Benchmarking Multilingual Temporal Reasoning in LLMs: The Temporal Reasoning Dataset

Vittorio Mazzia **Sandro Pollastrini** **Davide Bernardi**

Chiara Rubagotti **Daniele Amberti**

Amazon Alexa AI*

Correspondence: {vmazzia,kvantumo,dvdbe}@amazon.com

Abstract

Time reasoning is a make-or-break capability for Large Language Models (LLMs) aspiring to act as reliable personal and enterprise assistants. This paper introduces the Temporal Reasoning Dataset (TRD), a programmatically generated multilingual benchmark designed to evaluate temporal reasoning operational capabilities in LLMs across ten languages, with particular focus on basic operations relevant to conversational agents handling time-sensitive tasks. TRD utilizes human-curated carrier phrases to generate a resilient-to-overfitting dataset with diverse samples and controlled difficulty levels across five core task categories, each at five difficulty levels. Extensive experimentation shows consistent patterns in model performance across languages, with a strong linear decline in accuracy as task difficulty rises in reasoning-based tasks, while memorization-based tasks remain stable. Furthermore, reasoning tasks remain robust across temporal shifts, whereas memorization tasks show performance degradation. Additionally, contextual modifications to prompts influence model performance differently than human cognitive patterns.

1 Introduction

Time is a fundamental aspect of human cognition, allowing us to flexibly navigate between past, present, and future events. Similarly, temporal reasoning is an essential capability for large language models (LLMs), enabling them to effectively handle diverse tasks and interactions that rely on understanding and managing time-related information. From scheduling meetings and calculating durations to interpreting historical contexts, robust temporal reasoning is crucial for LLMs to serve effectively as personal and enterprise assistants.

This paper introduces the Temporal Reasoning Dataset (TRD), a novel multilingual benchmark specifically designed to evaluate operational temporal reasoning capabilities in LLMs across ten distinct languages spanning multiple language families. Leveraging human-curated carrier phrases, TRD programmatically generates extensive and diverse datasets that are resilient to data overfitting, allowing detailed analysis of temporal reasoning across various tasks and difficulty levels.

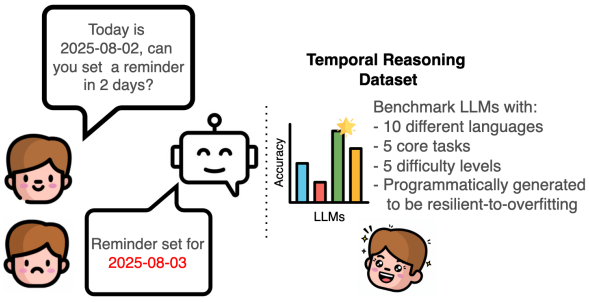

Figure 1: The Temporal Reasoning Dataset is a programmatically generated, multilingual benchmark designed to evaluate LLMs on practical scenarios involving time-sensitive tools and operations—capabilities essential for conversational agents.

In this study, we benchmarked each model using 104,000 generated samples distributed across five core temporal tasks. These tasks are grouped into three reasoning tasks: (1) arithmetic involving time and dates, (2) duration comprehension, (3) recurrence understanding; and two memorization tasks: (4) interval interpretation and (5) day-of-week recognition. Each task is carefully structured with five incremental levels of complexity to systematically assess model performance and adaptability (see Figure 1 for an overview). The multilingual design of our benchmark allows evaluation of temporal reasoning across diverse linguistic contexts. While this work primarily focuses on introducing the dataset and its initial findings, rather than providing an exhaustive analysis of LLMs' temporal capabilities, the evaluation across multiple languages offers valuable insights into how

* All authors were associated with Amazon at the time of publication.

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 168–181
February 26–March 1, 2026.

language structure influences temporal reasoning abilities and cross-lingual generalization.

Our benchmark facilitates rigorous experimentation and yields several key insights:

- Consistent patterns in base performance across languages, with models performing more robustly on Indo-European languages with Latin scripts compared to non-Latin scripts, and algorithmic reasoning abilities transferring more effectively across languages than memorized knowledge.
- A clear linear reduction in accuracy as the complexity of reasoning-based tasks increases, while memorization-based tasks demonstrate more stable performance patterns across difficulty levels. This contrasting behavior suggests that LLMs employ different strategies when handling temporal tasks that require explicit reasoning versus those that can be solved through pattern recognition or memorization.
- Temporal reasoning tasks remain stable across temporal shifts spanning a century, whereas memorization tasks exhibit significant performance degradation.
- Contextual additions to prompts negatively impact model performance in ways that differ from human cognition: while humans struggle more with topically related distractors, LLMs show varied sensitivity to both similar and dissimilar insertions, with performance drops varying unpredictably across models and task types.

Collectively, our dataset provides insights into temporal reasoning capabilities in multilingual LLMs, particularly highlighting the distinction between reasoning and memorization patterns. By focusing on practical cases relevant to conversational agents handling time-sensitive tools and operations, it offers a foundation for further investigation into making language models more reliable and effective in real-world applications. Furthermore, its programmatic nature makes it strongly resilient to data contamination and overfitting. All data used in this study (104,000 samples), along with the complete codebase for data generation and evaluation, is publicly available at[1].

[1] https://github.com/amazon-science

2 Related Works

Recent years have witnessed a growing interest in evaluating models' temporal understanding capabilities. A key line of research has focused on developing datasets explicitly designed to assess and enhance temporal reasoning in natural language processing (NLP) models. Thukral et al. (2021) and Hosokawa et al. (2023) created natural language inference datasets to probe pre-trained models' comprehension of common temporal expressions and concepts like event containment and state verification. To assess commonsense temporal reasoning, TimeDial (Qin et al. (2021)) and MC-TACO (Zhou et al. (2019)) were introduced, containing diverse situations and temporal expressions. Additionally, several recent question-answering datasets aim to evaluate temporal reasoning (Chen et al. (2021)). Furthermore, contemporary research explores time-aware training strategies and representations for language models (Wang et al. (2023); Cole et al. (2023); Kimura et al. (2021); Zhou et al. (2019); Kimura et al. (2021); Saxena et al. (2021)). The temporal knowledge graph completion domain has also investigated temporal reasoning within knowledge graphs (Dhingra et al. (2022); Jang et al. (2022)). Overall, there has been a notable expansion of temporal reasoning studies in natural language understanding (Wenzel and Jatowt (2023)).

While the proficiency of LLMs has been demonstrated across various tasks, their full capabilities and limitations remain unclear. Recent studies have benchmarked LLM performance in diverse scenarios and tasks. For instance, Asai et al. (2023) and Ahuja et al. (2023) conducted extensive evaluations of multiple LLMs on cross-lingual and multilingual tasks, respectively. Wadhwa et al. (2023) assessed two LLMs' capabilities on relation extraction tasks, while Yang et al. (2023) benchmarked ChatGPT in the context of mental health issues. Nay et al. (2024) comparatively analyzed ChatGPT and GPT-4 on legal tax problems. In summary, the latest research exhibits an increasing trend of probing LLM applications across various domains, languages, and tasks.

While recent research has extensively evaluated LLMs across various domains and languages, their temporal reasoning capabilities remain understudied, particularly in operational contexts. This gap is especially significant given the increasing deployment of LLMs as conversational agents and virtual assistants, where accurate temporal reason-

Table 1: Languages included in our benchmark, grouped by language family, branch, and ISO-639-1 codes.

