Time Awareness in Large Language Models: Benchmarking Fact Recall Across Time

Anonymous ACL submission

Abstract

Who is the US President? The answer changes depending on when the question is asked. While large language models (LLMs) are evaluated on various reasoning tasks, they often miss a crucial dimension: time. In real-world scenarios, the correctness of answers is frequently tied to temporal context. To address this gap, we present a novel framework and dataset spanning over 8,000 events from 2018 to 2024, annotated with day-level granularity and sourced globally across domains such as politics, science, and business. Our *TimeShift* evaluation method systematically probes LLMs for temporal reasoning, revealing that base models often outperform instruction-tuned and synthetic-trained counterparts on time-sensitive recall. Additionally, we find that even large-scale models exhibit brittleness in handling paraphrased facts, highlighting unresolved challenges in temporal consistency. By identifying these limitations, our work provides a significant step toward advancing time-aware language models capable of adapting to the dynamic nature of real-world knowledge.

1 Introduction

Large language models (LLMs) have revolutionized natural language understanding, reasoning, and factual recall, becoming foundational tools for applications such as chat bots (Brown et al., 2020; OpenAI et al., 2024; Touvron et al., 2023), search engines (Thakur et al., 2021), and automated fact-checkers (Petroni et al., 2019; Roberts et al., 2020). However, their ability to handle time-sensitive facts—a critical component of real-world knowledge—remains under-explored. In many scenarios, the correctness of an answer depends not only on the question but also on when it is asked. For example, "Who is the US President on November 9, 2020, versus January 21, 2021?" requires reasoning tied to specific dates, a capability that current benchmarks often overlook.

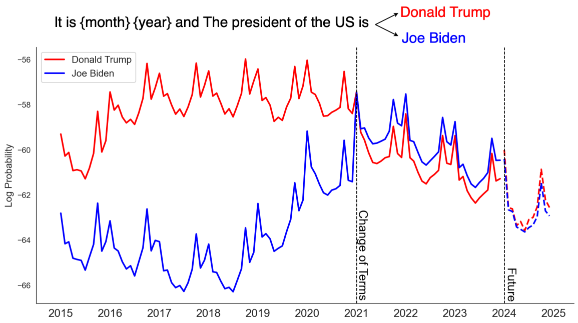

Figure 1: Temporal log probabilities of sentences predicting the U.S. president (Joe Biden or Donald Trump) using Llama 3.2 3B, showing a clear shift in predictions aligned with their terms. As the model's training data cuts off at the end of 2023, predictions beyond this point reflect extrapolated trends.

Time awareness is crucial for dynamic tasks such as real-time fact-checking, knowledge base maintenance, and temporal question answering. While LLMs excel at static factual recall and general reasoning, their performance on time-dependent queries remains an open challenge. To address this, our approach systematically probes models for temporal reasoning by measuring the log probabilities of time-sensitive sentences across different temporal contexts. For example, we evaluate whether the log probabilities of sentences like "Donald Trump is the US president" and "Joe Biden is the US president" shift appropriately as leadership changes over time. As illustrated in Figure 1, our approach captures these temporal dynamics, with models like Llama 3.1 8B (Dubey et al., 2024) showing partial success in adjusting predictions based on temporal prefixes. This highlights the importance of fine-grained temporal evaluation, which current benchmarks fail to capture comprehensively.

To address this gap, we introduce a novel dataset and evaluation framework designed to rigorously test daily temporal awareness in LLMs. Our dataset spans over 8,000 events from 2018 to 2024, annotated with day-level granularity and sourced globally across diverse domains such as politics, science, and business. Each event is paired with paraphrases to evaluate robustness in fact recall when phrasing varies. Using our *TimeShift* evaluation method, we systematically probe models by generating temporal variations to assess their ability to reason across time and paraphrased contexts.

Table 1 highlights examples from our dataset, showcasing the diversity of events and annotations. This fine-grained, systematic approach allows us to uncover limitations in temporal reasoning across model families, including instruction-tuned models and synthetic-trained architectures.

Our contributions are summarized as follows:

- We introduce a comprehensive dataset with over 8,000 events spanning seven years, annotated with day-level granularity and paired with paraphrases, enabling robust evaluation of time-sensitive fact recall.

- We propose *TimeShift*, a novel evaluation framework that systematically probes models' temporal reasoning capabilities, uncovering key limitations in handling time-dependent queries.

- We provide a detailed evaluation of over a dozen state-of-the-art open-source LLMs, revealing that base models often outperform instruction-tuned and synthetic-trained models. Surprisingly, even large models exhibit brittleness when paraphrased facts are tested.

- All data, code, and evaluation tools are open-sourced to encourage further research into temporal reasoning in LLMs.

By addressing a critical gap in current benchmarks, this work lays the groundwork for advancing time-aware LLMs capable of reasoning about the dynamic nature of real-world knowledge.

