NOT ALL TIME IS GREGORIAN: EVALUATING LLMS ON CULTURAL CALENDAR SYSTEMS

Deepon Halder
AI4Bharat, IIEST Shibpur
deeponh.2004@gmail.com

Adish Pandya
AI4Bharat, SRM University

Raj Dabre
AI4Bharat, IIT Madras

ABSTRACT

Large Language Models (LLMs) demonstrate strong temporal reasoning and historical fact retrieval, yet existing benchmarks rely almost exclusively on the Gregorian calendar, implicitly treating Western temporal standards as universal. This Gregorian-centric framing obscures a critical limitation: current foundation models fail to reason reliably within culturally diverse, non-Gregorian calendar systems used by billions worldwide. We introduce **ChronoBench**, a diagnostic benchmark for temporal reasoning across five major cultural calendars: **Vikram Samvat, Persian (Jalali), Hijri, Chinese Lunar, and Hebrew**[1]. The benchmark evaluates two core capabilities: *Event Date Retrieval*, measuring factual grounding in indigenous timelines, and *Date Arithmetic*, probing structural reasoning over non-linear temporal constructs such as intercalary months and lunar cycles. Evaluating several open-weight models, including *Gemma-3*, *DeepSeek-V3*, and *Qwen-32B*, reveals pronounced performance disparities. While reasoning-optimized models such as *DeepSeek-R1* show localized competence in solar calendars (e.g., Persian), performance collapses for lunisolar and purely lunar systems. Models consistently exhibit a Gregorian anchoring effect, defaulting to linear offsets or Western mathematical heuristics even when prompted within alternative calendar frameworks. These findings expose a deep-seated Gregorian bias in open foundation models, suggesting that temporal reasoning is often memorized rather than structurally learned. Our work identifies a key bottleneck in cultural alignment and provides a rigorous framework for developing more inclusive and robust temporal reasoning systems.

1 INTRODUCTION

The democratization of Large Language Models (LLMs) has necessitated a rigorous examination of their cultural inclusivity. While models excel at standard temporal tasks, such as calculating durations or ordering events, these capabilities are implicitly tied to the ISO 8601 standard and the Gregorian calendar. For billions of users globally, however, traditional calendars remain integral to cultural, religious, and historical identity.

While the impact of cultural commonsense in LLMs is gaining traction (Pawar et al., 2024), culturally grounded temporal reasoning remains under-explored. Early investigations, such as those by Sasaki et al. (2025), focused on the Japanese *wareki* system, revealing that even frontier models struggle with arithmetic across era boundaries. More recently, Miao et al. (2026) introduced the SPAN benchmark, which evaluates cross-calendar conversion and identifies critical failure modes like *Future-Date Degradation* and *Calendar Asymmetry Bias*. While SPAN utilizes a tool-augmented Time Agent to bypass these limitations, the underlying question of whether LLMs can natively internalize non-Gregorian logic persists. Our work builds on prior efforts by broadening both the evaluation scope and analytical depth through the following contributions:

- **Unified Cross-Cultural Benchmark:** We present a benchmark covering five distinct calendar systems with diverse mathematical structures: *Vikram Samvat, Persian (Jalali), Hijri,*

[1]We release the dataset used in this work as open source at https://huggingface.co/datasets/ai4bharat/chronobench.

Chinese Lunar, and Hebrew. Importantly, we incorporate the *Vikram Samvat* calendar, a lunar-based system integral to South Asian cultural and religious practices that is frequently absent from global temporal benchmarks.

- **Diagnostic Framework for Logic vs. Retrieval:** In contrast to earlier work that mainly emphasizes conversion accuracy or tool-use ability (Miao et al., 2026), we introduce a novel framework that examines *internal arithmetic consistency*. This enables us to differentiate between shallow memorization of particular dates and a model's capacity for structural reasoning within non-linear temporal systems.