Language Family	Branch	Languages (ISO-639-1)
Afro-Asiatic	–	Modern Standard Arabic (ar-SA)
Indo-European	Germanic	English (en-US), German (de-DE), Dutch (nl-NL)
	Romance	Spanish (es-ES), French (fr-FR), Italian (it-IT), Portuguese (pt-BR)
	Indo-Aryan	Hindi (hi-IN)
Japonic	–	Japanese (ja-JP)

ing is crucial for handling calendar-based tasks and time-sensitive operations. Our work addresses this need by introducing a programmatically generated benchmark that systematically evaluates both reasoning and memorization-based temporal capabilities across multiple languages. Through this study, we contribute to the growing body of LLM evaluation research by providing a resilient-to-overfitting framework specifically designed to assess how both open and closed-source LLMs handle practical temporal reasoning tasks in conversational contexts.

A complementary line of research studies LLMs on temporal data and time series. Chang et al. (2025) survey reasoning and agentic systems for *time-series* tasks, emphasizing evaluation designs and task topologies. Liu et al. (2025) introduce *Time-R1*, a 3B model trained with a staged RL curriculum and evaluated on *Time-Bench* for temporal understanding, future-event prediction, and creative scenario generation. Fons et al. (2024) propose a taxonomy and synthetic benchmark for LLM *time-series feature understanding*, analyzing sensitivity to formatting and sequence length. Potosnak et al. (2024) probe *implicit reasoning* in deep time-series forecasting via synthetic out-of-distribution composition, comparison, and inverse-search tasks. Our work is orthogonal and complementary: rather than numeric forecasting or open-world temporal prediction, we target *discrete calendar logic* with exact ground truth—date arithmetic, durations, recurrences, interval relations, and day-of-week.

3 Methodology

3.1 Overview

The TRD is a fully synthetic and highly configurable benchmark designed to evaluate time-related reasoning in LLMs. Because the dataset is programmatically generated, it allows precise control over task structure, temporal spans, difficulty levels, and linguistic diversity. This makes it especially suitable for analyzing model behavior under well-isolated experimental conditions.

To structure our evaluation, we first defined a baseline configuration corresponding to a medium level of temporal complexity (more info about configurations in Appendix A). This configuration includes moderate date ranges and recurrence values and serves as a practical midpoint across temporal reasoning tasks. From this anchor, we explored the dataset's flexibility through three experimental axes:

1. **Task Complexity Scaling:** We adjusted key temporal parameters, such as duration magnitudes and date offsets, to generate five increasing difficulty levels, from "short" to "very very long" timeframes (e.g., days/hours additions $\pm$, 1-4 days vs. 32-64). These ranged from simple near-term operations to complex, multi-step reasoning across long time horizons. This allowed us to observe how LLM performance evolves as temporal reasoning complexity increases in a multilingual setting.

2. **Reasoning vs. Memorization:** We divided tasks into two categories: those requiring internal computation (e.g., date arithmetic and recurrence) and those relying on factual calendar knowledge (e.g., day-of-week identification or interval boundaries). To further test memorization robustness, we evaluated model behavior on dates ranging from 2025 to 2095, extending beyond the likely boundaries of most pretraining corpora.

3. **Robustness to Contextual Noise:** We inserted additional context into prompts to simulate real-world distractions. Some insertions were topically related to time (e.g., "I always get confused with months that have 31 days"), while others were unrelated (e.g., "My brother

Table 2: English carrier phrases across core temporal reasoning tasks. Bolded sections are parametrized for temporal variations. TG indicates task group, that is how we group core tasks: Reasoning (R), and Memorization (M).

Core Task	TG	Object	Example
Temporal Arithmetic (±)	R	Date	Today is **2025-08-2**, what is the date going to be in **10** days?
		Time	It is now **19:29**, what will the time be in **1** hour and **26** minutes?
Duration	R	Date	If it was **2025-03-20** and now is **2025-03-29**, how many day(s) have passed?
		Time	If it was **13:32** and now is **14:21**, how much time has passed (minutes)?
Recurrence	R	Date	Today is **2025-02-23**, and I have a recurrence every **7** days. Without counting today, what is the date of the **2** occurrence?
Intervals	M	Date	If today is **2025-08-14**, what is the beginning and end of the following week?
Day of Weeks	M	Date	What day of the week (e.g., Monday, Tuesday, ...) is **2025-09-22**?

just bought a blue motorcycle"). This setup allowed us to assess how LLMs handle irrelevant or misleading context.

3.2 Carrier Phrase Design and Data Generation

At the core of TRD is a library of carrier phrases, which are sentence templates designed to represent various temporal reasoning tasks. These phrases were written and reviewed in collaboration with native speakers and language experts to ensure natural phrasing, grammatical correctness, and logical clarity. Each carrier phrase encodes a reusable template for a given reasoning task. For instance:

- *"Today is [DATE]. What is the date [X] days from now?"*
- *"It is now [TIME]. What time will it be in [Y] minutes?"*

Each template is then populated with dynamic parameters, including a reference date or time, a numerical offset or interval, and recurrence values (more info in Appendix A). The result is a set of semantically consistent, structurally diverse questions that span multiple levels of temporal reasoning. The combination of fixed logic and variable content makes TRD scalable while preserving interpretability and making it difficult to be overfitted.

TRD currently covers ten languages, selected to represent a wide range of linguistic families and structures. These include English, German, Spanish, French, Italian, Portuguese, Dutch, Hindi, Arabic, and Japanese (Table 1). Each language has its own dedicated set of carrier phrases adapted to its syntax and lexical conventions. This multilingual, parallel corpora setup enables direct comparison of LLM behavior across languages while maintaining high grammatical and semantic consistency.

The temporal tasks supported in TRD fall into five core categories: arithmetic, duration, recurrence, intervals, and day-of-week. These cover a spectrum of cognitive demands, from logic and arithmetic reasoning to factual retrieval. We group them into:

- **Reasoning tasks:** Temporal arithmetic, duration, and recurrence. These tasks require internal algorithmic computation.
- **Memorization tasks:** Day-of-week and interval reasoning. These depend more on static calendar knowledge.

Table 2 shows representative examples of each task and the cognitive abilities they engage.

Table 3: Evaluated model families and variants.

Model Family	Models
Anthropic	Claude 3.5 Sonnet v2
	Claude 3.5 Sonnet
	Claude 3.5 Haiku
	Claude 3 Haiku
Mistral	Mistral Large 2402
	Mixtral 8x7B
Cohere	Command R Plus
	Command R

3.3 Evaluation Setup

For this study, we generated a benchmark dataset comprising 104,000 multilingual samples. We evaluated eight models from three families (Table 3), selected to represent a diverse range of architectures and capabilities: frontier models (Claude 3.5 Sonnet), mid-tier options (Mistral Large, Command R Plus), and more efficient variants (Haiku, Mixtral 8x7B). Each model was evaluated on an identical

sample set spanning the full range of task categories introduced in Section 3.1: difficulty levels, reasoning vs. memorization, and contextual conditions. To ensure comparability and reproducibility:

- When applicable, models were instructed to output answers in strict, machine-readable formats (e.g., `YYYY-MM-DD`, `HH:MM`) following ISO 8601. This is not only convenient for the evaluation code, but it also allows staying close to a real-world scenario in which an LLM is used as an agent as part of an orchestration.

- Chain-of-thought (CoT), (Wei et al., 2022), prompting was disabled through in-context-learning instructions to measure raw temporal reasoning capacity without external scaffolding. The answer is practically produced within the layer of the models, leaving the exploration of CoT's impact on temporal reasoning performance for future research.

- Temperature was fixed at 0.0 to enforce deterministic behavior.

- Format compliance rates were consistently above 99% across all models when using the strict output instructions, with format errors counted as incorrect responses in our evaluation.