2 Related Work

Several datasets have been introduced to evaluate the temporal reasoning capabilities of LLMs. The *TempReason* dataset (Tan et al., 2023) and *TRAM* benchmark (Wang and Zhao, 2024) both focus on assessing LLMs' understanding of event order, duration, and frequency. However, these benchmarks primarily target broader temporal reasoning tasks rather than specific factual recall at finer time resolutions, such as determining the exact month when an event occurred.

An alternative approach involves modifying the self-attention mechanism (Vaswani et al., 2023) to incorporate temporal information (Rosin and Radinsky, 2022), improving performance on semantic change detection tasks (Schlechtweg et al., 2020; Hamilton et al., 2018). However, these adaptations have not been evaluated for their ability to recall specific temporal facts.

In addition, the *TempLAMA* dataset (Dhingra et al., 2022) probes LLMs on facts associated with specific years but does not extend to the month or day-level precision required for many real-world applications. Similarly, the *Test of Time* benchmark (Fatemi et al., 2024) explores event relationships over time but lacks the focus on precise, time-bound factual recall.

3 Dataset

Our dataset is designed to assess LLMs' temporal awareness, specifically their ability to recall facts tied to specific dates. It comprises over 8,000 significant events from 2018 to 2024 across politics, business, science, art, and crime, ensuring geographical and cultural diversity. As an English-language dataset, geographically the highest event concentration is in the United Stated (3,700+), followed by global (≈ 950) and UK (≈ 330) as illustrated in Figure 2.

Events are evenly distributed across months and days, though seasonal variations exist (e.g., increased reporting in summer, slight weekend decline). Each event is concisely represented by a headline of no more than 30 words, ensuring clarity and brevity, and is sourced from reputable and authoritative outlets (Section 3.1) to ensure accuracy and credibility.

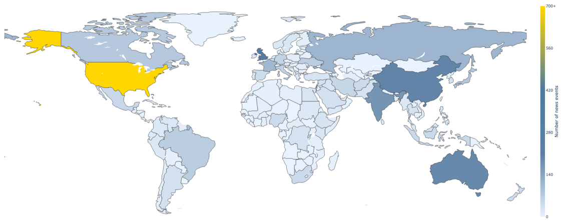

Figure 2: World map showing the amount of news per country, US is in the first place with over 3,700 events across the 7 years.

Original Sentence	Paraphrase 1	...	...	...	Year	Month	Day	Category
Rolling Stone magazine co-founder Jann Wenner...	Jann Wenner, co-founder of...	...	...	...	2023	9	16	Entertainment & Arts
Meta launches Threads - Instagram's new...	Meta introduces Threads, a new app.."	...	...	...	2023	7	5	Science & Technology

Table 1: Examples from our dataset containing over 8,000 events with precise timestamps and paraphrases. For clarity, we display only a subset of paraphrases, omitting some metadata (country, source URLs) from this table.

3.1 Data Collection and Structure

The dataset was constructed by employing a custom web-scraping pipeline that extracted headlines from major global news outlets (e.g., BBC (BBC News, 2023), Reuters (Reuters, 2023), The New York Times (nyt, 2023)), academic journals (e.g., Nature (Nature Editorial Board, 2022)), and government publications (e.g., official government websites, United Nations reports (United Nations, 2022)). To ensure accuracy, automated filtering mechanisms cross-referenced timestamps and removed duplicates, while heuristic-based checks discarded ambiguous events lacking clear temporal markers. Events with conflicting date information across sources were excluded to maintain consistency.

Each event in the dataset is annotated with its exact day, month, and year and is accompanied by four paraphrased versions. These paraphrases were generated through a combination of text transformation models and cross-source comparisons, ensuring variation in expression while preserving factual accuracy and similar length distribution (Appendix A.1). This variation is essential for evaluating the robustness of LLMs in factual recall when events are expressed differently. The dataset is specifically designed to assess whether models can recognize events despite rewording. Table 1 provides an example of the dataset structure.

To categorize events, metadata tags were extracted during the scraping. If not available, we used a lightweight LLM-based classifier trained on labeled event data to infer these attributes.

By employing rigorous filtering, multi-source validation, and LLM-assisted classification, our dataset provides a high-fidelity benchmark for evaluating LLMs' ability to recall time-sensitive facts with precision.

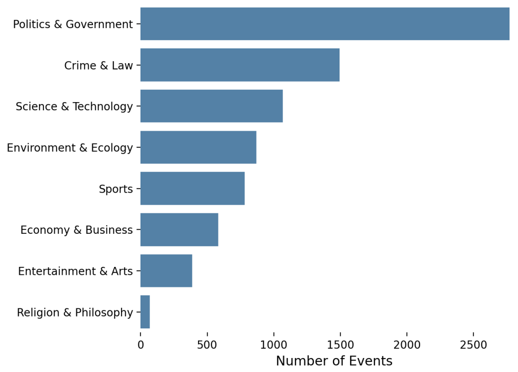

Figure 3: Distribution of events across categories, showing the highest concentration in Politics & Government and Crime & Law categories.