2 Non-Gregorian Calendar Mechanisms and Conversion

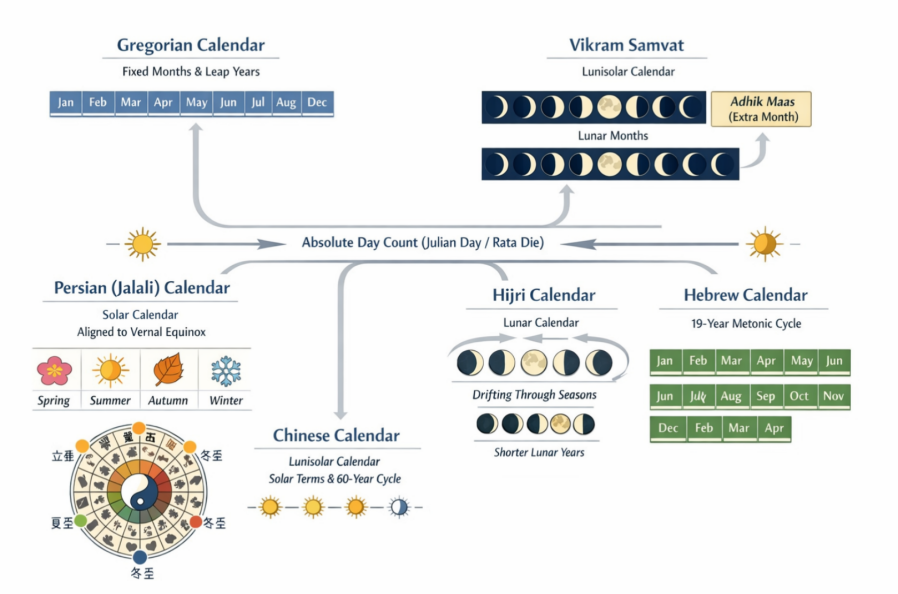

Figure 1: Visual Representation of Calendar structures.

Conversion typically follows a two-stage process: mapping a Gregorian date to an absolute, calendar-independent count (e.g., Julian Day Number) to remove variable month/leap year complications, then reconstructing the target date using system-specific rules.

- **Vikram Samvat (Lunisolar):** Synchronizes lunar months with solar years via astronomical modeling. Beyond a rough 57-year offset, accurate conversion requires calculating lunar new moons and inserting intercalary months (*Adhik Maas*) based on solar ingress.
- **Persian/Jalali (Solar):** A highly deterministic system tracking the vernal equinox. After normalization, years and months are derived via an accurate leap-year scheme and cumulative day subtractions, ensuring minimal long-term drift.
- **Hijri (Lunar):** Strictly lunar and decoupled from solar cycles, resulting in 354 or 355-day years. It employs an alternating 29–30 day month pattern, causing dates to drift approximately 11 days per Gregorian year.
- **Chinese (Lunisolar):** Combines lunar cycles with solar terms and a 60-year sexagenary cycle. Conversion is computationally heavy, as leap months are inserted only when a lunar month lacks a principal solar term.
- **Hebrew (Lunisolar):** Purely arithmetic but complex, utilizing a 19-year Metonic cycle with seven leap years. Month lengths vary based on the cycle position and the specific year type determined from the Hebrew epoch.

Algorithmically, solar calendars are the most predictable. Purely lunar systems remain straightforward despite seasonal drift, while lunisolar systems represent the highest complexity due to their reliance on either astronomical data or layered cyclical logic.

3 METHODOLOGY

3.1 DATA COLLECTION AND CURATION

The dataset was built via a multi-stage pipeline combining synthetic generation, external conversion, and human validation. We used Gemini 3.0 Flash to produce an initial list of historical events with their Gregorian dates, which were then converted into natural language sentences for evaluation prompts. Ground-truth labels for cultural calendar systems were obtained using specialized online date converters such as *nanakshahi.net* and other regional tools. To ensure benchmark integrity, both the conversion steps and final dates were human-verified against historical and cultural records. Appendix B details the temporal distribution of events, while Appendix D outlines the astronomical and mathematical standardization protocols.

3.2 TASK DEFINITION

We evaluate models on two distinct tasks designed to probe different aspects of temporal reasoning: *Event Date Retrieval* (factual grounding) and *Date Arithmetic* (logical computation). A comprehensive breakdown of prompt templates and sample distributions is available in Appendix A.

3.3 EXPERIMENTAL SETUP

We evaluated six leading open-weights large language models: Deepseek v3 (DeepSeek-AI (2024)), Deepseek R1 (DeepSeek-AI (2025)), Gemma 3 27B (Gemma Team & Google DeepMind (2024)), GPT OSS 120B (OpenAI (2024)), Mistral Small 24B (Jiang et al. (2023)), and Qwen 32B (Qwen Team & Alibaba Group (2024)). All models were evaluated under a strict zero-shot prompting setup, and given a thinking budget of 8196 tokens.

4 RESULTS

The performance of the evaluated models is summarized in Table 1. The results indicate a systemic failure to accurately reason in non-Western temporal frameworks.

Table 1: Zero-shot Exact Match (EM) Accuracy. "Retr." denotes Task 1 (Event Retrieval) and "Arith." denotes Task 2 (Date Arithmetic).

