TRAVELER:
A Benchmark for Evaluating Temporal Reasoning across Vague, Implicit and Explicit References

Paper Authors:

Svenja Kenneweg[1] Jörg Deigmöller[2] Philipp Cimiano[1]
Julian Eggert[2]

[1]Bielefeld University, Bielefeld, Germany
[2]Honda Research Institute Europe, Offenbach, Germany

Presenter:

Binheng Zheng

binheng.zheng@stud.uni-heidelberg.de

Heidelberg, 25.11.2025

Outline

Why Temporal Reasoning Matters

- Human communication constantly refers to **past and future events**.
- Understanding temporal references is essential for NLU:
  - *"on 13 January 2023"* (explicit)
  - *"the last time Peter cooked"* (implicit)
  - *"Who just prepared risotto?"* (vague)
- These expressions require **different levels of reasoning difficulty**.

Challenges in Temporal Reference Resolution

- Temporal expressions vary in explicitness:
  - **Explicit**: direct timestamps, easy to resolve
  - **Implicit**: depends on speech time or context
  - **Vague**: fuzzy boundaries, inherently ambiguous
- Resolving these requires **Event Temporal Reasoning**:
  - comparing timestamps
  - tracking event sequences
  - identifying relevant events among many

Existing Benchmarks are Limited

- Existing temporal QA datasets:
  - do **not** systematically test explicit vs. implicit vs. vague
  - do **not** vary event set length (only few events)
  - do **not** include **vague temporal adverbials**
- **Gap:** No benchmark evaluates LLMs on **multi-event, multi-step temporal reasoning**.

Our Hypotheses

- **H1:** Performance decreases as temporal explicitness decreases.
- **H2:** Vague temporal references lead to the **lowest accuracy**.
- **H3:** Larger event sets significantly reduce model performance.

Categories of Temporal Questions

- Temporal questions differ in how time is referenced:
  - **Explicit**: clear timestamps (e.g., 2023-09-20)
  - **Implicit**: depend on context or speech time (e.g., *yesterday*, *last time*)
  - **Vague**: fuzzy boundaries (e.g., *just*, *recently*)
- Prior work distinguishes multiple implicit subtypes:
  - commonsense (e.g., Christmas 2023)
  - relative to speech time (e.g., two days ago)
  - relative to other anchors (e.g., two days before Christmas 2022)
  - personal knowledge (e.g., Tom's birthday)
- **TRAVELER**[1] **focuses on:** explicit, implicit (speech-time), and vague.

[1] The benchmark is publicly available at: `https://gitlab.ub.uni-bielefeld.de/s.kenneweg/TRAVELER`.

Vague Temporal Expressions

- Vagueness arises when boundaries are unclear: *tall*, *many*, *some time ago*.
- Vague temporal adverbials are underexplored in NLP: *just*, *recently*, *long time ago*.
- Prior work (Kenneweg et al. 2024):
  - human surveys map vague terms to probability curves
  - e.g., probability that "recently" applies at different time distances
- **TRAVELER adopts this method to define probabilistic ground truth.**

Existing Temporal QA Benchmarks

- Prior benchmarks:
  - TimeQA, TempQuestions, TimeQuestions
  - BIG-bench: Date Understanding, Temporal Sequences
- Limitations:
  - lack **systematic control** over explicit vs. implicit vs. vague
  - do not vary **event set length**
  - no coverage of **vague temporal adverbials**
  - limited support for **multi-event** reasoning
- **Gap:** No benchmark tests temporal reasoning over **extended event sets** with **varying explicitness**.

LLMs for Reasoning

- LLMs show strong performance on many reasoning tasks.
- Chain-of-Thought improves symbolic and mathematical reasoning.
- However, LLMs struggle with:
  - multi-step reasoning
  - logical consistency
  - commonsense and planning tasks
  - temporal reasoning (e.g., in TimeBench)
- **Motivation:** need targeted benchmarks to test event-based temporal reasoning.