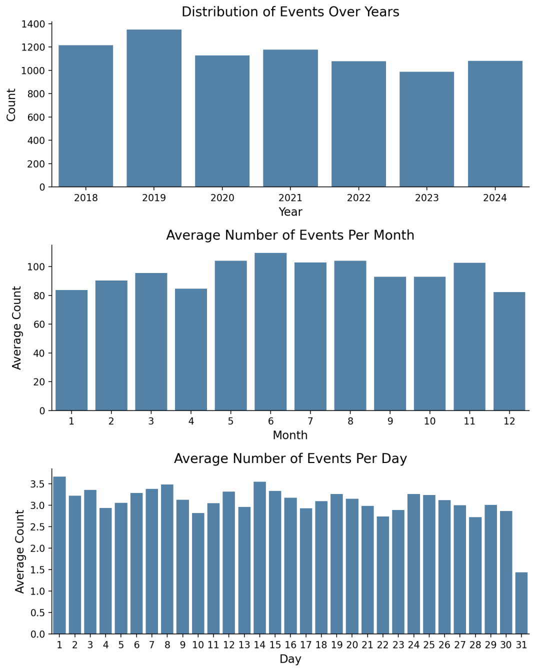

Figure 4: Even distribution of events across years, months, and days, ensuring balanced temporal coverage for evaluation.

3.2 Category and Temporal Distribution

The dataset spans a diverse range of categories, as illustrated in Figure 3. On average, each day includes approximately three events, with some seasonal variations—such as a slight increase during summer months and a decline on weekends. The temporal distribution across years, months, and days is shown in Figure 4, ensuring a balanced representation that prevents any specific period from disproportionately affecting the results.

Prefix Format	Year Acc.	Month Acc.	Day Acc.	Final Acc.
`It is {day} {month}, {year}. {event}`	31.5%	4.9%	0.4%	**12.3%**
`It is {month} {day}, {year}. {event}`	31.5%	4.9%	0.2%	12.2%
`It is {year} {month}, {day}. {event}`	31.5%	2.5%	0.2%	11.4%
`It is {year} {month}, {day} and {event}`	31.5%	2.5%	0.0%	11.3%
`{day}.{month}.{year}, {event}`	29.6%	3.4%	0.2%	11.1%
`On {month} {day}, {year}, {event}`	26.9%	4.4%	0.5%	10.6%
`On {day}/{month}/{year}, {event}`	26.9%	4.4%	0.1%	10.5%
`{year}-{month}-{day}: {event}`	26.7%	3.2%	0.0%	10.0%

Table 2: Comparison of selected date-prefix formats based on accuracy in predicting time-sensitive facts. We tested a wide range of prefixes and report the best-performing ones.

3.3 Public Availability

The dataset, along with the evaluation framework, is publicly available on HuggingFace and GitHub, providing the research community with an accessible resource to further explore time-sensitive fact recall in LLMs.[1]

4 Experiments

The core hypothesis driving our dataset is that an LLM should assign the highest probability to the sentence describing an event with the correct temporal prefix—specifically, the day, month, and year in which the event occurred. This hypothesis underpins the evaluation setup, where the model is tested on its ability to select the correct temporal context from a range of possibilities.

For example, consider the sentence: "It is April 13, 2022. Rolling Stone magazine co-founder Jann Wenner..." Here, the temporal prefix ("It is April 13, 2022.") explicitly situates the event within a specific timeframe, providing a clear basis for the model's probabilistic assessment. This specific prefix was selected based on additional experiments in Section 4.1, where it best aligned predictions with temporal context.

4.1 Prefix selection

Selecting the optimal prefix for evaluating temporal awareness in LLMs is crucial, as phrasing affects how models interpret time-sensitive queries. To identify the best-performing prefix, we tested various prefix formulations on 10% of the dataset using Llama-3.2 1B, Llama-3.2 3B, and Gemma-2 2B. These models, spanning different parameter sizes and architectures, provided a representative assessment of prefix impact on the performance. We explored variations in word order (e.g., year-first, day-first), separators (e.g., commas, dashes, slashes), and explicit prepositions (e.g., "On date" vs. "It is date").

From this extensive search, Table 2 reports the best-performing prefixes. The highest final accuracy of 12.3% was achieved with "It is {day} {month}, {year}. {event}", making it the optimal choice for probing LLMs' temporal recall. Notably, prefixes starting with the year (e.g., "It is year month, day and event") reduced accuracy, suggesting models overemphasized the year while struggling with finer details. Similarly, while numerical date formats using separators (e.g., ”day.month.year, event” or ”year-month-day: event”) performed reasonably, they exhibited slightly lower day-level accuracy. Based on these findings, we adopt the top-performing prefix across all subsequent experiments to ensure reliable and consistent temporal evaluation of LLMs.

