NAACL 2025

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

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

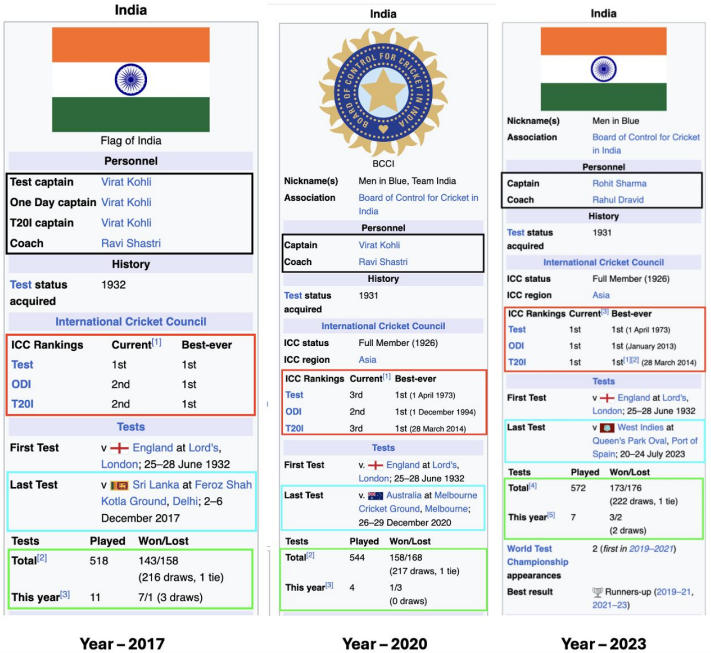
Affiliations: UC San Diego, University of Utah,

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?*



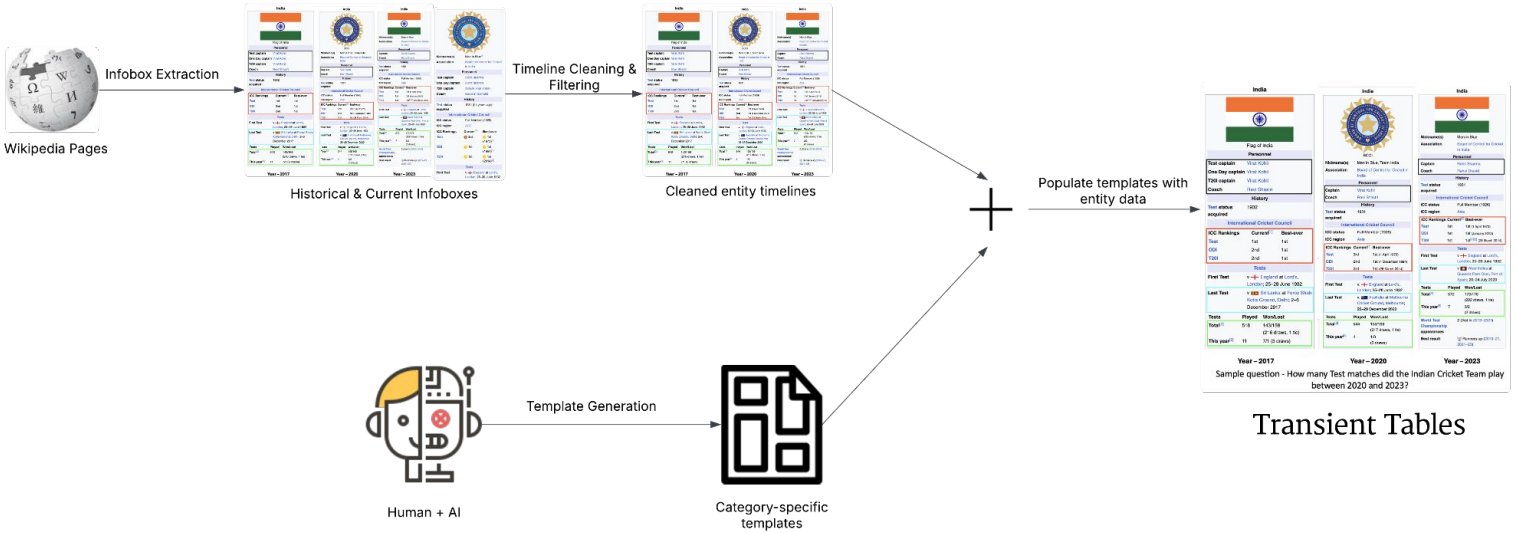
Sample question - How many Test matches did the Indian Cricket Team play between 2020 and 2023?