Benchmark Design Overview

- Goal: evaluate LLMs' **event-temporal reasoning** ability.
- Two key dimensions:
  - **Temporal explicitness**: explicit vs. implicit (speech-time) vs. vague
  - **Event set length**: from 5 to 100 events
- Pipeline:
  - Generate synthetic event sets
  - Generate temporal questions from templates
  - Define ground truth (crisp vs. probabilistic)
  - Design prompting strategies for LLM evaluation

Generation of Synthetic Event Sets

- Events are sampled in a **synthetic home environment**.
- Each event is a tuple:

$$\langle \text{Event Type}, \text{Subject}, \text{Location}, \text{Timestamp} \rangle$$

- Timestamps:
  - Unix timestamps[2] between 2023-01-01 and 2023-09-29
- Event set lengths:

$$n \in \{5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$$

[2] A Unix timestamp t is defined as the number of seconds elapsed since 1970-01-01 00:00:00 UTC, i.e. $t = \text{seconds(time)} - \text{seconds(1970-01-01 00:00:00 UTC)}$. For example, 2023-01-01 00:00:00 UTC corresponds to $t = 1672531200$.

Synthetic Event Types, Subjects, Locations (Table 1)

Event Type	Subject	Location
watch film	Mary, Tom, Robot	living room, kitchen
eat risotto	Mary, Tom, Robot	living room, kitchen
read book	Mary, Tom, Robot	living room, kitchen
dance lively salsa	Mary, Tom, Robot	living room, kitchen
store wine bottle	Mary, Tom, Robot	living room, kitchen
drink juice	Mary, Tom, Robot	living room, kitchen
chat with friend	Mary, Tom, Robot	living room, kitchen

Question Generation

- For each event set, questions are generated from:
  - **Question templates** (Table 2)
  - **Temporal expressions** (Table 3)
- Temporal categories:
  - Explicit
  - Implicit relative to speech time
  - Vague
- For each combination of:
  - event-set length (11 values) and
  - question category (3 types)

  generate **100 questions** $\Rightarrow 100 \times 11 \times 3 = 3300$ questions.

Question Templates (Table 2)

Template	Return Type
Who ⟨event type⟩⟨location⟩⟨temporal expr.⟩?	String
Did ⟨subject⟩⟨event type⟩⟨location⟩⟨temporal expr.⟩?	Bool
How often did ⟨subject⟩⟨event type⟩⟨location⟩⟨temporal expr.⟩?	Integer
When was the last time ⟨subject⟩⟨event type⟩⟨location⟩?	Date

The last template is only used for the **implicit relative to speech time** category.

Temporal Expressions by Question Category (Table 3)

Question Category	Temporal Expressions
Explicit	on yyyy-mm-dd in yyyy-mm in the year yyyy
Implicit relative to speech time	today, yesterday this year, this month last month
Vague	just, recently some time ago, long time ago

Example: Question Generation from One Event

Given an event:

$$\langle \text{wash mug}, \text{Tom}, \text{kitchen}, \text{2023-09-29 20:27} \rangle$$

we can generate:

- Explicit:
  - Who washed a mug in the kitchen on 2023-09-29?
- Implicit (speech-time):
  - Did Tom wash a mug in the kitchen yesterday?
  - How often did Tom wash a mug in the kitchen this month?
  - When was the last time Tom washed a mug in the kitchen?
- Vague:
  - Did Tom just wash a mug in the kitchen?
  - Did Tom wash a mug in the kitchen some time ago?

Ground Truth: Explicit & Implicit (Speech-Time)

- For **explicit** and **implicit relative to speech time**:
  - ground truth is computed automatically from the event set.
- Example: *Did Tom eat risotto in the living room on 2023-09-29?*
  - Answer = "yes" iff there exists an event:

$$\langle S = \text{Tom}, T = \text{eat risotto}, L = \text{living room}, D = \text{2023-09-29} \rangle$$

- Model responses (for *Did*, *Who*, *How often*, *Last time* [3]) are evaluated using **binary accuracy**:
  - correct answer $\Rightarrow 1$, incorrect $\Rightarrow 0$.