4.2 TimeShift Algorithm

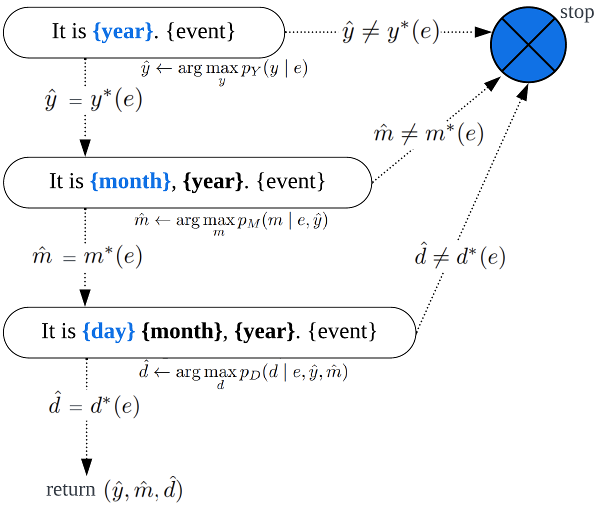
It is {year}. {event}
$\hat{y} \neq y^*(e)$
stop
$\hat{y} \leftarrow \arg\max_y p_Y(y \mid e)$
$\hat{y} = y^*(e)$
$\hat{m} \neq m^*(e)$
It is {month}, {year}. {event}
$\hat{m} \leftarrow \arg\max_m p_M(m \mid e, \hat{y})$
$\hat{m} = m^*(e)$
$\hat{d} \neq d^*(e)$
It is {day} {month}, {year}. {event}
$\hat{d} \leftarrow \arg\max_d p_D(d \mid e, \hat{y}, \hat{m})$
$\hat{d} = d^*(e)$
return $(\hat{y}, \hat{m}, \hat{d})$

Figure 5: Schema of the *TimeShift* algorithm. Nodes represent sentences for which probabilities are computed with varying temporal prefixes (in blue). The sentence with the highest probability is selected as the prediction.

[1] https://huggingface.co/datasets/

The *TimeShift* algorithm evaluates a model's ability to correctly predict the time of occurrence of an event. Given an event e, our goal is to predict its correct date, structured as:

- $y^*(e)$ — the true year of the event,
- $m^*(e)$ — the true month of the event,
- $d^*(e)$ — the true day of the event.

The model generates probability distributions, specifically representing the likelihood of the entire sentence, including the temporal prefix. To ensure numerical stability, we compute the sum of the logarithms of these probabilities for each time unit:

- $p_Y(y \mid e)$ — probability of the event occurring in year y,
- $p_M(m \mid e, \hat{y})$ — probability of the event occurring in month m, given the predicted year $\hat{y}$,
- $p_D(d \mid e, \hat{y}, \hat{m})$ — probability of the event occurring on day d, given the predicted year $\hat{y}$ and month $\hat{m}$.

Instead of evaluating all possible (year, month, day) combinations, we apply a sequential filtering approach as described in Algorithm 1 and Schema 5 improving efficiency while preserving accuracy.

Algorithm 1 TimeShift

```
Input: Event $e$
Output: Predicted date $(\hat{y}, \hat{m}, \hat{d})$
Step 1: Predict Year
$\hat{y} \leftarrow \arg\max_{y} p_Y(y \mid e)$
if $\hat{y} \neq y^*(e)$ then
    Stop (Incorrect Year)
end if
Step 2: Predict Month (If Year is Correct)
$\hat{m} \leftarrow \arg\max_{m} p_M(m \mid e, \hat{y})$
if $\hat{m} \neq m^*(e)$ then
    Stop (Incorrect Month)
end if
Step 3: Predict Day (If Month is Correct)
$\hat{d} \leftarrow \arg\max_{d} p_D(d \mid e, \hat{y}, \hat{m})$
if $\hat{d} \neq d^*(e)$ then
    Stop (Incorrect Day)
end if
Return $(\hat{y}, \hat{m}, \hat{d})$
```

4.3 Stability Measurement Algorithm

The Stability Measurement algorithm evaluates the robustness of a model's predictions under minor input variations. Instead of measuring absolute accuracy, it quantifies whether the model maintains consistent predictions when an event description is paraphrased.

Given an event e, the dataset provides:

- **Original sentence:** e (news event).
- **Paraphrased sentences:** e'_1, e'_2, e'_3, e'_4 (four paraphrases of e).
- **True label:** $y^*(e)$ (ground-truth year, month, or day).

Given these inputs, the model produces:

- **Prediction for the original sentence:** $\hat{y}(e)$.
- **Predictions for the paraphrased sentences:** $\hat{y}(e'_i)$ for $i \in \{1, 2, 3, 4\}$.