More information on the number of samples generated for each experiment and the prompts adopted can be found in Appendix B. Note that TRD is fully customizable, and these sample sizes are arbitrary and do not constrain future iterations.

4 Experimental Results and Discussion

4.1 Base Performance Across Languages

We begin by examining the baseline performance of models across all 10 languages using the medium difficulty level. Table 4 presents an overview of model performance, separated into reasoning and memorization tasks.

Table 4 reveals several important patterns. Claude 3.5 Sonnet v2 and Claude 3.5 Sonnet consistently achieve the highest accuracy across languages for both reasoning and memorization tasks, maintaining impressive performance even for non-Latin script languages. Models generally perform more consistently across Indo-European languages with Latin scripts than on Japanese (Japonic family), Arabic (Afro-Asiatic family), and Hindi (Indo-European but non-Latin script).

Languages with non-Latin scripts frequently show lower performance, especially for lower-capacity models, suggesting that script differences create additional challenges for temporal reasoning. For reasoning tasks, the performance gap between languages is typically smaller than for reasoning tasks, suggesting that algorithmic reasoning transfers more effectively across languages than memorized knowledge.

Larger models demonstrate more consistent performance across languages, suggesting that increased parameter count contributes to more robust cross-lingual capabilities. Occasionally, we observe performance anomalies, such as Claude 3.5 Haiku's exceptionally high performance on Hindi memorization tasks (0.980) compared to English (0.380), which may reflect specifics of the training data distribution. This anomaly could stem from overrepresentation of certain date-weekday associations in Hindi corpora, or differences in how temporal expressions are tokenized and encoded across scripts. Furthermore, this pattern holds only for memorization tasks, providing additional evidence that LLMs employ different strategies when solving memorization versus reasoning tasks.

Extended results across tasks and languages can be found in Table 6 - 7 of Appendix C.

4.2 Performance Across Difficulty Levels

Our experimental axis of model performance across the five difficulty levels reveals a very interesting pattern: a clear linear trend in accuracy reduction as task difficulty increases for reasoning-based tasks. This pattern holds consistently across all evaluated models and languages.

Figure 2 illustrates this trend across the average of all Indo-European languages. We observe a consistent linear decline in accuracy as we move from short to very-very-long temporal complexity (refer to Appendix A for configuration definitions).

The slope of this decline varies across models, with larger models generally showing a more gradual decline than smaller ones. Instead, for memorization tasks, such as day-of-the-week determination and intervals, there is an almost stable trend across difficulties.

This distinction between stable memorization performance and declining reasoning performance as complexity increases mirrors human cognitive

Table 4: Accuracy across languages and models for both reasoning and memorization tasks with medium difficulty. Each language is evaluated with 900 samples, one for each task introduced in Section 3.1.

Model	Task	Language									
		pt-BR	de-DE	es-ES	fr-FR	hi-IN	it-IT	ja-JP	nl-NL	ar-SA	en-US
claude-3-5-sonnet-v2	REASONING	0.974	**0.990**	0.986	0.964	**0.903**	0.970	**0.970**	0.977	0.967	0.949
claude-3-5-sonnet		**0.989**	0.989	**0.987**	**0.983**	0.903	**0.971**	0.967	**0.987**	**0.979**	**0.973**
mistral-large-2402		0.823	0.849	0.824	0.851	0.714	0.803	0.709	0.801	0.763	0.824
mixtral-8x7b		0.633	0.619	0.580	0.591	0.517	0.629	0.509	0.624	0.510	0.649
claude-3-5-haiku		0.836	0.841	0.823	0.830	0.813	0.840	0.674	0.829	0.840	0.851
claude-3-haiku		0.839	0.861	0.849	0.833	0.827	0.817	0.697	0.836	0.830	0.851
command-r-plus		0.811	0.769	0.804	0.767	0.754	0.807	0.731	0.739	0.761	0.771
command-r		0.609	0.626	0.621	0.600	0.579	0.639	0.561	0.594	0.569	0.613
claude-3-5-sonnet-v2	MEMORIZATION	0.965	0.925	**0.975**	**0.940**	0.955	0.965	0.965	0.970	0.830	0.775
claude-3-5-sonnet		**0.980**	**0.975**	0.960	**0.940**	0.955	**0.985**	**0.985**	**1.000**	**0.835**	0.925
mistral-large-2402		0.795	0.770	0.800	0.700	0.280	0.785	0.605	0.545	0.455	**0.990**
mixtral-8x7b		0.490	0.590	0.580	0.545	0.290	0.565	0.580	0.320	0.300	0.800
claude-3-5-haiku		0.905	0.775	0.935	0.900	**0.980**	0.885	0.910	0.960	0.750	0.380
claude-3-haiku		0.835	0.850	0.850	0.890	0.530	0.860	0.585	0.735	0.670	0.450
command-r-plus		0.460	0.420	0.430	0.445	0.310	0.425	0.460	0.480	0.480	0.600
command-r		0.410	0.425	0.365	0.380	0.260	0.330	0.235	0.425	0.270	0.335

patterns. While both humans and LLMs often struggle with increasingly complex temporal calculations, factual knowledge about calendars and time intervals tends to remain accessible regardless of temporal distance.

In Appendix D are reported all two remaining language families, showing the same patterns.

4.3 Temporal Stability and Memorization

To assess how LLMs handle temporal shifts, we tested models on dates spanning from 2025 to 2095 (maintaining all remaining medium-timeframe configurations), revealing a fascinating dichotomy between reasoning and memorization tasks. Figure 3 shows the result across models, but by aggregating per language, with error bars showing the standard deviation.

For reasoning tasks, such as calculating the duration between dates, performance remains remarkably stable across all temporal periods. Whether calculating the number of days between dates in 2025 or 2095, models show consistent accuracy levels. This stability suggests that the arithmetic operations underlying these tasks are well-learned and generalize effectively regardless of the specific years involved. Moreover, this is in accordance with what was observed in Section 4.2.

In contrast, memorization tasks show significant degradation for dates far from the training distribution. Performance on day-of-week and intervals determination drops dramatically for years beyond 2050 for most models. This pattern suggests that models are relying on memorized associations between dates and weekdays, rather than implementing algorithmic solutions like Zeller's congruence (Tonapi, 2023) that would generalize across any date. The notable performance drop observed around 2055 for memorization tasks likely represents a boundary effect where training data coverage diminishes significantly, as most web corpora contain fewer explicit references to dates beyond the mid-century.

4.4 Impact of Contextual Insertions

Our insertion experiments reveal how contextual additions affect model performance on temporal reasoning tasks. Figure 4 presents results in an aggregated form for all languages and difficulties.

Contrary to human cognition, where topically related distractors typically cause more interference than unrelated ones, LLMs show varied responses to both similar and dissimilar insertions. Some models show greater disruption from similar insertions, while others are more affected by dissimilar ones, and the pattern varies across different task types.