[3] For "Who …?", the gold answer is the set $S_{\text{gold}} = \{S_e \mid e \in G, \text{ conditions hold}\}$; for "How often …?", it is the count $|G'(S, T, L)|$ of matching events; for "When was the last time …?", it is the latest date $\max_{e \in G'(S,T,L)} D_e$.

Ground Truth: Vague Temporal References (Idea)

- Vague adverbials (*just*, *recently*, *some time ago*, *long time ago*) have **no crisp boundary**.
- Use **human surveys** (Prolific) to estimate applicability:
  - For each event type and adverbial:
  - Ask native speakers how appropriate the adverbial is
  - for events that happened Δt time units ago.
- Responses $\Rightarrow$ probability curves

$$p_{e,A}(\Delta t) \in [0, 1]$$

indicating how likely adverbial A applies to an event e at time distance Δt.

Example Probability Curve (Fig. 1)

- For each event type (e.g., *eating risotto*) and adverbial (e.g., *long time ago*), median survey responses define a probability curve $p_{e,A}(\Delta t)$.

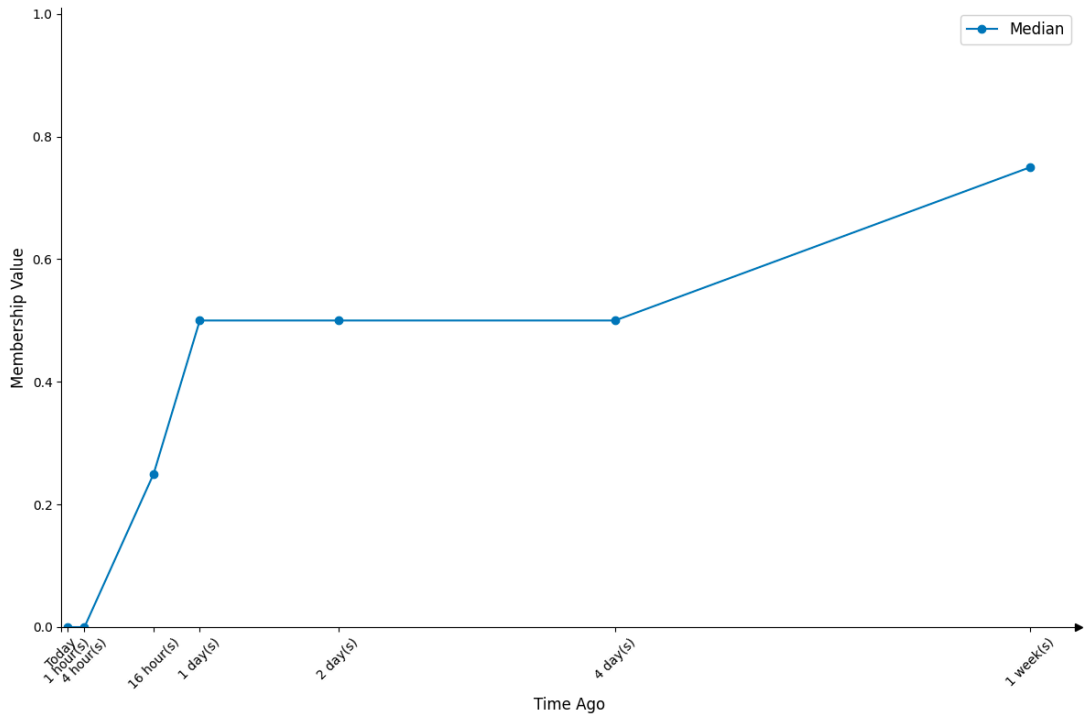


Fig. 1: Median responses for event type *eating risotto* and adverbial *long time ago*.

Formal Ground Truth for Vague References

- Let G be the set of all events in the prompt. Each event e has attributes:

$$\{S_e, T_e, L_e, D_e\}$$

- For a subject S, event type T, and location L, define:

$$G'(S, T, L) = \{e' \in G \mid S_{e'} = S,\ T_{e'} = T,\ L_{e'} = L\}.$$

- For each $e' \in G'(S, T, L)$, compute time difference

$$\Delta t = D_{\text{ref}} - D_{e'}.$$

- From the survey, obtain $p_{e',A}(\Delta t)$: probability that adverbial A applies to e' at Δt.