Model	**Vikram**		**Persian**		**Hijri**		**Chinese**		**Hebrew**	
	Retr.	Arith.	Retr.	Arith.	Retr.	Arith.	Retr.	Arith.	Retr.	Arith.
DeepSeek-R1	36.1%	10%	73.03%	65%	12.2%	23%	56.55%	52%	44.12%	28%
DeepSeek-V3	6.83%	8%	45.37%	42%	11.71%	9%	4.14%	0%	33.82%	8%
Gemma-3-27B	2.44%	1%	21.95%	2%	0.98%	4%	0.69%	0%	3.92%	1%
GPT-OSS-120B	5.85%	7%	18.05%	53%	10.98%	18%	4.2%	0%	4.19%	4%
Mistral-Small	6.83%	1%	14.15%	14%	0.0%	1%	1.38%	0%	4.41%	0%
Qwen-32B	5.85%	1%	11.45%	6%	0.0%	0%	0.0%	0%	4.35%	1%

4.1 ANALYSIS BY TASK

Task 1: Event Date Retrieval Factual grounding in non-Gregorian timelines remains a major challenge for all models. As shown in Table 1, although *DeepSeek-R1* achieves relatively strong retrieval accuracy in the Persian (73.03%) and Chinese (56.55%) systems, this performance is inconsistent across calendars. Qualitative error analysis shows that models frequently hallucinate using lazy heuristics. For example, when asked for the Vikram Samvat date of the *Battle of Plassey*, models often identify the correct Gregorian year (1757) but then merely add 57 years, ignoring the complex lunar month and day alignments. This Gregorian anchoring is further illustrated in Figure 2, where **Year Accuracy** remains high for most models, while **Month** and **Day** accuracies drop sharply. Sample JSON data structures for these tasks are provided in Appendix A.

Task 2: Date Arithmetic In date arithmetic tasks, models exhibit systematic failures across non-Gregorian calendars. For the Hijri calendar, they do not account for the alternating 29/30-day lunar

cycle and instead frequently default to Gregorian month lengths of 30 or 31 days. Similarly, in lunisolar systems such as the Hebrew and Vikram Samvat calendars, models consistently fail to recognize leap years or insert intercalary months (*Adhik Maas*). The heatmap data confirms this pattern: even the strongest models display a sharp gradient from green in **Year** accuracy to red in **Day** accuracy, demonstrating that as reasoning shifts from simple retrieval to structural arithmetic, the models' temporal logic fails to generalize beyond Western calendrical standards.

4.2 Granularity Analysis: Year vs. Month vs. Day

Beyond aggregate scores, we analyze model performance across varying levels of temporal granularity: *Year*, *Month*, and *Day* accuracy. As illustrated in Figure 2, there is a precipitous decline in performance as the required precision increases.

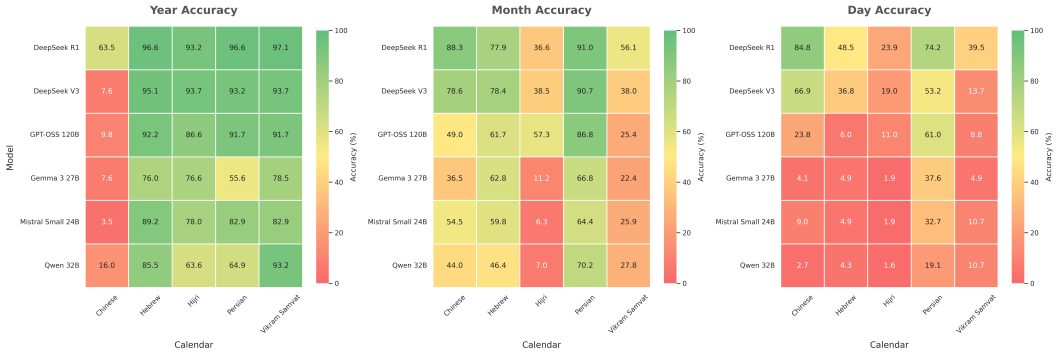

Figure 2: Heatmap of accuracy across different temporal granularities. Performance degrades significantly from Year-level retrieval to exact Day-level arithmetic across all models and calendars.

