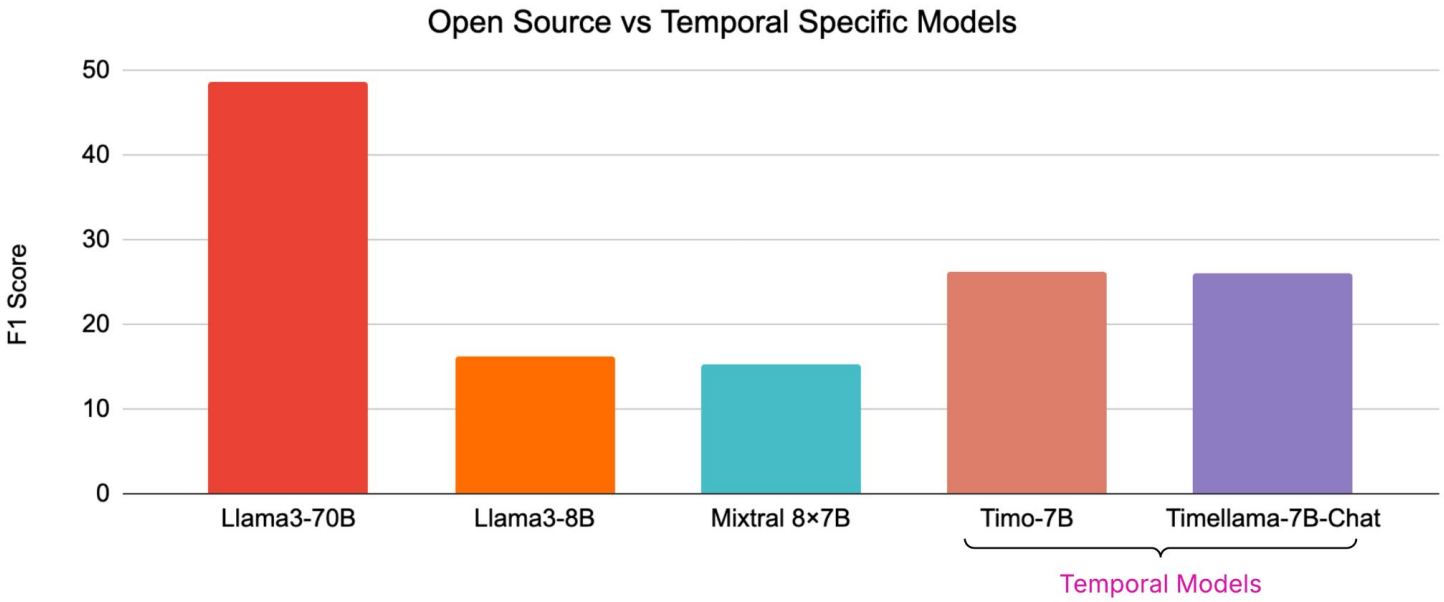
Results in different in-context variations and different intermediate task decompositions with various prompting methods.

Results : Temporal Models



CoT prompting with oracle tables and Key Extraction for task decomposition.

Results : Temporal Models



CoT prompting with oracle tables and Key Extraction for task decomposition.

Conclusion & Future Work

- A novel task of question answering on temporally evolving tables.
- A new **Transient Tables** dataset
    - 3,971 question-answer pairs.
    - From over 14k tables and 1,238 entities across various time periods.
- First study on LLM reasoning over entity-centric temporal tables.

In future:

- Currently its confined to Wikipedia infoboxes. Extending it to diverse structures beyond tables.
- Neuro-symbolic learning for better interpretability.

Transient Tables Paper

Thanks

Transient Tables Data

*seeking Summer Internship opportunities — let's connect!

ashankarampeta@ucsd.edu