We define the following probability metric:

Stability Probability: The probability that the model predicts the same correct result for a paraphrased event, given that it was correct for the original event:

$$S = P(\hat{y}(e'_i) = y^*(e) \mid \hat{y}(e) = y^*(e))$$

Here, S quantifies the stability of predictions under paraphrasing. A high S indicates that the model is robust to rewording, whereas a low S suggests that the model is sensitive to variations in input phrasing.

4.4 Why Use Log Probabilities Instead of Direct QA?

A natural alternative would be to directly prompt the model with an open-ended question, such as *"When did this event occur?"*, and compare the generated response to the ground truth. However, model responses in a free-form QA setting are inherently non-deterministic—even at temperature zero, variations in rounding, tokenization, and parallelization can lead to inconsistencies. This makes direct QA evaluation difficult to reproduce.

Log probabilities offer a structured, reproducible alternative by ranking all possible dates for an event in a probabilistic framework. While one could attempt to rank QA-based responses (e.g., via likelihoods or multiple-choice scoring), such approaches introduce additional challenges, such as inconsistent specificity (e.g., "early 2023" vs. "January 5, 2023") and reliance on post-processing heuristics. Our method avoids these pitfalls by ensuring a consistent evaluation framework across models.

4.5 Metrics

We evaluate models using two key metrics:

- **Accuracy**: Accuracy is evaluated at multiple temporal granularities:
  - **Yearly Accuracy**: The probability that the model correctly predicts the year of an event.
  - **Monthly Accuracy**: Computed only for instances where the year is correctly predicted. Since it is conditioned on a correct yearly prediction, its sample size is smaller.
  - **Daily Accuracy**: Assessed only when both the year and month predictions are correct.
  - **Approximate Daily Accuracy**: Similar to daily accuracy but allowing for a ±1-day margin of error.
  - **Total Daily Accuracy**: The overall probability of correctly predicting an event's exact date, computed as the product of yearly, monthly, and daily accuracies.

- **Stability**: Stability measures the model's robustness to input paraphrasing, evaluating its consistency in predicting the correct date across reworded event descriptions. Stability is computed at the following levels:
  - **Yearly Stability**: The probability that the model predicts the correct year for a paraphrased event, given that it was correctly predicted for the original event.
  - **Monthly Stability**: Evaluated only for instances where the model correctly predicts the year.
  - **Daily Stability**: Assessed when both the year and month predictions are correct.
  - **Approximate Daily Stability**: Similar to daily stability but allowing for a ±1-day margin of error.

5 Results

Our experiments reveal several clear trends in LLM performance on the time-sensitive fact recall task. We evaluated base and instruction-tuned variants, with parameter sizes ranging from 1B to 72B. Accuracy results are presented in Table 3, while stability results are detailed in Table 4.

5.1 Instruction-Tuned Models Underperform

Across all model families, instruction-tuned variants underperform compared to base models on this task. For instance, Gemma-27B achieves 55.16% Yearly accuracy but drops to 44.11% after instruction tuning. Similarly, Llama-3.1 8B outperforms its instruction-tuned counterpart, achieving 38.17% versus 33.05%.

We hypothesize that the broad generalization achieved during instruction tuning dilutes time-specific factual recall, prioritizing task flexibility over detailed temporal knowledge.

5.2 Impact of Model Size on Performance

Model size exhibits a strong correlation with performance on our time-awareness benchmark, with larger models consistently outperforming smaller ones across all metrics. However, Qwen-2.5 72B deviates from this trend, underperforming relative to its parameter count. For instance, Gemma-2 27B Base achieves a Yearly accuracy of 55.16%, surpassing Gemma-2 9B (45.45%) and Gemma-2 2B (35.24%). This pattern aligns with broader findings that larger models more effectively capture nuanced information, including temporal dependencies. Qwen-2.5's weaker performance may stem from differences in training data quality, suboptimal fine-tuning, or a focus on multilingual generalization over precise factual retrieval.

5.3 Underperformance of Synthetic-Training Models

Despite excelling in reasoning and generation, synthetic-trained models like the Phi family struggle with temporal recall. Phi-3-mini 3.8B achieves only 26.00% Yearly accuracy, while the larger Phi-4 14B reaches 34.26%. Notably, the 1B parameter Llama-3.2 nearly matches Phi-4 with 33.35%, suggesting that increased parameter size alone cannot offset the limitations of synthetic data. This highlights a key weakness: synthetic datasets, though effective for general knowledge, often lack the real-