- **Year-Level Resilience:** Most models, particularly *DeepSeek-R1* and *V3*, maintain high accuracy (90%+) in retrieving the correct year for Persian, Hebrew, and Hijri systems. This suggests that the offset (e.g., Gregorian Year -622 for Hijri) is well-represented in the weights, likely due to the high frequency of year-only mentions in training corpora.
- **The Chinese-Calendar Anomaly:** Notably, almost all models except *DeepSeek-R1* fail at Year-level retrieval for the Chinese calendar ($<$10% accuracy). This is attributed to the sexagenary cycle and lunar-based New Year, which prevents a simple linear year-to-year mapping.
- **Arithmetic Decay:** The transition from *Month Accuracy* to *Day Accuracy* reveals the structural blindspot. For the Hijri and Hebrew calendars, day-level accuracy drops by over 50% compared to month-level. This confirms that while models may know an event occurred in a certain month, they lack the astronomical logic to calculate the exact day within non-30/31 day cycles.

For a more granular mechanistic analysis of how solar versus lunisolar logic affects these error rates, see the ablation study in Appendix C, and for an analysis of operational directionality (addition vs. subtraction), refer to Appendix E.

5 Conclusion

Our evaluation reveals that current Large Language Models suffer from a profound "Gregorian bias," where temporal reasoning is treated as a retrieval task rather than a structural logic. While models like DeepSeek-R1 and Gemma-3 show some proficiency in year-level retrieval, they face a systemic collapse in Date Arithmetic, with accuracy frequently dropping to near zero when faced with non-linear cycles such as intercalary months or lunar drift. This failure highlights that as AI scales, it remains tethered to Western-centric temporal standards, masking a significant bottleneck in cultural alignment. By introducing this multi-calendar benchmark, we provide a rigorous framework for moving beyond surface-level memorization toward models that can natively navigate the heterogeneous logical systems used by billions of people worldwide.

REFERENCES

DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in large language models via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

Gemma Team and Google DeepMind. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. URL https://arxiv.org/abs/2403.08295.

Albert Jiang, Alexandre Sablayrolles, Antoine Roux, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Zhongjian Miao, Hao Fu, and Chen Wei. SPAN: Benchmarking and improving cross-calendar temporal reasoning of large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv preprint arXiv:2511.09993.

OpenAI. Gpt-oss: Open-source large language models at scale. Technical report, 2024. URL https://github.com/openai/gpt-oss.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*, 2024. URL https://arxiv.org/abs/2411.00860.

Qwen Team and Alibaba Group. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2024. URL https://arxiv.org/abs/2309.16609.

Mutsumi Sasaki, Go Kamoda, Ryosuke Takahashi, Kosuke Sato, Kentaro Inui, Keisuke Sakaguchi, and Benjamin Heinzerling. Can language models handle a non-gregorian calendar? the case of the japanese wareki. *arXiv preprint arXiv:2509.04432*, 2025. URL https://arxiv.org/abs/2509.04432.

A DATASET COMPOSITION AND GROUND TRUTH

Our benchmark is designed to disentangle superficial factual recall from structural temporal reasoning. We curate a total of **304 samples** per calendar split across two primary tasks, ensuring that the ground-truth labels account for the complex astronomical rules inherent in non-Gregorian systems.

TASK 1: EVENT DATE RETRIEVAL

This task evaluates the model's ability to anchor historical facts directly within an indigenous timeline. By requiring the precise date (Day, Month, Year) in a cultural calendar, we test if the model has internalized the relationship between events and non-Gregorian cycles, rather than merely relying on Gregorian-to-offset heuristics.

- **Data:** 204 curated historical event entities (e.g., the *Dandi March* or the *Wuchang Uprising*).
- **Core Challenge:** Probing factual grounding and the model's capacity to handle regional variations in date nomenclature (e.g., Chinese sexagenary years).

TASK 2: DATE ARITHMETIC

This task serves as a "stress test" for the model's algorithmic understanding of a calendar's structural mechanics. Models are provided with a starting cultural date and a day-offset Δ, requiring them to compute the final date D_{target}.

- **Data:** 100 computationally verified samples involving significant temporal jumps.

- **Core Challenge:** Successfully modeling non-linear transitions, such as the insertion of intercalary months (*Adhik Maas* or *Runyue*) and the fluctuating month lengths of purely lunar systems.