Probabilistic Accuracy for Vague "Did ...?"

- Probability that *none* of the events in G' fits A:

$$\bar{P}_{G',A} = \prod_{e' \in G'} \left(1 - p_{e',A}\right).$$

- Probability that *at least one* event in G' fits A:

$$P_{G',A} = 1 - \bar{P}_{G',A}.$$

- For a question of the form:

"Did $\{S\}$ do $\{T\}$ in the $\{L\}$ $\{A\}$?"

and a model response:

$$P_{\text{response}}(\text{yes} \mid S, T, L, A) = P_{G',A},$$
$$P_{\text{response}}(\text{no} \mid S, T, L, A) = 1 - P_{G',A}.$$

- This probability is used as a **soft accuracy** score for "Did ...?"[4].

[4] Other vague question types are treated analogously: for "Who ... A?" we compute $P_{G'_s,A} = 1 - \prod_{e' \in G'_s}(1 - p_{e',A})$ per subject s; "How often ... A?" uses a soft count $\sum_{e' \in G'(S,T,L)} p_{e',A}$ as the expected number of A-events.

Example: "Did Tom eat risotto in the kitchen a long time ago?"

- Event set G contains, among others:

$$e_1 : \langle \text{Tom}, \text{eat risotto}, \text{kitchen}, \text{2023-09-29 20:27} \rangle,$$
$$e_2 : \langle \text{Tom}, \text{eat risotto}, \text{kitchen}, \text{2023-09-22 22:27} \rangle.$$

- Then

$$G' = \{e_1, e_2\}$$

  for subject Tom, event type *eat risotto*, location *kitchen*.
- Reference time: $D_{\text{ref}} = \text{2023-09-29 22:27}$.
- Time differences:

$$\Delta t_1 \approx 2 \text{ hours}, \quad \Delta t_2 \approx 1 \text{ week}.$$

- From survey (for "long time ago"):

$$p_{e_1,A}(\text{2h}) = 0, \quad p_{e_2,A}(\text{1week}) = 0.75.$$

Example: Soft Accuracy for the Response

- Probability at least one event fits "long time ago":

$$P_{G',A} = 1 - (1 - 0) \cdot (1 - 0.75) = 0.75.$$

- If the model answers **"yes"** to:

  Did Tom eat risotto in the kitchen a long time ago?

  then its **accuracy score** for this question is:

$$P_{\text{response}}(\text{yes} \mid S, T, L, A) = 0.75.$$

- If it answers **"no"**, the score would be:

$$P_{\text{response}}(\text{no} \mid S, T, L, A) = 1 - 0.75 = 0.25.$$

Prompting Strategies

- Three dimensions of prompt design:
  - **Prompt type**:
    - Zero-shot
    - Chain-of-Thought (**CoT** Review)
    - CoT Step-by-Step
  - **Date information**:
    - Date-Only (date + time)
    - Date-Extended (plus weekday, calendar week)
  - **Event presentation**:
    - Json format
    - Natural Language
- In total: $3 \times 2 \times 2 = \mathbf{12}$ different prompt configurations.
  - All 12 will be evaluated in the experiments.

Prompting Strategies: Intuition

- **Zero-shot**:
  - model directly answers based on the event list and the question.
- **CoT Review**:
  - explicitly asks the model to **inspect each event** and record matches.
- **CoT Step-by-Step**:
  - adds "Let's think step by step." to further encourage explicit reasoning.
- **Date-Extended** vs. **Date-Only**:
  - tests if extra details (weekday, calendar week) help or just add noise.
- **Json** vs. **Language**:
  - tests whether LLMs benefit more from structured data or natural language descriptions.

Example Prompts: Zero-shot vs. CoT Review

Event set (Language, Date-Only):

- On September 29, 2023 at 08:01, Mary watched a film in the living room.
- On September 28, 2023 at 14:27, Tom ate a risotto in the kitchen.
- On June 11, 2023 at 12:44, Ria read a book in the living room.
- On August 11, 2023 at 10:57, Mary danced a lively salsa in the kitchen.
- On September 1, 2023 at 20:44, Tom stored a wine bottle in the living room.