Key Contributions

- Transient Tables Dataset
    - A novel QA dataset with **3,971 questions** from **14,000+ tables** spanning **1,238 entities**
    - **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

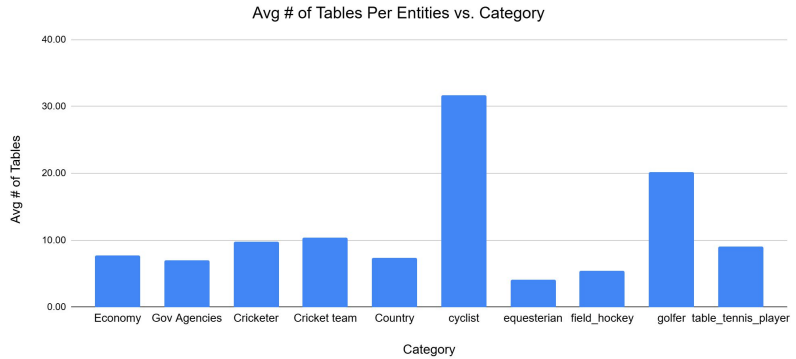
Dataset Creation

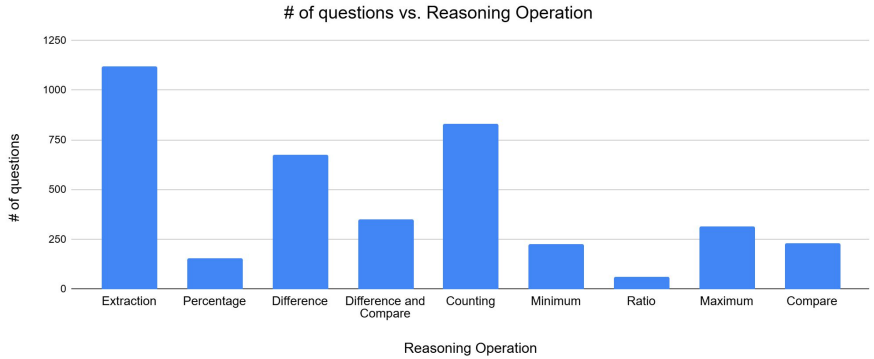
- Entity timeline selection from Wikipedia infoboxes
- Timeline cleaning and filtering (8-12 tables per entity)
- Query-answer generation through templates
- Statistics: 3,971 questions, 1,238 entities, average 11.42 tables per entity



Question Categorization

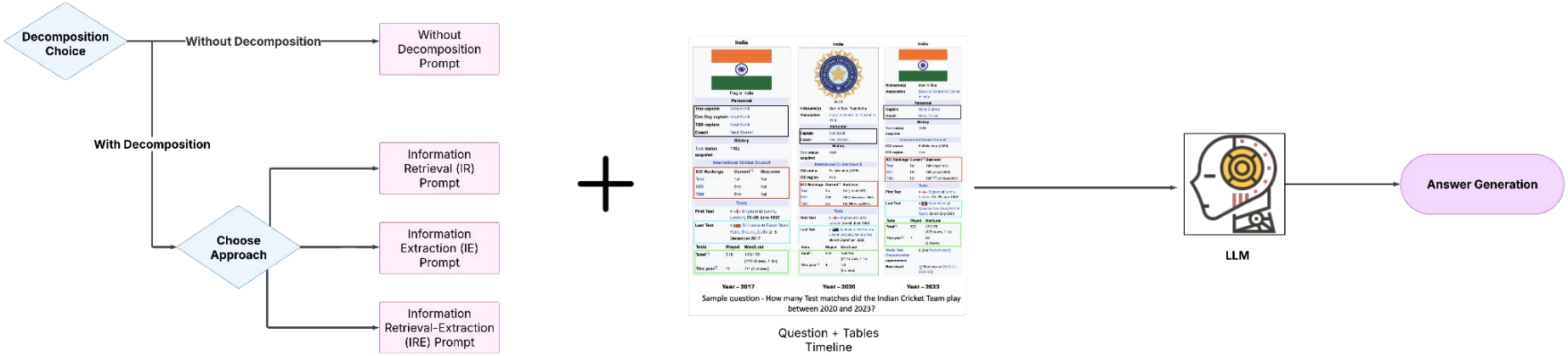
- Time information:
  - 2,985 implicit vs. 986 explicit questions
- Reasoning types
  - extraction, counting, comparison, etc.
- Complexity:
  - 2,113 single key vs. 1,858 multiple key questions