GROUND-TRUTH GENERATION AND TOOLING

To ensure high-fidelity labels, we leveraged a suite of specialized astronomical converters to map Gregorian dates to their cultural counterparts. These tools were chosen for their adherence to regional standards (e.g., the Purnimanta system for Vikram Samvat):

- **Vikram Samvat/Nanakshahi:** `https://nanakshahi.net/convert/`
- **Chinese Lunar:** `https://www.prokerala.com/general/calendar/chinese-year-converter.php`
- **Hebrew:** `https://www.hebcal.com/converter`

Figures 3 and 4 demonstrate the JSON-formatted structure and natural language prompts used during evaluation.

```
{
  "Task": "Event Retrieval",
  "Calendar": "Hebrew",
  "Question": "What did the Atomic bombing of Hiroshima
  happen according to the Hebrew Calendar?",
  "Ground Truth": "27th of Av, 5705"
}
```

Figure 3: Example data structure for Task 1: Event Retrieval.

```
{
  "Task": "Date Arithmetic",
  "Calendar": "Vikram Samvat",
  "Question": "What was the exact date
              22 days before 1 June 2024
              according to the Hindi Vikram Samvat Calendar?
              ",
  "Ground Truth": "28 Vaisakh, 2081"
}
```

Figure 4: Example data structure for Task 2: Date Arithmetic.

B TEMPORAL DISTRIBUTION OF EVENTS

To ensure the benchmark covers a broad historical range while remaining grounded in the models' likely pre-training knowledge, we curated the dataset with historical events with a distribution shown in Figure 6. The events span nearly a millennium, ranging from the late 11th century to the present day.

- **Historical Depth:** The dataset includes sparse but significant events from the medieval period (1100–1400 CE) to test the models' retrieval of deep historical facts in cultural calendars.
- **Early Modern and Colonial Era:** There is a steady increase in event density between 1500 and 1800 CE, coinciding with major global shifts, colonial history, and documented regional transitions in the target calendar regions.
- **Modern Bias:** A significant concentration of events is observed in the 19th and 20th centuries, peaking between 1900 and 2000 CE. This modern bias is intentional, as it targets the era of highest data density in LLM training corpora, thereby minimizing failures due to simple lack of factual knowledge and highlighting failures in calendar-specific reasoning.

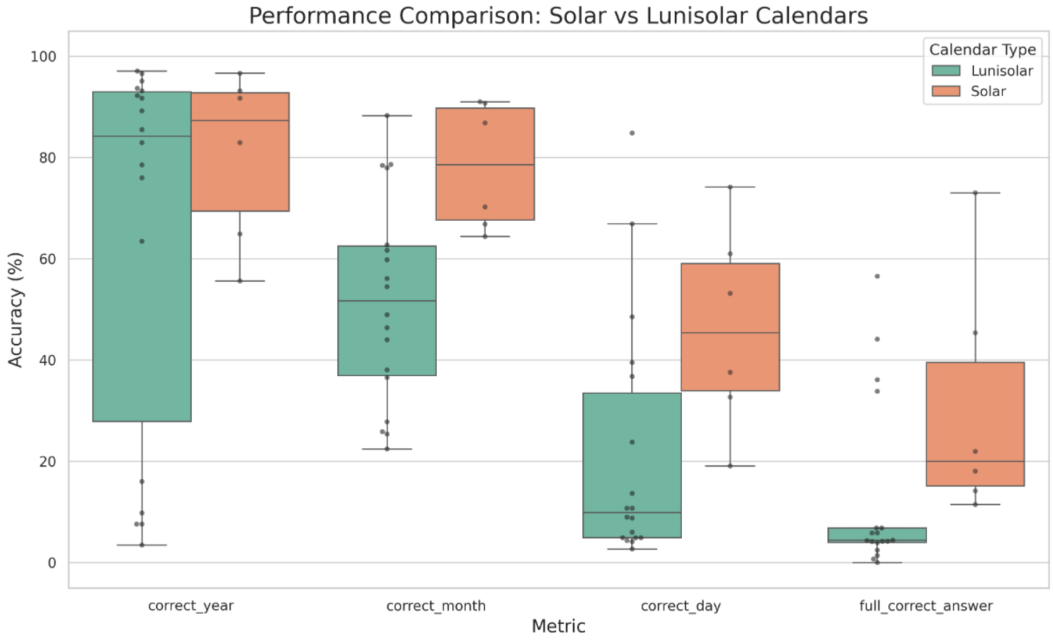

Figure 5: Comparison of Lunisolar v/s Solar Calendars

C Ablation Study: Mechanistic Complexity and Granularity

To isolate the root causes of the Gregorian bias, we perform an ablation analysis focusing on the mathematical foundation of the calendar systems (Solar vs. Lunisolar) and the precision level of the reasoning task.

C.1 Impact of Calendar Type: Solar vs. Lunisolar

We categorize the benchmarked systems into Solar (Persian) and Lunisolar or Lunar (Vikram Samvat, Hebrew, Chinese, Hijri) to observe how structural mechanics influence model accuracy.