Models	Yearly Acc	Monthly Acc	Daily Acc	Daily Acc Approx	Daily Total
Qwen-2.5 1.5B	22.07%	11.98%	3.30%	10.38%	0.09%
Qwen-2.5 1.5B	23.48%	11.74%	3.62%	13.57%	0.10%
Llama-3.2 1B	25.89%	11.61%	4.15%	5.81%	0.12%
Phi-3.5-mini 3.8B	27.57%	14.16%	3.83%	9.90%	0.15%
Phi-3-mini 3.8B	26.00%	15.16%	4.11%	10.76%	0.16%
Gemma-2 2B	29.65%	14.34%	4.99%	13.20%	0.21%
Qwen-2.5 7B	29.94%	14.45%	5.48%	12.68%	0.24%
Phi-4 14B	34.26%	18.70%	3.89%	10.70%	0.25%
Qwen-2.5 7B	32.28%	14.87%	5.45%	13.51%	0.26%
Llama-3.2 1B	33.35%	13.57%	6.06%	11.57%	0.27%
Llama-3.2 3B	32.81%	17.75%	4.71%	14.56%	0.27%
Gemma-2 2B	35.24%	17.13%	5.58%	11.36%	0.34%
Llama-3.2 3B	40.86%	20.75%	4.12%	11.18%	0.35%
Gemma-2 9B	40.11%	23.25%	4.68%	13.37%	0.44%
Gemma-2 9B	45.45%	24.28%	6.55%	15.82%	0.72%
Llama-3.1 8B	46.92%	27.66%	5.67%	12.30%	0.74%
Llama-3.1 8B	48.47%	27.42%	5.72%	12.10%	0.76%
Mistral-Nemo 12B	50.62%	31.70%	7.38%	14.69%	1.18%
Gemma-2 27B	44.11%	35.98%	7.78%	17.75%	1.23%
Qwen-2.5 72B	46.35%	32.66%	10.63%	21.66%	1.61%
Qwen-2.5 72B	48.47%	38.15%	**11.19%**	**22.59%**	2.07%
Gemma-2 27B	**55.16%**	**42.09%**	8.97%	18.58%	**2.08%**

Table 3: Performance of large language models on the time-awareness benchmark. The table reports accuracy metrics at different granularities: yearly, monthly, daily, and an approximate daily measure that allows for a ±1-day error margin. The daily total accuracy reflects the model's likelihood of correctly assigning an event to its exact day, as further detailed in Section 4.5. The results highlight the performance gap between instruction-tuned models (rows highlighted in blue) and their non-tuned counterparts, as well as the general accuracy improvement with increasing model size—except for Qwen-2.5 72B, which underperforms relative to smaller models. Gemma-2 27B achieves the highest overall accuracy.

Models	Yearly stability	Monthly stability	Daily stability	Daily stability Approx
Qwen-2.5 1.5B	11.14%	0.93%	0.02%	0.02%
Phi-3-mini 3.8B	12.31%	1.11%	0.02%	0.02%
Qwen-2.5 1.5B	13.16%	0.84%	0.03%	0.03%
Phi-3.5-mini 3.8B	14.72%	1.38%	0.03%	0.04%
Llama-3.2 1B	16.34%	0.98%	0.03%	0.03%
Qwen-2.5 7B	17.99%	2.02%	0.07%	0.07%
Gemma-2 2B	18.75%	1.91%	0.07%	0.09%
Gemma-2 2B	20.31%	2.33%	0.07%	0.07%
Llama-3.2 1B	20.42%	1.95%	0.06%	0.07%
Phi-4 14B	20.92%	2.81%	0.07%	0.09%
Qwen-2.5 7B	22.34%	3.40%	0.07%	0.09%
Llama-3.2 3B	22.60%	2.26%	0.04%	0.05%
Llama-3.2 3B	25.77%	3.72%	0.07%	0.07%
Gemma-2 9B	32.31%	9.02%	2.50%	2.78%
Llama-3.1 8B	33.05%	9.12%	2.61%	2.52%
Qwen-2.5 72B	33.52%	9.12%	2.00%	2.00%
Qwen-2.5 72B	35.05%	9.29%	2.33%	2.37%
Gemma-2 27B	36.39%	9.02%	2.12%	2.52%
Mistral-Nemo 12B	37.16%	9.31%	2.72%	2.75%
Llama-3.1 8B	38.17%	10.33%	2.69%	2.72%
Gemma-2 9B	38.19%	10.67%	2.47%	2.91%
Gemma-2 27B	**39.39%**	**11.02%**	**3.12%**	**3.12%**

Table 4: Model stability analysis. Again, the instruct-tuned models are highlighted in blue. The table illustrates a general trend of increasing stability as the number of parameters grows. It also highlights the challenges of prompting the model for finer-grained time horizons—rephrasing a sentence while expecting the model to predict the exact same year, month, and day as in the original phrasing proves to be particularly difficult.

world temporal grounding needed for accurate recall of time-sensitive facts.

5.4 Performance over the years

Nearly all models evaluated in this study were trained on datasets with a cut-off date of December 2023, as is explicitly stated for the Llama model family. For other models, which were released in the first half of 2024 but do not specify a precise cut-off date, it is reasonable to assume their training data extends to early 2024. This temporal boundary is clearly reflected in their performance, with a marked decline in accuracy for more recent facts, as illustrated in Figure 6. Interestingly, all models perform worse on recent events and better on older ones, likely due to the higher availability, stability, and reinforcement of past information in training data, whereas newer events are less represented and may undergo evolving interpretations.