Zero-shot prompt (simplified):

Today is 2023-09-29 22:18. I will give you a list of events (event set) that have taken place in the past: […]. Who watched a film in the living room on 2023-09-29? Answer with the name of the subject or say "nobody".

CoT Review prompt (simplified):

Review each event in the event set sequentially. If the action, object, location and date match the information in the question, record the subject of that event. At the end, return the subjects of all matched events. Today is 2023-09-29 22:18. [same event set …] Who watched a film in the living room on 2023-09-29?

Experimental Setup

- Two-stage evaluation:
  - **Stage 1:** Prompting strategies (GPT-4 only)
  - **Stage 2:** Model comparison (4 LLMs)
- Models:
  - Gemma-7B-it
  - Llama3-8B-Instruct
  - Llama3-70B-Instruct
  - GPT-4-0125
- Evaluation dimensions:
  - temporal explicitness: explicit, implicit (speech-time), vague
  - event set length: from 5 up to 100 events
  - question templates: Who / Did / How often / When last time

Stage 1: Prompting Strategy Evaluation

- Goal: compare **12** prompt configurations on GPT-4:
  - Prompt type: Zero-shot / CoT Review / CoT Step-by-Step
  - Date info: Date-Only / Date-Extended
  - Event format: Json / Language
- Setup:
  - Question categories: **Explicit** and **Implicit (speech-time)**
  - Event set lengths: **5** and **50**
- Metrics:
  - accuracy for explicit & implicit questions
  - averaged over lengths and temporal categories

Results: Prompting Strategies (GPT-4, Table 4)

- Table 4 (paper): accuracy of GPT-4 for all 12 configurations.
- Key observations:
  - **CoT prompts** slightly outperform Zero-shot.
  - **Date-Only** performs as well as or better than Date-Extended.
  - **Language** descriptions perform at least as well as Json.
  - Accuracy drops when event set length increases from 5 to 50.

Prompting Strategy	Date Information	Event Presentation	Events 5	Events 50	Average
Zero-Shot	Date-Only	Json	.97	.67	.82
Zero-Shot	Date-Only	Language	.96	.67	.82
Zero-Shot	Date-Extended	Json	.97	.64	.81
Zero-Shot	Date-Extended	Language	.96	.68	.82
CoT Review	**Date-Only**	**Json**	.97	.71	**.84**
CoT Review	**Date-Only**	**Language**	.94	.71	**.83**
CoT Review	Date-Extended	Json	.95	.68	.82
CoT Review	Date-Extended	Language	.93	.71	.82
CoT Step-by-Step	**Date-Only**	**Json**	.94	.71	**.83**
CoT Step-by-Step	**Date-Only**	**Language**	.95	.71	**.83**
CoT Step-by-Step	Date-Extended	Json	.94	.66	.80
CoT Step-by-Step	Date-Extended	Language	.94	.70	.82

Stage 2: Model Comparison

- Use the four best-performing prompts from Stage 1:
  - CoT Review / CoT Step-by-Step
  - Date-Only
  - Json / Language
- Evaluate all four LLMs:
  - Gemma-7B-it, Llama3-8B, Llama3-70B, GPT-4-0125
- Same setting:
  - Question categories: Explicit & Implicit (speech-time)
  - Event set lengths: 5 and 50
- Goal:
  - Assess how model size and architecture affect event-temporal reasoning.

Results: Model Comparison (Table 5)

- Table 5 (paper): average accuracy of 4 LLMs across 4 best prompts.
- Findings:
  - **Llama3-70B** achieves the highest accuracy overall.
  - **GPT-4** performs slightly worse but still close to 70B.
  - **Llama3-8B** and **Gemma-7B** perform significantly worse.
  - Larger models are more robust to longer event sets.