Modeling Techniques

- Information granularity variations: closed book, single table, full timeline, oracle timeline
- 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?"

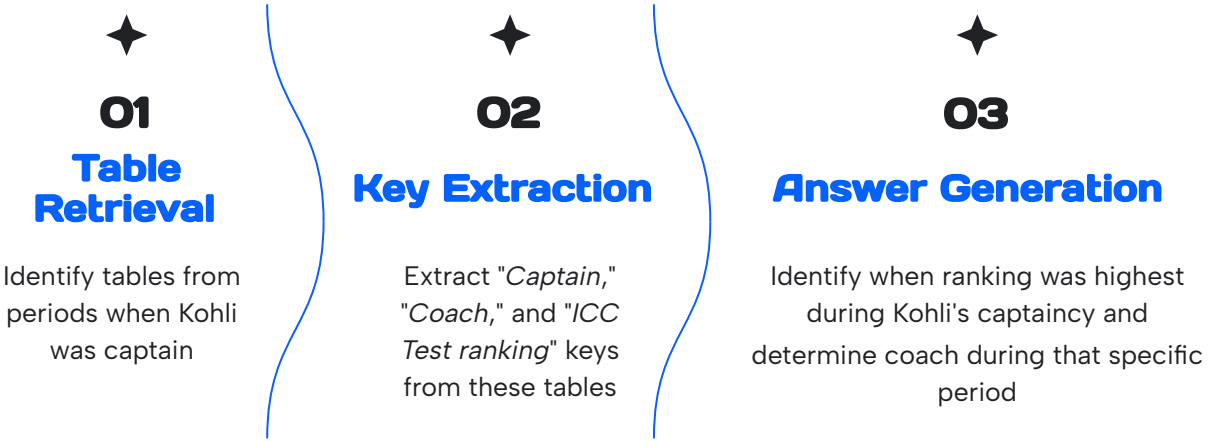
Approach 1: Without Decomposition

- 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 : 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)