The task type significantly influences the impact of insertions. The day-of-the-week task shows

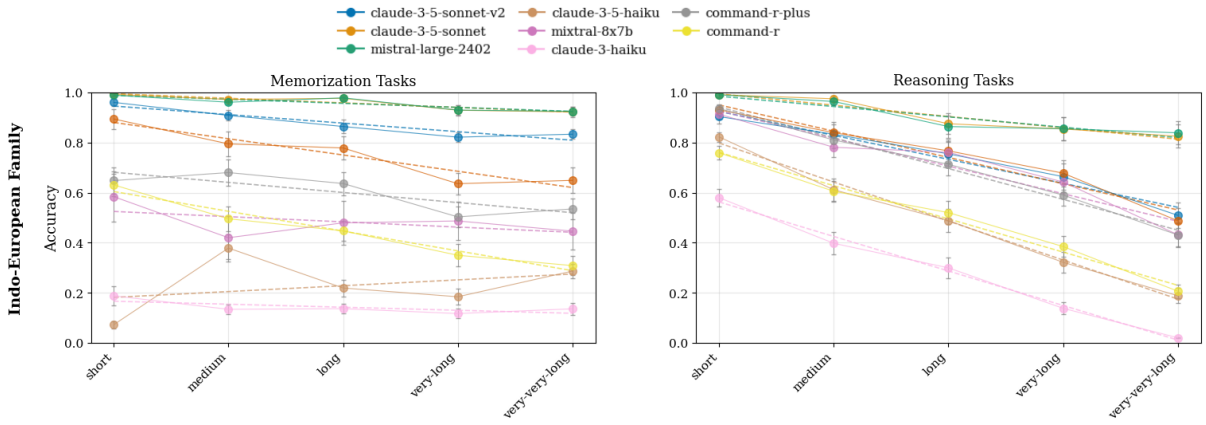

Figure 2: Accuracy by difficulty levels for Indo-European languages (including Hindi, which uses Devanagari script). The dotted line shows a linear regression of the distribution. Tasks are aggregated by average.

higher sensitivity to insertions across most models, with absolute performance drops of up to 10-22 percentage points when insertions are present. Simpler tasks like time addition often show more resilience to insertions.

As shown in Figure 4, the most capable models (e.g., Claude 3.5 Sonnet v2) generally show greater resilience to both types of insertions. This suggests that improved reasoning capabilities correlate with better contextual filtering, allowing these models to more effectively distinguish between relevant information and distractions.

Across all models, the average performance impact of insertions is small but consistently with a negative average. This consistent negative effect indicates that current LLMs lack robust mechanisms for filtering out irrelevant information, processing all input context together, and trying to give a meaning to each instruction in the prompt.

Results aggregated for family languages can be found in Appendix E.

5 Conclusion

This paper introduced the Temporal Reasoning Dataset (TRD), a large-scale, multilingual benchmark designed to assess how well LLMs understand and reason about time, with particular focus on operational scenarios relevant to conversational agents. By combining programmatically generated samples with human-curated linguistic structures, TRD provides a resilient-to-overfitting framework that enables detailed analysis of model behavior

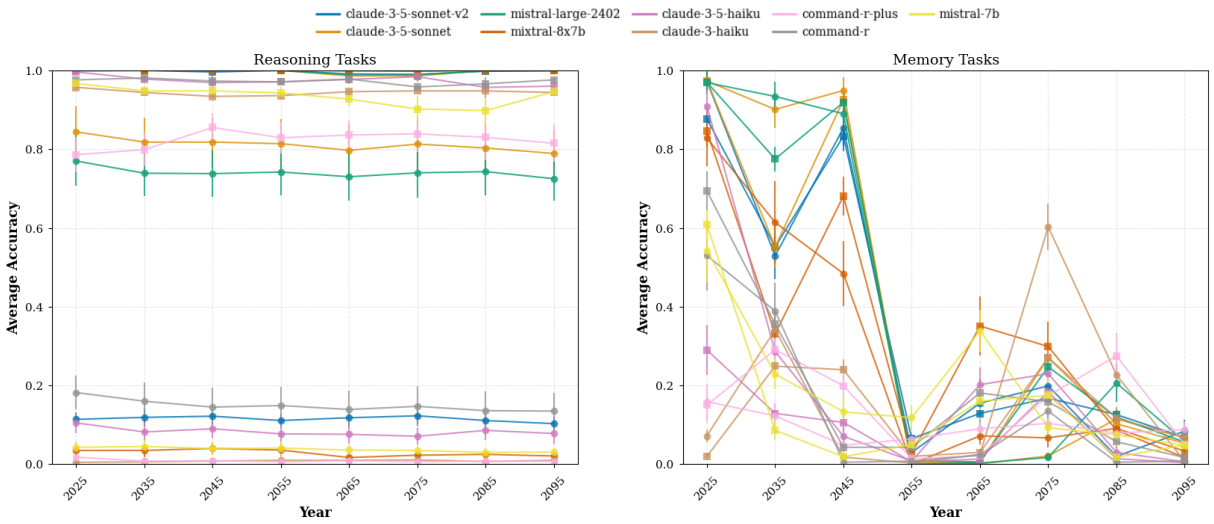

Figure 3: Reasoning and memory tasks performance across years. Languages are averaged together, and error bars show the standard deviation.

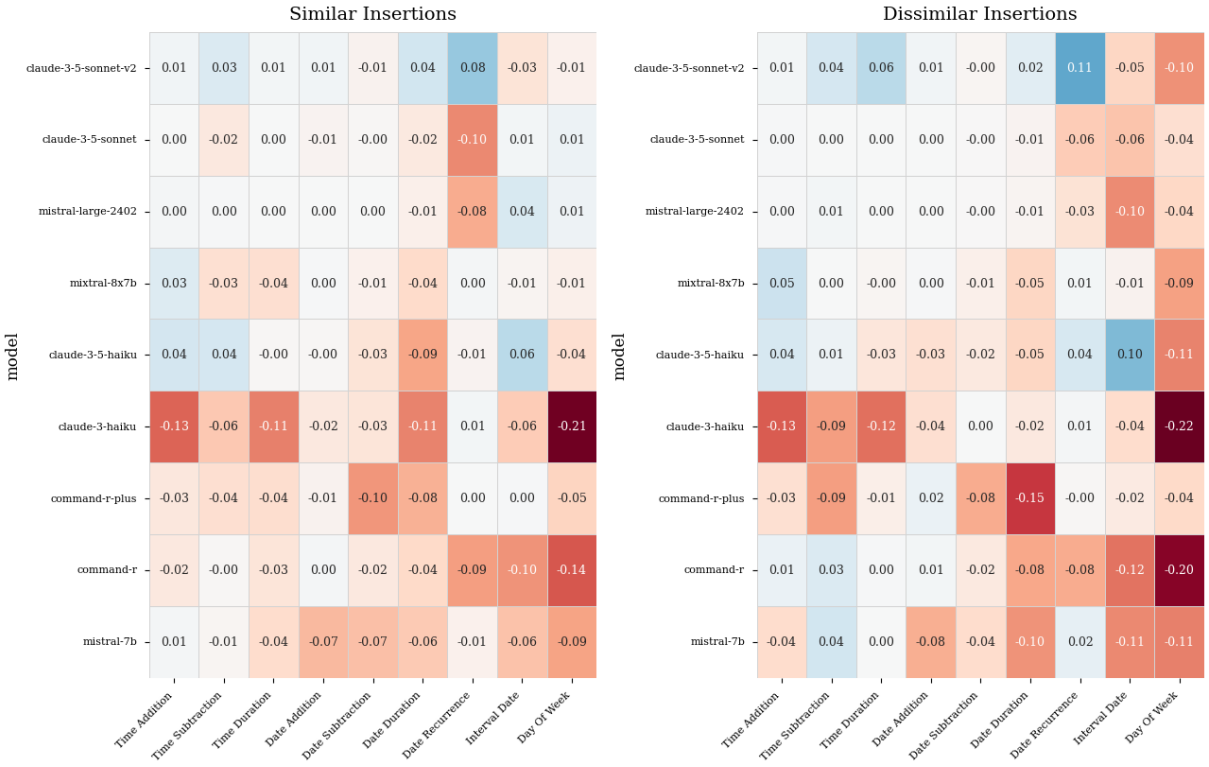

Figure 4: Impact of insertions on model performance, aggregating per language and difficulties. Each block reports the absolute difference between the baseline evaluation without insertions.

across a variety of temporal tasks, languages, and difficulty levels.