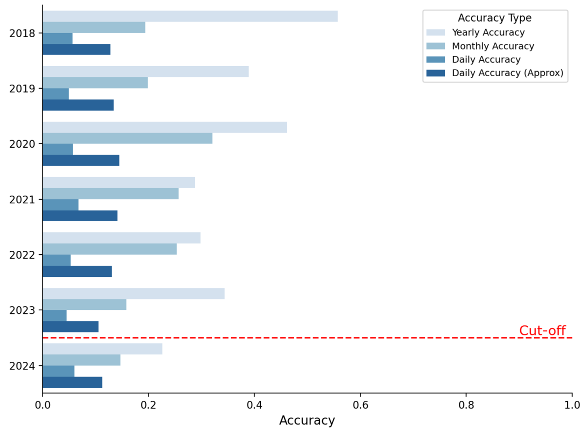

Figure 6: Averaged performance of all models across years, showing a clear drop in performance as the cut-off date approaches.

5.5 Stability Across Paraphrases

Our stability evaluation highlights a significant susceptibility of models to variations in phrasing. Even the best-performing model in our study, Gemma-2 27B Base, correctly predicted the year in only 39.39% of paraphrased cases, despite accurately classifying the original event. Smaller and instruction-tuned models, such as Llama-3.2 1B Base (20.42%) and Qwen-2.5 1.5B Instruct (11.14%), exhibited even greater sensitivity, frequently altering their predictions in response to minor rewording. This trend becomes even more pronounced as temporal granularity increases, with accuracy dropping to single-digit percentages for the best models in Daily stability assessments.

These findings reinforce a well-documented issue in LLMs: prompt sensitivity (Zhuo et al., 2024; Sclar et al., 2024; Zhan et al., 2024). Larger models tend to demonstrate greater stability than their smaller counterparts, but even they struggle with consistency when faced with slight linguistic variations, emphasizing the ongoing challenge of robustness in factual recall.

6 Conclusion

In this paper, we introduced a novel dataset and evaluation benchmark specifically designed to assess LLMs' ability to handle time-sensitive facts, addressing a critical gap in existing evaluation methods that primarily focus on static factual recall. Our dataset, comprising over 8,000 events spanning from 2018 to 2024, provides a structured framework for testing models' temporal awareness by evaluating their ability to correctly associate events with their respective time periods.

Beyond assessing temporal accuracy, our dataset also enables the evaluation of model stability, offering insights into their robustness against variations in prompt phrasing. This aspect is crucial for understanding how reliably a model can maintain consistency in its predictions.

Our findings indicate that larger models consistently outperform smaller ones in time-sensitive tasks, reinforcing the role of scale in factual recall. However, instruction-tuned models, despite their strengths in general-purpose reasoning, struggle with temporal reasoning. Additionally, models trained primarily on synthetic data, such as those in the Phi family, demonstrate notable limitations in real-world temporal understanding. Furthermore, our analysis highlights the significant impact of prompt formulation on model behavior, revealing how slight variations in wording can lead to different predictions.

Time awareness is essential for real-world applications such as virtual assistants, fact-checking, and temporal question-answering. By publicly releasing our dataset and evaluation framework, we aim to support further research in this area and encourage the community to extend this work.

Limitations

While our benchmark provides a rigorous evaluation of time-sensitive factual recall, it has certain limitations. First, the dataset is exclusively in English, which may limit its applicability to multilingual LLMs. Future work should extend the dataset to include non-English events, enabling broader linguistic coverage. Second, our evaluation framework focuses on open-source models, as log probability access is required to perform a structured ranking of temporal claims. This constraint prevents us from directly assessing closed-source models, such as GPT-4 or Claude, unless they provide likelihood scores. Third, while our dataset covers diverse event categories and global sources, there remains an overrepresentation of Western events, particularly from the United States and the United Kingdom, due to the nature of English-language reporting. Expanding the dataset with multilingual and regionally balanced sources could mitigate this bias.

References

2023. The new york times. Accessed: 2024-08-16.

BBC News. 2023. Bbc news. Accessed: 2024-08-16.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. *Preprint*, arXiv:2406.09170.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Diachronic word embeddings reveal statistical laws of semantic change. *Preprint*, arXiv:1605.09096.

Nature Editorial Board. 2022. Nature journal. *Nature*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *Preprint*, arXiv:1909.01066.

Reuters. 2023. Key economic and business events of 2023. Accessed: 2024-08-16.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *Preprint*, arXiv:2002.08910.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. *Preprint*, arXiv:2306.08952.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Preprint*, arXiv:2104.08663.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

United Nations. 2022. United nations annual report 2022. `https://www.un.org/en/annualreport`. Accessed: 2024-08-16.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Yuqing Wang and Yun Zhao. 2024. Tram: Benchmarking temporal reasoning for large language models. *Preprint*, arXiv:2310.00835.