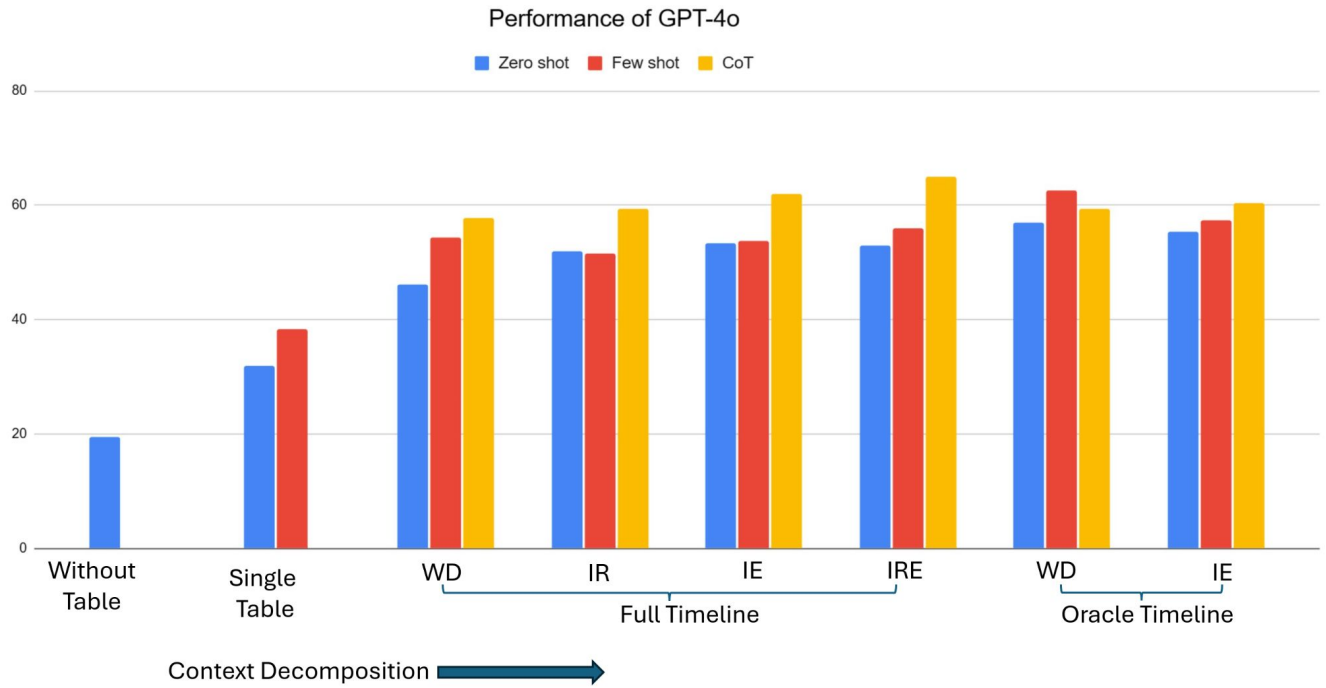
Stage 2: "Key Extraction" (extract relevant attributes from retrieved tables)