The results reveal clear and consistent patterns. Temporal reasoning tasks show a gradual but steady decline in accuracy as complexity increases, suggesting that models struggle increasingly with multi-step or abstract time-related computations. However, these same tasks tend to remain stable even when applied to dates far outside the training distribution, indicating that arithmetic-based reasoning generalizes well. Memorization tasks, on the other hand, appear more brittle. Their accuracy often depends on familiar data ranges, and performance drops noticeably when models are tested on unfamiliar or far-future dates.

Contextual distractions, whether relevant to the task or entirely unrelated, reduce on average model performance. This indicates a tendency in current models to process all input equally, without effectively filtering out irrelevant information. While the impact of such insertions varies across models and task types, it remains a challenge across the board.

The multilingual aspect of this study highlights how linguistic diversity shapes model performance. Models perform more consistently on languages with Latin scripts and tend to struggle more with non-Latin scripts, especially in tasks that depend on memorized calendar knowledge. Larger models generally demonstrate more robust and uniform behavior across languages.

Together, these findings emphasize the importance of temporal reasoning as a core capability for LLMs and point to current limitations in both generalization and contextual understanding. The distinct patterns observed between reasoning and memorization tasks, along with the impact of contextual modifications, provide valuable insights for developing more reliable conversational agents capable of handling time-sensitive operations.

Future work will explore the impact of Chain-of-Thought prompting on temporal reasoning performance, extend the benchmark to additional languages and task categories, and investigate the integration of TRD evaluations within deployed conversational systems.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed

Axmed, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*.

Ching Chang, Yidan Shi, Defu Cao, Wei Yang, Jeehyun Hwang, Haixin Wang, Jiacheng Pang, Wei Wang, Yan Liu, Wen-Chih Peng, and Tien-Fu Chen. 2025. A survey of reasoning and agentic systems in time series with large language models. *arXiv preprint arXiv:2509.11575*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *arXiv preprint arXiv:2108.06314*.

Jeremy R Cole, Aditi Chaudhary, Bhuwan Dhingra, and Partha Talukdar. 2023. Salient span masking for temporal understanding. *arXiv preprint arXiv:2303.12860*.

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.

Elizabeth Fons, Rachneet Kaur, Soham Palande, Zhen Zeng, Tucker Balch, Manuela Veloso, and Svitlana Vyetrenko. 2024. Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark. *arXiv preprint arXiv:2404.16563*. Accepted to EMNLP 2024.

Taishi Hosokawa, Adam Jatowt, and Kazunari Sugiyama. 2023. Temporal natural language inference: Evidence-based evaluation of temporal text validity. In *European Conference on Information Retrieval*, pages 441–458. Springer.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. *arXiv preprint arXiv:2204.14211*.

Mayuko Kimura, Lis Kanashiro Pereira, and Ichiro Kobayashi. 2021. Towards a language model for temporal commonsense reasoning. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 78–84.

Zijia Liu, Peixuan Han, Haofei Yu, Haoru Li, and Jiaxuan You. 2025. Time-r1: Towards comprehensive temporal reasoning in llms. *arXiv preprint arXiv:2505.13508*.

John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. 2024. Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A*, 382(2270):20230159.

Willa Potosnak, Cristian Challu, Mononito Goswami, Michal Wilinski, Nina Zukowska, and Artur Dubrawski. 2024. Implicit reasoning in deep time series forecasting. *arXiv preprint arXiv:2409.10840*. NeurIPS 2024 Workshop: Time Series in the Age of Large Models.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. Timedial: Temporal commonsense reasoning in dialog. *arXiv preprint arXiv:2106.04571*.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. *arXiv preprint arXiv:2106.01515*.

Shivin Thukral, Kunal Kukreja, and Christian Kavouras. 2021. Probing language models for understanding of temporal expressions. *arXiv preprint arXiv:2110.01113*.

Anushka Tonapi. 2023. Zeller's congruence. *At Right Angles*, pages 68–71.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566. NIH Public Access.

Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa, and Yi Cai. 2023. Bitimebert: Extending pre-trained language representations with bi-temporal information. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 812–821.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Georg Wenzel and Adam Jatowt. 2023. An overview of temporal commonsense reasoning and acquisition. *arXiv preprint arXiv:2308.00002*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. " going on a vacation" takes longer than" going for a walk": A study of temporal commonsense understanding. *arXiv preprint arXiv:1909.03065*.

A Difficulty Configurations

To systematically evaluate how model performance scales with temporal complexity, we defined five levels of difficulty based on the span and granularity of temporal values. These configurations control the range of date and time shifts, recurrence intervals, and duration calculations across tasks. The medium configuration was used as the baseline setting in our main evaluation. Table 5 summarizes the full set of difficulty levels used throughout the benchmark.

B Additional Experiment Details

Each model tested was evaluated on 104,000 samples divided as follows:

- **Variations**: 9 tasks * 100 samples * 10 languages * 1 difficulty = 9000
- **Difficulties**: 9 tasks * 100 samples * 10 languages * 5 difficulty = 45,000
- **Insertions**: 9 tasks * 100 samples * 10 languages * 1 difficulty * 2 variations = 18,000
- **Memorization**: 4 tasks * 100 samples * 10 languages * 8 years = 32,000

However, it is important to note that TRD is fully customizable, and the number of samples used in these experiments was chosen arbitrarily and does not constrain future iterations.

On the other hand, Figure 5 illustrates the template used for all experiments with TRD. The evaluated model is conditioned to provide answers in a machine-readable format without additional explanations. While this method avoids explanations and model thoughts, which could increase false positives by using exact match as the evaluation criterion, it also prevents the application of reasoning schemas like Chain of Thought, (Wei et al., 2022), that might improve performance. Exploring these schemas is a topic for future research.

C Base Performance Across Languages: Results per Task

In this appendix, we provide in Tables 6 and 7 results per task across languages and models. These tables present the detailed breakdown of model performance for each temporal reasoning task in all ten languages evaluated in our benchmark.

D Performance Across Difficulty Levels: Additional Languages

In this appendix, we provide in Figure 6 the performance results across difficulty levels for the remaining language families not covered in the main text. These visualizations demonstrate how model accuracy changes with increasing task difficulty across our multilingual dataset, showing similar patterns of linear performance degradation observed in the primary Indo-European languages.

E Impact of Contextual Insertions: Results Across Language Families

In this appendix, we provide detailed results showing how contextual insertions affect model performance across different language families. Figures 7, 8, and 9 illustrate the performance impact of both similar (time-related) and dissimilar (unrelated) insertions on Indo-European, Afro-Asiatic, and Japonic language families respectively.

Table 5: Difficulty configurations by temporal span and recurrence parameters. Medium was used as baseline.

Timeframe	**Start**	**End**	**Days (±)**	**Hours (±)**	**Every**	**Q**	**Date Δ**	**Time Δ**
Short	2025-01-01	2025-12-31	1–4	1–4	1–4	1–2	1–4	1–60
Medium	2025-01-01	2028-12-31	4–8	4–8	4–8	2–4	4–8	60–120
Long	2025-01-01	2030-12-31	8–16	8–16	8–16	4–8	8–16	120–240
Very Long	2025-01-01	2033-12-31	16–32	16–32	16–32	8–16	16–32	240–480
Very Very Long	2025-01-01	2036-12-31	32–64	32–64	32–64	16–32	32–64	480–960

```
"""Answer the following ‘Question‘ and provide an answer after the ‘Answer‘ keyword.
    When needed, use machine format, YYYY-MM-DD or HH:MM. Do not add further
   explanations or comments.
Question: question Answer: """
```

Figure 5: Template used to condition evaluated models with the temporal reasoning QA dataset.