Prompting Strategy	Date Information	Event Pres.	`Gemma -7b-it`	`Llama3 -8B-Instr.`	`Llama3 -70B-Instr.`	`GPT-4 -0125`	Average
CoT Rev.	Date-Only	Json	**.68**	.68	.86	**.84**	.76
CoT Rev.	**Date-Only**	**Lang.**	**.68**	**.74**	.88	.83	**.78**
CoT Step.	Date-Only	Json	.63	.69	.84	.83	.75
CoT Step.	Date-Only	Lang.	.65	.72	**.90**	.83	.77

Final Prompt Configuration

- Based on Stage 1 and Stage 2:
  - CoT prompts > Zero-shot
  - Date-Only ≥ Date-Extended
  - Language ≥ Json
- **Final choice for all detailed analyses:**
  - **CoT Review + Date-Only + Language** encoding
- This configuration is:
  - competitive for GPT-4
  - robust across all four models

Effects of Question Category & Template (Table 6)

- Now fix the final prompt configuration and evaluate:
  - 3 temporal categories:
    - Explicit
    - Implicit (relative to speech time)
    - Vague
  - 4 question templates:
    - Who ...?
    - Did ...?
    - How often did ...?
    - When was the last time ...?
- Table 6 (paper) reports accuracies for all combinations and models.

		Gemma -7b-it	Llama3-8B -Instruct	Llama3-70B -Instruct	GPT-4 -0125	Average
Question Category	Explicit	**.84**	**.75**	**.92**	**.84**	.84
	Implicit relative...	.34	.58	.74	.64	.58
	Vague	.26	.45	.41	.42	.39
Question Templates	Who ...?	.44	.46	.65	.47	.51
	Did ...?	**.68**	**.77**	**.82**	**.86**	.78
	How often did ...?	.35	.52	.65	.57	.52
	When was the ...?	.34	.53	.66	.75	.57

Main Findings: Temporal Categories

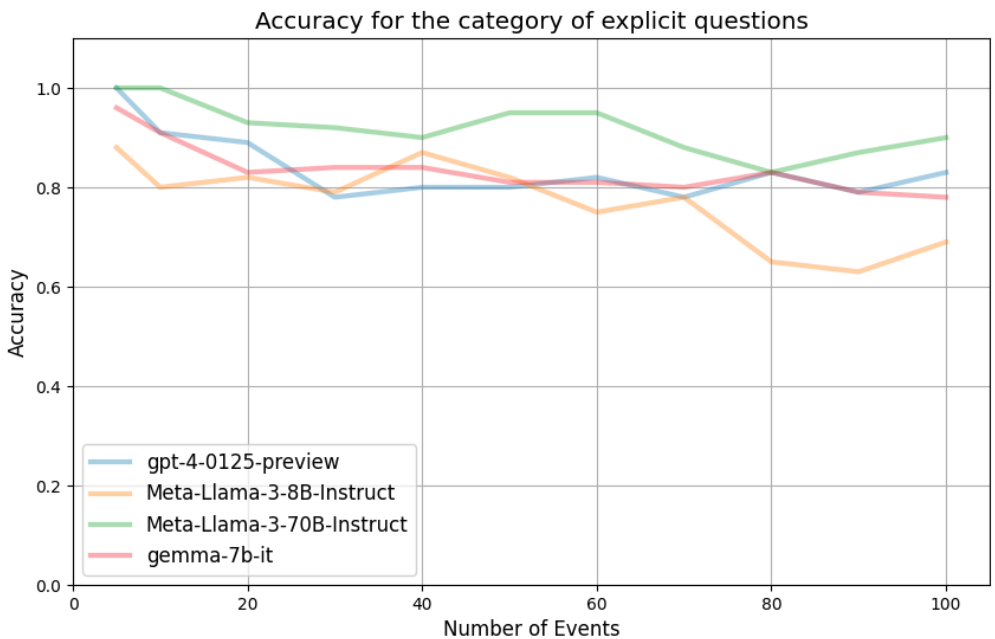
- Across all models:
  - **Explicit** questions: highest accuracy.
  - **Implicit (speech-time)**: significantly lower.
  - **Vague** questions: lowest accuracy.
- This supports:
  - **H1:** performance decreases as temporal explicitness decreases.
  - **H2:** vague temporal references are the hardest.
- Interpretation:
  - explicit timestamps are easy to match,
  - implicit references require reasoning over speech time and context,
  - vague references require fuzzy, probabilistic reasoning.