- **Solar Resilience:** Models demonstrate higher median accuracy and lower variance in Solar systems across all metrics. This is likely due to the deterministic nature of tracking the vernal equinox.
- **Lunisolar Fragility:** Performance in Lunisolar systems exhibits higher variance at the year level and a sharper decline in month/day accuracy. This reflects the failure to model non-linear rules like *Adhik Maas* (intercalary months) or the Metonic cycle.

C.2 Granularity Decay Analysis

The ablation across temporal granularities, Year, Month, and Day, exposes the limits of surface-level memorization.

- **Year-Level Anchoring:** Median accuracy for `correct_year` remains high ($> 80\%$) for most systems. Models appear to utilize linear offsets well-represented in training weights.
- **Precision Collapse:** There is a precipitous drop-off as granularity increases. While Solar models maintain a median accuracy of $\sim 45\%$ for `correct_day`, Lunisolar models collapse to $\sim 10\%$.
- **Logical Ceiling:** The `full_correct_answer` metric confirms a systemic floor; models lack the structural logic to calculate exact days within non-standard lunar cycles.

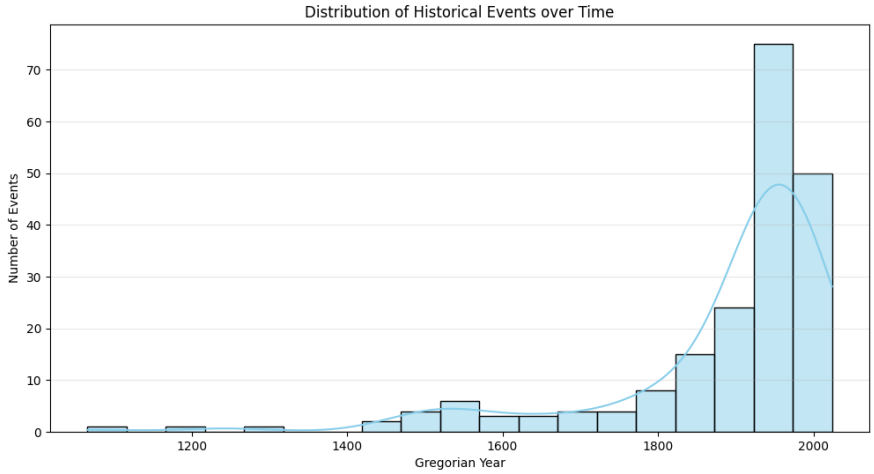

Figure 6: Distribution of the Dates over Time

D Data Standardization and Annotation

D.1 Standardization and Disambiguation

Managing regional and astronomical variations in non-Gregorian systems is critical for ground-truth integrity. Because cultural calendars often vary by geography or sect, we standardized our labels using the following rigorous mathematical and astronomical conventions:

- **Vikram Samvat:** We adopted the **Purnimanta** system (where the month ends on the full moon), which is prevalent in Northern India. For the New Year, we followed the **Chaitra Sukladi** convention.
- **Hijri:** We prioritized the **Tabular Islamic Calendar** (arithmetic calculation) over local lunar sightings (*Ru'yat*) to ensure mathematical reproducibility. This eliminates the ± 1-day variance inherent in physical sightings, providing a stable ground truth for algorithmic evaluation.
- **Chinese Lunar:** Calculations were standardized to the **China Standard Time (UTC+8)** zone. Intercalary months (*Runyue*) were determined using the "no-mid-term" rule (a lunar month lacking a *Zhongqi* solar term).
- **Hebrew:** We followed the standard fixed arithmetic of the **Hillel II** system, which utilizes the 19-year Metonic cycle to reconcile lunar months with the solar year.

D.2 Annotation and Verification Protocol

To ensure the highest level of linguistic and cultural nuance, the annotation and verification process was conducted by the **authors themselves**.

E Ablation Study: Which is harder? Adding or Subtracting Days

To move beyond aggregate scores, we disentangle model performance across three temporal granularities, **Year**, **Month**, and **Day**, and evaluate the *Full Correct* exact match (EM) rate. The following heatmaps illustrate performance across both **Addition** $(+)$ and **Subtraction** $(-)$ tasks for all evaluated models and calendar systems.

The visualization of these metrics reveals several systemic failure modes, with a particular emphasis on the performance divergence between **Addition** and **Subtraction** tasks:

- **Directional Asymmetry (The Subtraction Penalty):** Models consistently demonstrate a "Subtraction Penalty," where accuracy drops when calculating dates in the past compared

to the future. As seen in Figure 7, models like *DeepSeek-V3* and *GPT-OSS-120B* show a marked decrease in day-level accuracy for Subtraction tasks in the Hijri and Persian systems, likely due to the added cognitive load of calculating alternating month lengths and leap years in reverse.