Table 6: Accuracy across languages and tasks for Antrhopic LLMs.

Model	Task	Language									
		pt-BR	de-DE	es-ES	fr-FR	hi-IN	it-IT	ja-JP	nl-NL	ar-SA	en-US
claude-3-5-haiku	date_addition	0.980	**0.990**	0.980	0.980	**0.990**	**0.990**	**0.990**	**0.990**	0.980	**0.990**
	date_duration	0.990	0.960	0.990	0.980	0.960	0.990	0.280	0.970	**1.000**	**1.000**
	date_recurrence	0.130	0.170	0.050	0.080	0.000	**0.210**	0.130	0.110	0.150	0.100
	date_subtraction	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	day_of_week	0.970	0.750	0.910	0.990	0.990	0.850	0.980	**1.000**	0.680	**1.000**
	interval_date	0.840	0.800	0.960	0.810	**0.970**	0.920	0.840	0.920	0.820	0.850
	time_addition	**0.960**	0.940	0.950	0.930	0.940	0.860	0.850	0.900	0.920	0.950
	time_duration	0.960	0.970	0.970	0.960	0.940	0.950	0.540	0.960	0.970	**0.990**
	time_subtraction	0.830	0.860	0.820	0.880	0.860	0.880	**0.930**	0.870	0.860	**0.930**
claude-3-5-sonnet	date_addition	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	date_duration	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	date_recurrence	**0.930**	0.920	0.910	0.890	0.360	0.820	0.800	0.910	0.880	0.820
	date_subtraction	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	day_of_week	**1.000**	**1.000**	0.990	0.990	0.940	0.980	0.980	**1.000**	0.700	**1.000**
	interval_date	0.960	0.950	0.930	0.890	0.970	0.990	0.990	**1.000**	0.970	0.980
	time_addition	0.990	**1.000**	**1.000**	**1.000**	0.970	0.980	0.980	**1.000**	0.970	**1.000**
	time_duration	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	time_subtraction	**1.000**	**1.000**	**1.000**	0.990	0.990	**1.000**	0.990	**1.000**	**1.000**	0.990
claude-3-5-sonnet-v2	date_addition	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	date_duration	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.970	**1.000**	**1.000**	**1.000**
	date_recurrence	0.820	**0.930**	0.900	0.790	0.360	0.800	0.850	0.840	0.800	0.660
	date_subtraction	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	day_of_week	**1.000**	0.980	**1.000**	0.990	0.950	0.990	0.980	**1.000**	0.700	**1.000**
	interval_date	0.930	0.870	0.950	0.890	0.960	0.940	0.950	0.940	0.960	**0.970**
	time_addition	**1.000**	**1.000**	**1.000**	0.980	0.980	**1.000**	0.990	**1.000**	0.970	0.990
	time_duration	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	time_subtraction	**1.000**	**1.000**	**1.000**	0.980	0.980	0.990	0.980	**1.000**	**1.000**	0.990
claude-3-haiku	date_addition	**1.000**	0.990	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.990	**1.000**	0.990
	date_duration	**1.000**	0.940	0.980	0.990	0.960	0.980	0.010	0.980	0.950	0.990
	date_recurrence	0.020	**0.250**	0.090	0.020	0.000	0.010	0.110	0.090	0.030	0.090
	date_subtraction	**1.000**	0.990	**1.000**	0.990	0.990	0.990	0.980	**1.000**	0.990	**1.000**
	day_of_week	0.990	0.960	**1.000**	0.980	0.330	0.990	0.330	0.980	0.700	**1.000**
	interval_date	0.680	0.740	0.700	0.800	0.730	0.730	**0.840**	0.490	0.640	0.600
	time_addition	0.930	0.920	0.940	0.930	0.940	0.900	0.880	0.940	**0.950**	0.920
	time_duration	0.980	0.970	0.960	0.950	0.970	0.950	0.980	0.930	0.950	**0.990**
	time_subtraction	0.940	0.970	0.970	0.950	0.930	0.890	0.920	0.920	0.940	**0.980**

Table 7: Accuracy across languages and tasks for Mistral and Cohere models.

Model	Task	Language									
		pt-BR	de-DE	es-ES	fr-FR	hi-IN	it-IT	ja-JP	nl-NL	ar-SA	en-US
mistral-7b	date_addition	0.840	0.830	**0.910**	**0.910**	0.440	0.830	0.890	0.840	0.790	0.900
	date_duration	0.350	0.320	0.570	0.510	0.440	0.380	0.400	0.490	0.370	**0.620**
	date_recurrence	0.000	0.000	0.000	0.000	0.000	0.000	**0.110**	0.000	0.010	0.000
	date_subtraction	0.870	0.870	0.860	0.890	0.820	0.850	**0.940**	0.860	0.870	0.930
	day_of_week	0.060	0.160	0.080	0.160	0.100	0.090	0.000	0.070	0.100	**0.340**
	interval_date	0.130	0.140	**0.240**	0.090	0.020	0.180	0.110	0.120	0.040	0.180
	time_addition	0.400	0.490	0.370	0.380	0.120	0.200	0.340	0.310	0.270	**0.510**
	time_duration	0.150	0.120	0.120	0.100	0.050	0.090	0.060	0.110	0.050	**0.160**
	time_subtraction	0.250	0.150	**0.300**	0.150	0.170	0.160	0.200	0.030	0.140	0.270
mistral-large-2402	date_addition	0.990	**1.000**	**1.000**	0.990	0.990	0.990	0.960	0.990	0.990	0.990
	date_duration	0.940	0.950	0.960	**0.970**	0.770	**0.970**	0.350	0.930	0.940	0.960
	date_recurrence	0.320	0.450	0.240	**0.480**	0.000	0.090	0.190	0.340	0.070	0.250
	date_subtraction	**1.000**	0.990	**1.000**	0.990	**1.000**	**1.000**	0.990	0.970	**1.000**	**1.000**
	day_of_week	0.760	0.720	0.780	0.730	0.250	0.750	0.610	0.270	0.170	**0.860**
	interval_date	**0.830**	0.820	0.820	0.670	0.310	0.820	0.600	0.820	0.740	0.690
	time_addition	0.900	**0.950**	0.900	0.910	0.830	0.940	0.890	**0.950**	0.850	0.940
	time_duration	0.720	0.660	**0.760**	0.700	0.670	0.730	0.680	0.610	0.680	0.680
	time_subtraction	0.890	0.940	0.910	0.920	0.740	0.900	0.900	0.820	0.810	**0.950**
mixtral-8x7b	date_addition	**1.000**	0.960	0.970	0.980	0.860	0.970	0.970	**1.000**	0.960	0.990
	date_duration	0.750	0.720	0.600	0.590	0.600	**0.880**	0.620	0.780	0.670	0.640
	date_recurrence	0.030	**0.100**	0.070	0.000	0.000	0.000	0.010	0.020	0.000	0.090
	date_subtraction	0.910	0.920	0.950	0.920	0.890	0.910	0.770	0.950	0.600	**0.970**
	day_of_week	0.460	0.700	0.570	**0.780**	0.110	0.750	0.570	0.480	0.110	0.720
	interval_date	0.520	0.480	**0.590**	0.310	0.470	0.380	**0.590**	0.160	0.490	0.470
	time_addition	0.710	0.710	0.630	0.750	0.590	0.660	0.500	**0.760**	0.590	**0.760**
	time_duration	**0.400**	0.380	0.270	0.360	0.270	0.320	0.140	0.360	0.330	**0.400**
	time_subtraction	0.630	0.540	0.570	0.540	0.410	0.660	0.550	0.500	0.420	**0.690**
command-r	date_addition	**0.990**	0.970	**0.990**	0.970	0.950	0.980	0.960	0.980	**0.990**	**0.990**
	date_duration	0.490	0.620	0.380	0.400	0.440	**0.650**	0.390	0.510	0.230	0.550
	date_recurrence	0.000	0.000	0.000	0.000	0.000	0.000	**0.030**	0.000	0.000	0.000
	date_subtraction	0.930	0.920	0.960	0.940	0.930	0.950	0.930	0.940	**0.970**	0.940
	day_of_week	0.610	0.650	0.530	0.640	0.310	0.470	0.330	0.640	0.380	**0.730**
	interval_date	**0.210**	0.200	0.200	0.120	**0.210**	0.190	0.140	**0.210**	0.160	0.170
	time_addition	0.760	0.810	**0.850**	0.760	0.740	0.810	0.770	0.790	0.750	0.780
	time_duration	0.300	0.300	0.320	0.370	0.260	0.300	0.070	**0.410**	0.240	0.170
	time_subtraction	0.790	0.760	0.850	0.760	0.730	0.780	0.780	0.530	0.800	**0.860**
command-r-plus	date_addition	**0.990**	**0.990**	**0.990**	**0.990**	**0.990**	0.980	0.980	**0.990**	**0.990**	0.950
	date_duration	0.990	0.970	0.990	0.950	0.920	0.990	0.520	0.960	**1.000**	0.970
	date_recurrence	0.150	0.130	0.100	0.030	0.000	0.180	**0.230**	0.080	0.100	0.030
	date_subtraction	0.920	0.930	0.960	0.940	**0.970**	0.940	0.940	0.890	0.930	0.940
	day_of_week	**0.820**	0.790	0.810	0.760	0.370	0.790	0.470	0.750	0.620	0.750
	interval_date	0.100	0.050	0.050	0.130	0.250	0.060	**0.450**	0.210	0.340	0.010
	time_addition	0.880	0.860	0.880	0.890	**0.920**	0.830	**0.920**	0.770	0.820	0.910
	time_duration	0.810	0.670	**0.820**	0.730	0.690	**0.820**	0.630	0.700	0.720	0.790
	time_subtraction	**0.940**	0.830	0.890	0.840	0.790	0.910	0.900	0.780	0.770	0.810