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

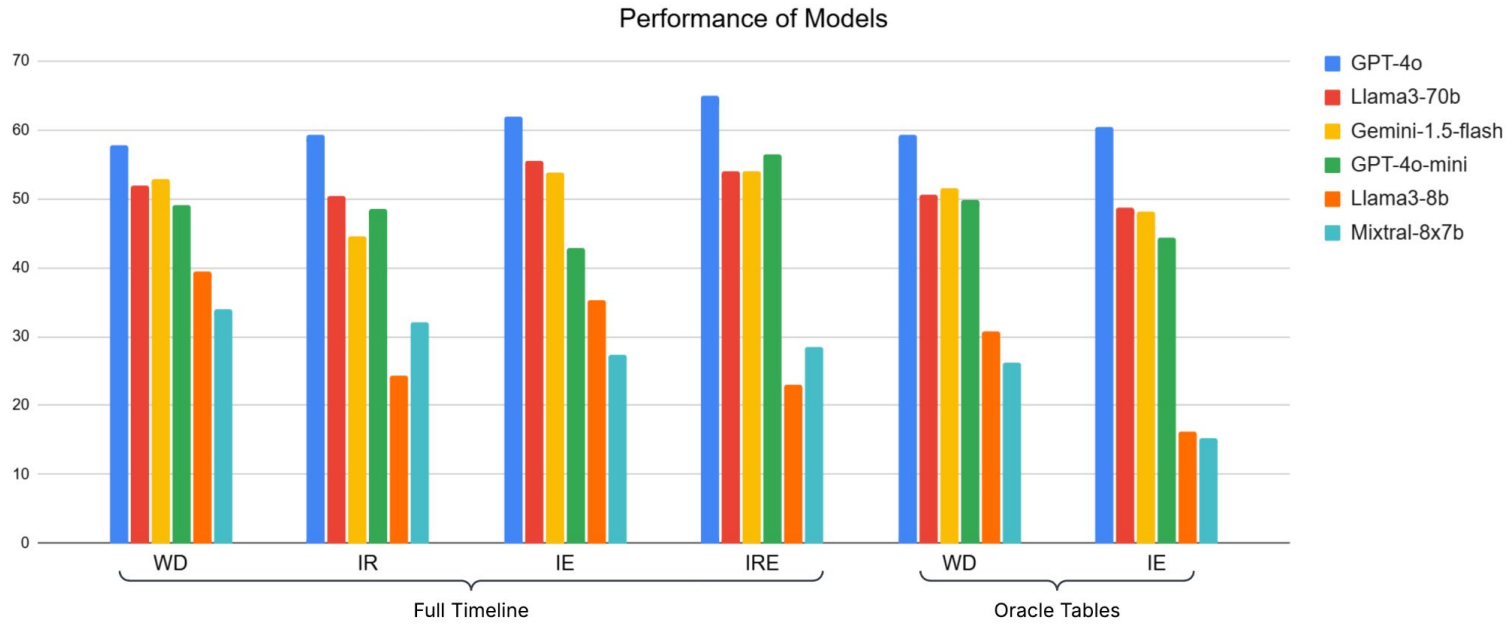
Experimental Setup

- Models evaluated
    - GPT-4o, Llama3-70B, Gemini-1.5-flash, GPT-4o-mini, Llama3-8B, Mixtral-7x8B
- 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

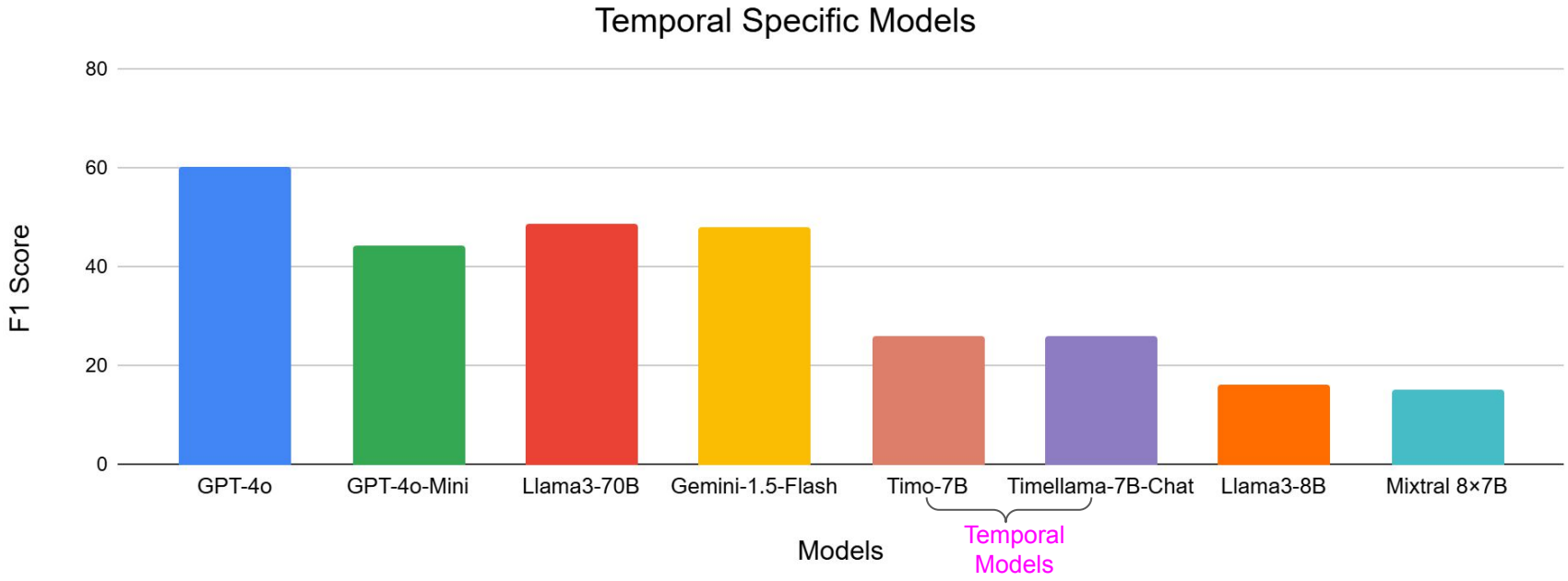


Results: Different Models



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.

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.

Scan the QR to checkout Transient Tables

Thanks