- **Arithmetic vs. Retrieval in Solar Systems:** In the Persian (Jalali) calendar, *DeepSeek-R1* maintains relatively stable performance across both Addition and Subtraction. However, other models show higher variance in Subtraction, suggesting they rely on forward-facing heuristics that fail when the temporal logic requires regressive computation.
- **Lunisolar Fragility in Reverse:** The collapse in **Subtraction** is most acute in lunisolar systems like *Vikram Samvat* and *Hebrew*. In Figure 8, month-level accuracy for Subtraction in the Hebrew calendar is significantly lower than Addition across almost all models. This indicates that the complex rules governing the 19-year Metonic cycle and intercalary months are not structurally internalized, making backward traversal algorithmically inaccessible for the models.
- **The Chinese Anomaly across Operations:** The Chinese calendar shows a near-total collapse in both Addition and Subtraction. Figure 9 reveals that even at the year level, models cannot handle either operation, confirming that the sexagenary cycle and lunar-based New Year shifts represent a fundamental bottleneck that simple linear offsets cannot solve.
- **Precision Decay across Operators:** As illustrated by the transition from Figure 9 to 10, the collapse is present in both operations but is exacerbated in Subtraction. While models may retrieve a correct year for a past event, they fail significantly more often to identify the specific day (Subtraction) than to calculate a future date (Addition), highlighting a lack of bidirectional temporal logic.

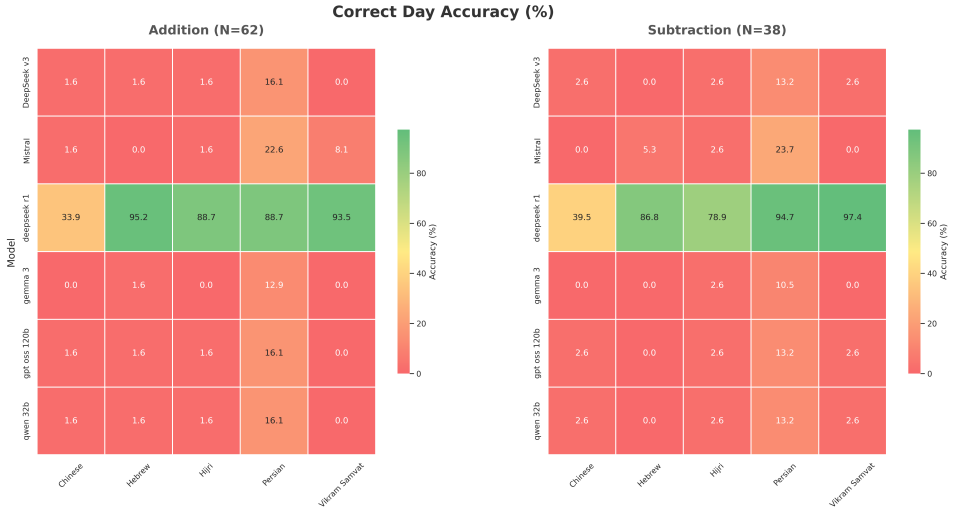

Figure 7: Heatmap of **Correct Day Accuracy**. This metric represents the model's ability to identify the correct numerical day within the target calendar. Note the significant "red zone" for lunisolar systems like Vikram Samvat and Chinese Lunar.