Main Findings: Question Templates

- Template-wise:
  - **Did ...?** (yes/no) performs best:
    - only requires existence checking.
  - **Who ...?** performs worst:
    - requires retrieving and correctly outputting subject names.
  - **How often ...?** and **When was the last time ...?** are in between:
    - need counting or identifying the most recent event.
- Conclusion:
  - more complex reasoning operations (counting, ordering, open answers) are more error-prone than simple existence checks.

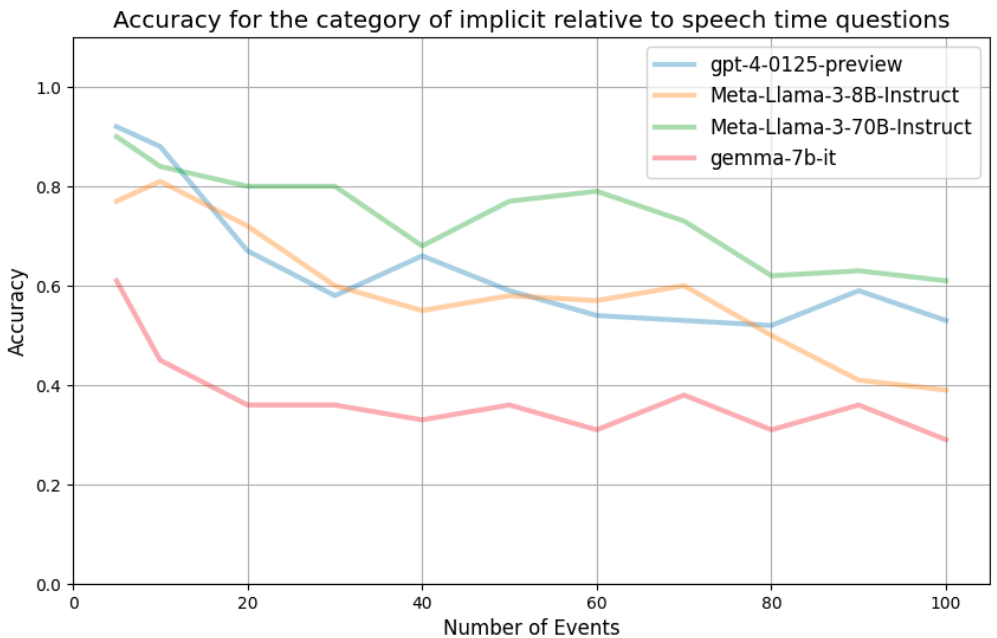
Effect of Event Set Length: Explicit Questions (Fig. 4)

- Fig. 4 (paper): accuracy vs. event set length for **explicit** questions.
- Trends:
  - all models degrade as event set length increases from 5 to 100.
  - **Llama3-70B** is most robust, with the smallest drop.
  - smaller models (Gemma, Llama3-8B) degrade more strongly.