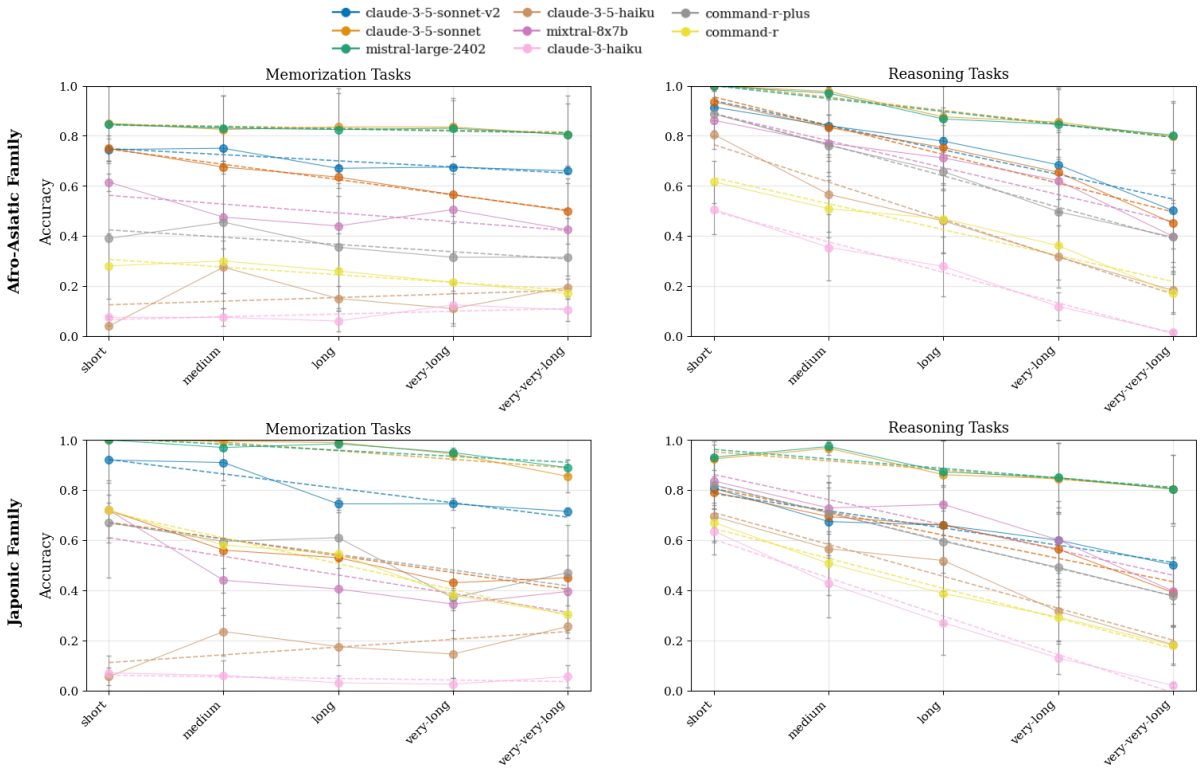

Figure 6: Accuracy by difficulty level across remaining language families.