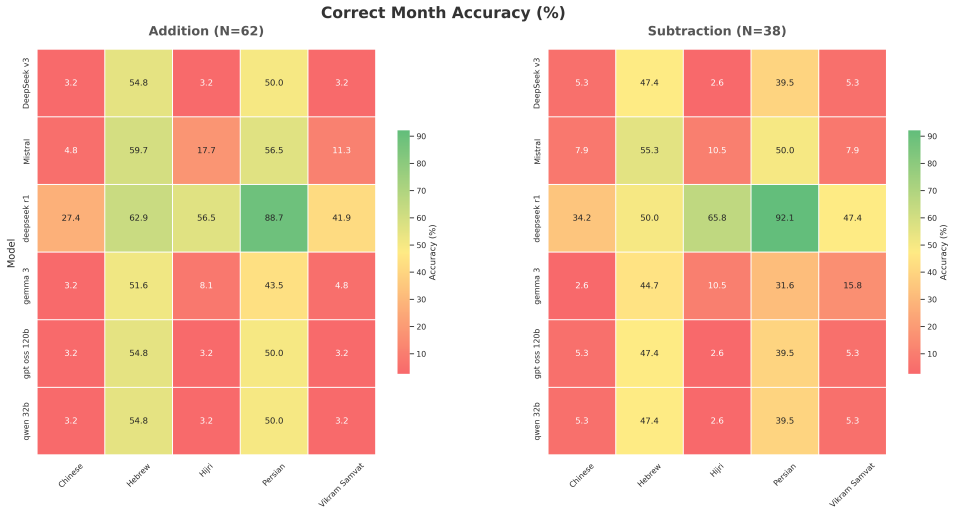

Figure 8: Heatmap of **Correct Month Accuracy**. Performance remains moderate for solar systems but shows a steep decline for the Hijri and Hebrew calendars due to the fluctuating nature of lunar months.

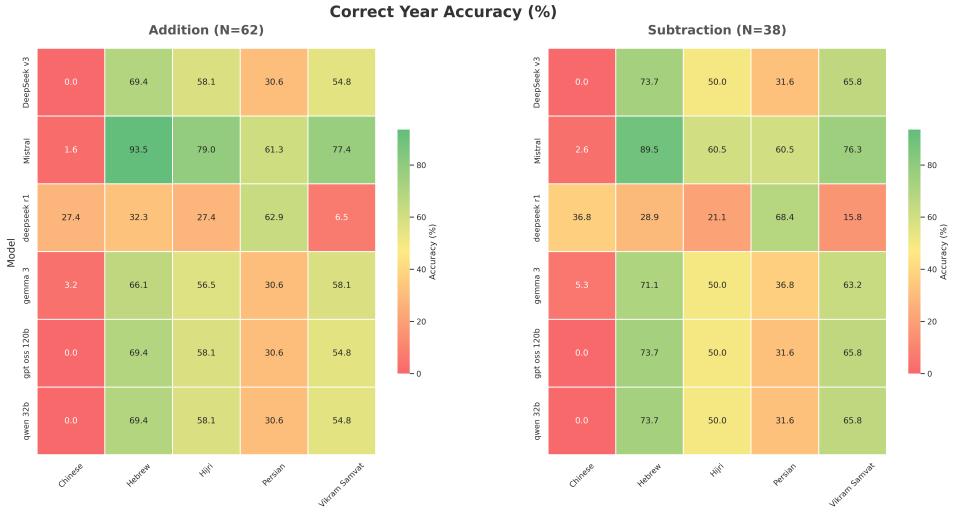

Figure 9: Heatmap of **Correct Year Accuracy**. This represents the highest level of model resilience, indicating that models often successfully retrieve or calculate fixed year-level offsets (e.g., $+57$ for Vikram Samvat).

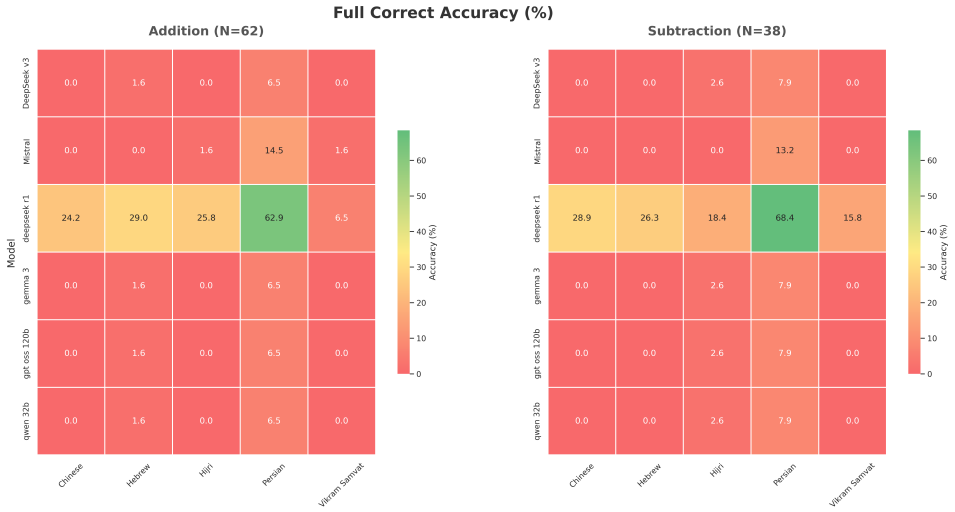

Figure 10: Heatmap of **Full Correct Date Accuracy**. This final metric requires an exact match for Day, Month, and Year simultaneously. The near-total collapse across most models highlights the lack of integrated temporal logic.