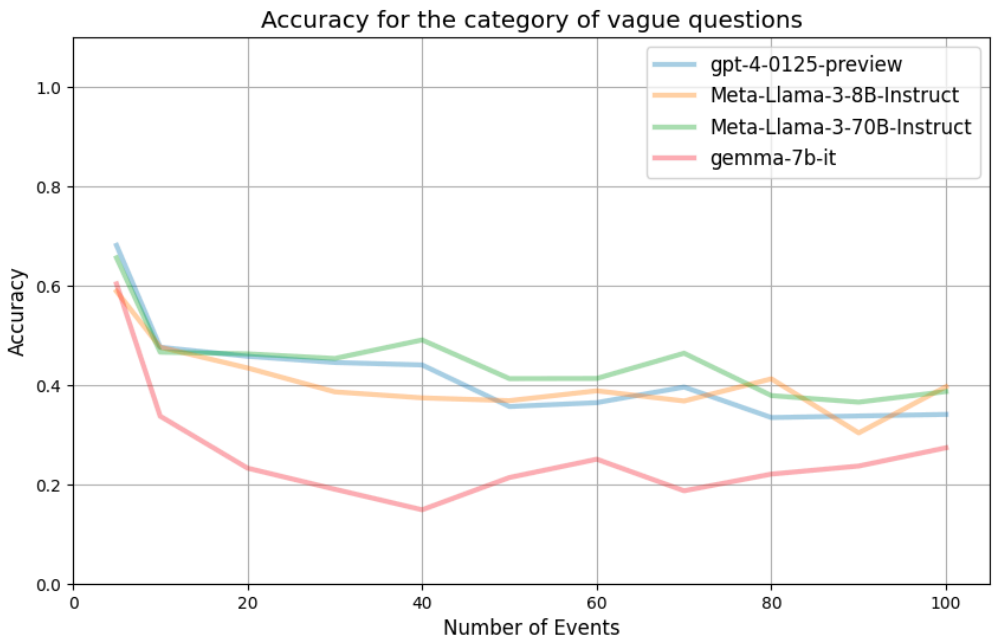
Effect of Event Set Length: Implicit Questions (Fig. 5)

- Fig. 5 (paper): accuracy vs. event set length for **implicit (speech-time)** questions.
- Trends:
  - much stronger degradation compared to explicit questions.
  - some models lose almost 40 percentage points from 5 to 100 events.
  - again, Llama3-70B is most robust, but still clearly affected.
- Interpretation:
  - combining speech-time reasoning with long event sets is very challenging.



Effect of Event Set Length: Vague Questions (Fig. 6)

- Fig. 6 (paper): accuracy vs. event set length for **vague** questions.
- Trends:
  - all models start from a relatively low accuracy even at 5 events.
  - performance further decreases with longer event sets.
  - differences between model sizes are less pronounced than for explicit questions.
- Interpretation:
  - the combination of vagueness and long event sets is the most difficult setting.



Summary of Experimental Findings

- **Prompting:**
  - CoT Review + Date-Only + Language encoding works best overall.
- **Models:**
  - larger models (Llama3-70B, GPT-4) clearly outperform smaller ones.
- **Temporal explicitness:**
  - Explicit > Implicit (speech-time) > Vague in terms of accuracy.
- **Question templates:**
  - yes/no questions are easiest; open or counting questions are harder.
- **Event set length:**
  - performance degrades significantly as the number of events increases, especially for implicit and vague questions (supports H3).

Discussion: What Does TRAVELER Show?

- **Temporal reasoning is fragile**
  - Small changes in how time is expressed
  - or how many events are present
  - can strongly affect LLM performance.
- **Surface matching vs. real reasoning**
  - Models cope well with simple, explicit questions.
  - Performance drops when they must chain events, align with speech time, or interpret vague expressions.
- **Event-set length as a stress test**
  - Long event lists expose limits of in-context memory and attention, even for large models.

Limitations and Implications

- **Simplified, synthetic setting**
  - Restricted home-environment events and vocabularies.
  - Does not yet cover open-domain narratives or real documents.
- **Partial coverage of temporal phenomena**
  - Focus on explicit, speech-time implicit, and vague references.
  - Excludes commonsense temporal knowledge and personal, user-specific timelines.
- **Implications for model design**
  - Chain-of-thought helps only moderately: better prompts alone are not enough.
  - Robust event-temporal reasoning likely needs explicit temporal structure or dedicated components, not just larger base models.

Future Directions

- **Conclusion**
  - TRAVELER isolates temporal explicitness and event-set length and shows that even strong LLMs struggle with long histories and vague references.
- **Towards better temporal reasoning**
  - Integrate explicit temporal memory or temporal graphs to store and query events.
  - Combine LLMs with formal temporal reasoners or function calls for date comparison and ordering.
  - Develop temporal-aware training pipelines that emphasise temporal spans, relations, and time-sensitive feedback.
- **Extending TRAVELER**
  - More diverse event types and domains, additional temporal categories, and richer, more realistic contexts.