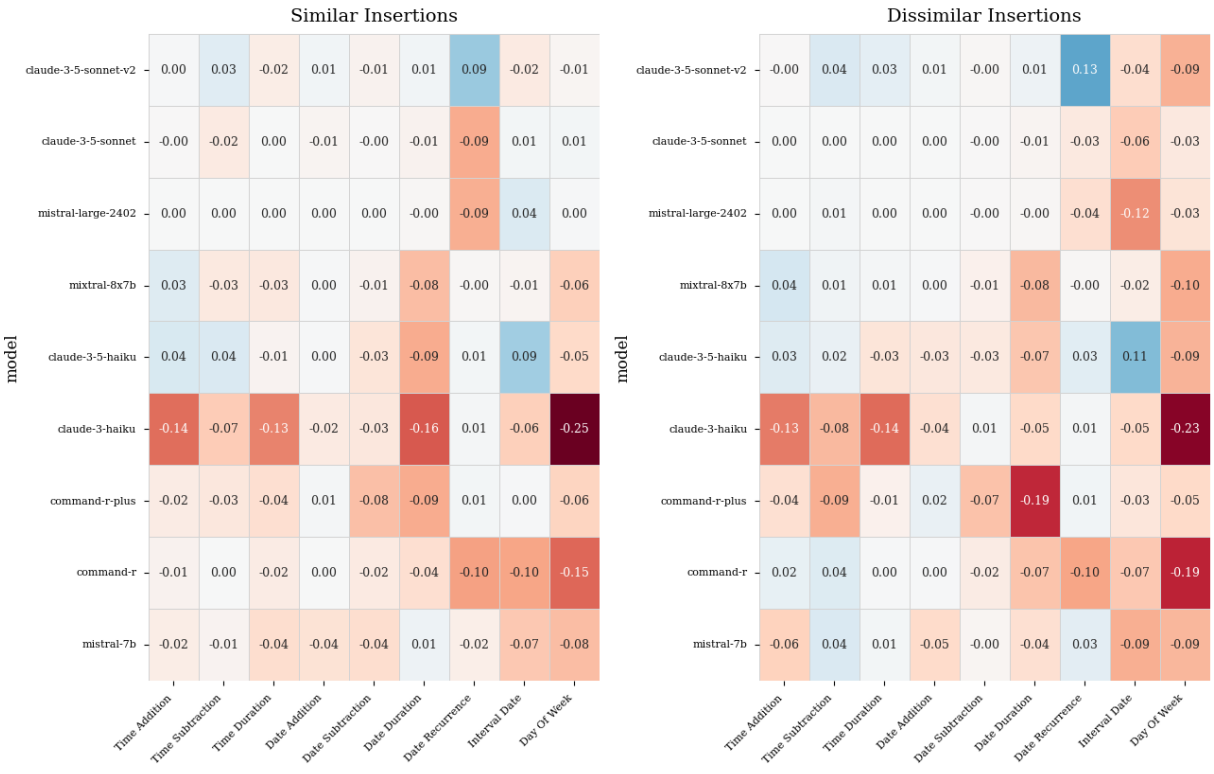

Figure 7: Impact of Insertions on Model Performance for Indo-European language families. Each block reports the absolute difference between the baseline evaluation without insertions.

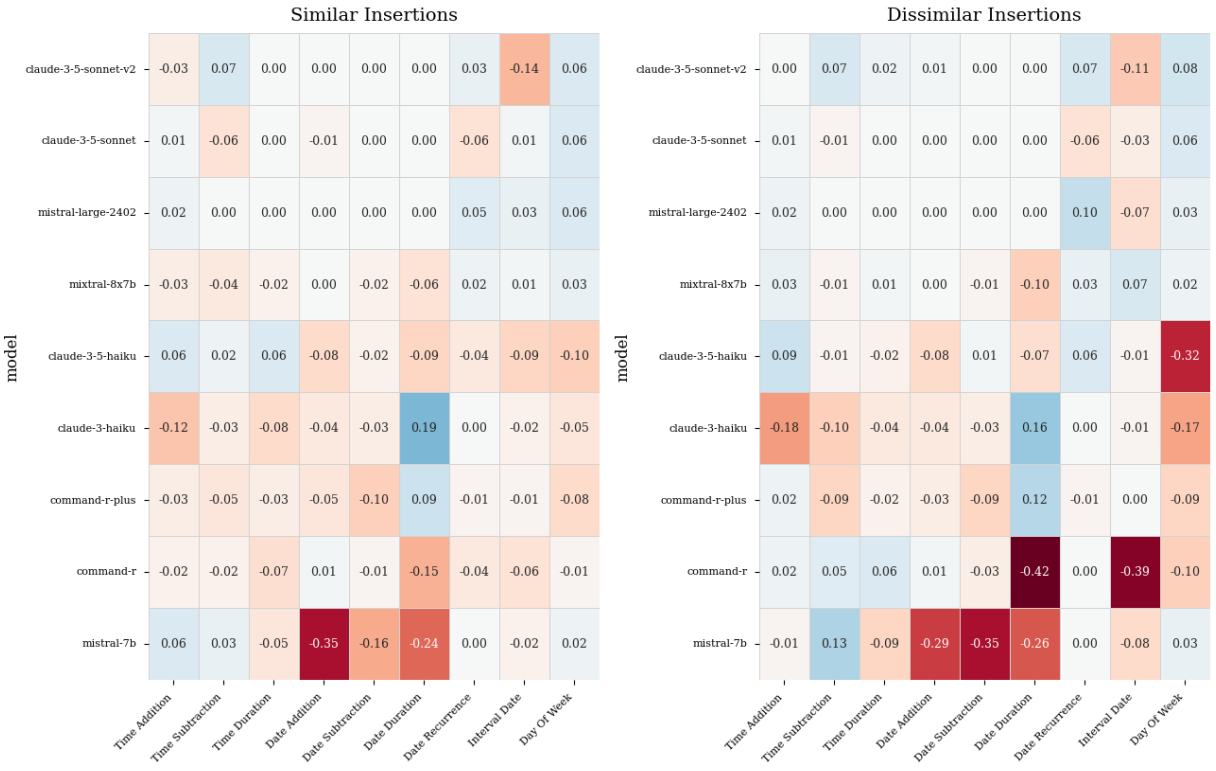

Figure 8: Impact of Insertions on Model Performance for Afro-Asiatic language families. Each block reports the absolute difference between the baseline evaluation without insertions.

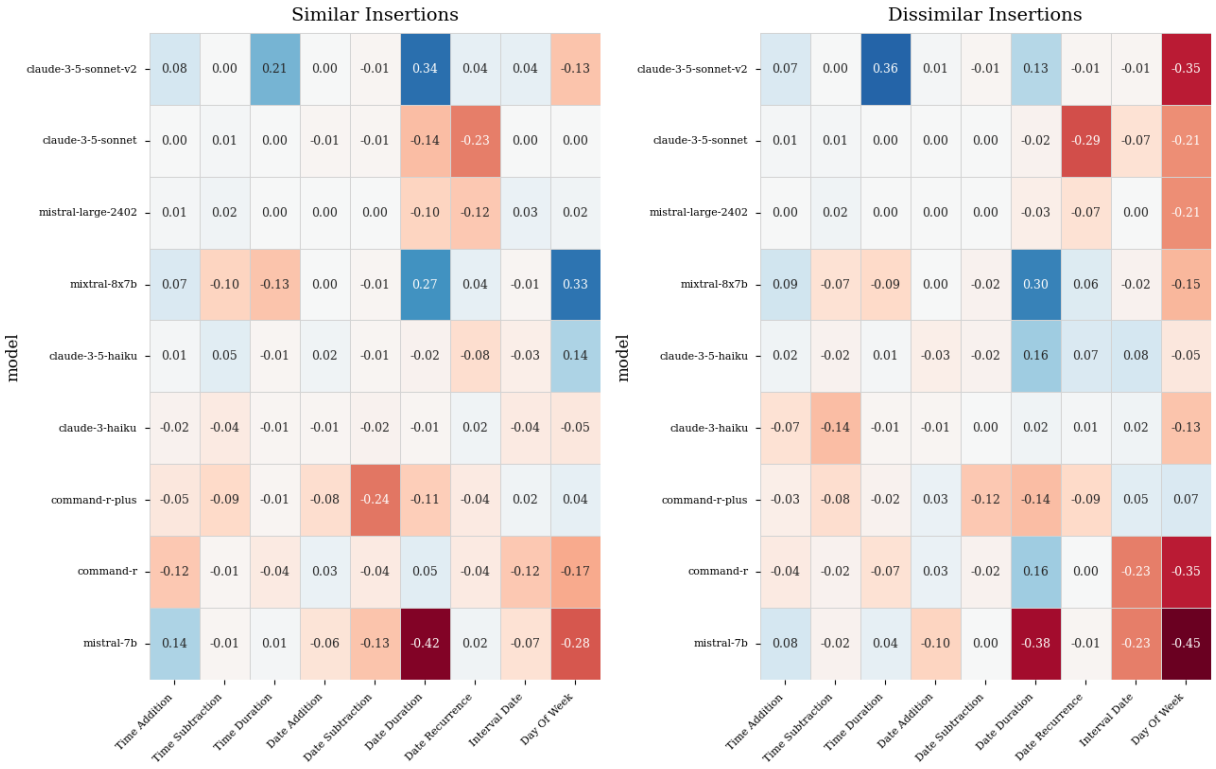

Figure 9: Impact of Insertions on Model Performance for Japonic language families. Each block reports the absolute difference between the baseline evaluation without insertions.