Pengwei Zhan, Zhen Xu, Qian Tan, Jie Song, and Ru Xie. 2024. Unveiling the lexical sensitivity of llms: Combinatorial optimization for prompt enhancement. *Preprint*, arXiv:2405.20701.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

In this section, we provide additional details on dataset properties, evaluation distributions, and experimental results. These supplementary figures further illustrate key aspects of our dataset's structure and the robustness of our evaluation framework.

A.1 Paraphrase Length Distribution

To ensure fair evaluation, we maintain similar length distributions between original event descriptions and their paraphrased variants. As shown in Figure 7, the paraphrased sentences closely follow the length distribution of the original events. This minimizes potential biases where longer or shorter phrasings could disproportionately affect model performance.

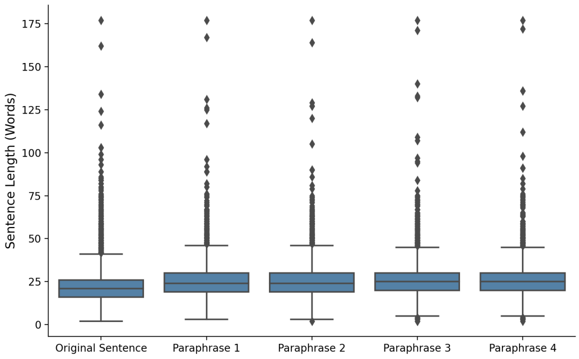

Figure 7: Comparison of sentence length distributions between original event descriptions and their paraphrased counterparts. The alignment of distributions ensures that paraphrases do not introduce systematic biases in model evaluation.

A.2 Accuracy Across Event Categories

To assess whether certain event categories are easier for models to predict, we analyze accuracy distributions across different domains. As depicted in Figure 8, the model performance remains relatively stable across categories such as politics, science, and entertainment. This suggests that the temporal reasoning task is not inherently skewed toward specific domains, reinforcing the general applicability of our benchmark.

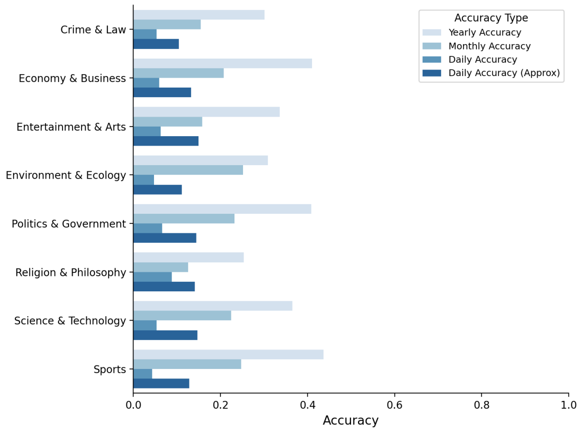

Figure 8: Accuracy distribution across different event categories. The relatively uniform performance indicates that no single category disproportionately influences model accuracy, confirming the dataset's balanced composition.

A.3 Accuracy Across Geographical Regions

Similarly, we analyze model performance across different continents to ensure that temporal reasoning is not biased toward specific geographic regions. Figure 9 shows that accuracy is relatively consistent across continents, further supporting the robustness of our evaluation framework. While English-language news sources naturally introduce an overrepresentation of Western events (e.g., USA and UK), our dataset remains diverse enough to challenge models across a wide range of global events.

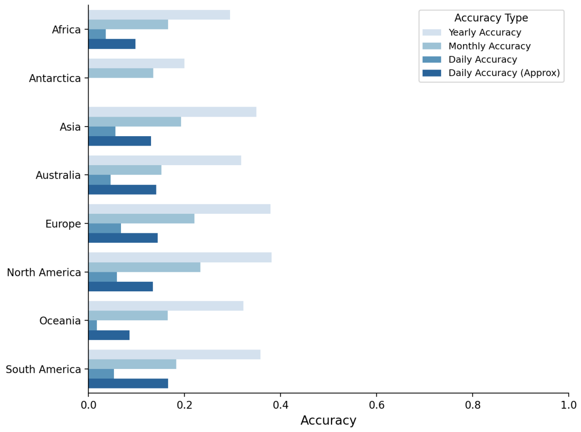

Figure 9: Accuracy distribution across continents. The balanced accuracy levels suggest that the dataset provides a fair temporal reasoning challenge across diverse geographic regions, despite the English-language focus.

A.4 Additional Insights and Future Work

Overall, the even accuracy distribution across event categories and geographic regions reinforces the robustness of our dataset. The alignment of paraphrased sentence lengths with original event descriptions ensures that variations in sentence structure do not introduce unintended evaluation biases.

Future work could extend the dataset by incorporating multilingual event sources, enabling evaluation across different linguistic and cultural contexts. Additionally, refining the event selection pipeline to ensure better coverage of underrepresented regions could further enhance the dataset's utility.