LOST IN TIME: CLOCK AND CALENDAR UNDERSTANDING CHALLENGES IN MULTIMODAL LLMS

Rohit Saxena **Aryo Pradipta Gema** **Pasquale Minervini**
ILCC, School of Informatics, University of Edinburgh Miniml.AI
{rohit.saxena, aryo.gema, p.minervini}@ed.ac.uk

ABSTRACT

Understanding time from visual representations is a fundamental cognitive skill, yet it remains a challenge for multimodal large language models (MLLMs). In this work, we investigate the capabilities of MLLMs in interpreting time and date through analogue clocks and yearly calendars. To facilitate this, we curated a structured dataset[1] comprising two subsets: (1) *ClockQA*, which comprises various types of clock styles—standard, black-dial, no-second-hand, Roman numeral, and arrow-hand clocks—paired with time-related questions; and (2) *CalendarQA*, which consists of yearly calendar images with questions ranging from commonly known dates (*e.g.,* Christmas, New Year's Day) to computationally derived ones (*e.g.,* the 100th or 153rd day of the year). We aim to analyse how MLLMs can perform visual recognition, numerical reasoning, and temporal inference when presented with time-related visual data. Our evaluations show that despite recent advancements, reliably understanding time remains a significant challenge for MLLMs.

1 INTRODUCTION

The ability to interpret and reason about time from visual inputs is critical for many real-world applications—ranging from event scheduling to autonomous systems. Despite advances in multimodal large language models (MLLMs), most work has focused on object detection (Wu et al., 2024), image captioning (McKinzie et al., 2024), or scene understanding (Fu et al., 2024), leaving temporal inference underexplored (Zhang et al., 2025). In particular, analogue clock reading and calendar comprehension involve intricate cognitive steps: they demand fine-grained visual recognition (*e.g.,* clock-hand position, day-cell layout) and non-trivial numerical reasoning (*e.g.,* calculating day offsets).

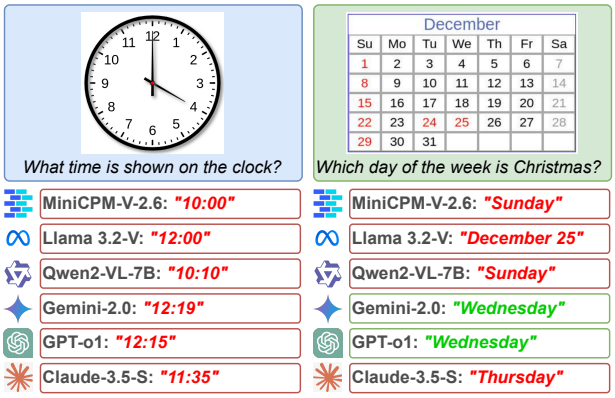


Figure 1: Predictions on ClockQA and CalendarQA.

In recent years, a variety of vision-language benchmarks were proposed to evaluate multimodal reasoning on diverse tasks such as geometry, logic, coding, and advanced mathematics (Yue et al., 2024; Kazemi et al., 2024a;b; Lu et al., 2024). Additional efforts have been made to automatically read analogue clocks and other dials (Yang et al., 2022; Alexeev et al., 2020; Bao et al., 2019; Cai et al., 2020; Howells et al., 2021), showing that dial or gauge interpretation is a cognitively complex skill requiring visual-spatial understanding and arithmetic reasoning. However, clock and calendar readings remain underexplored in these existing large-scale benchmarks, and comprehensive evaluations of MLLMs on such tasks are lacking (see Appendix A for more information on related works).

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

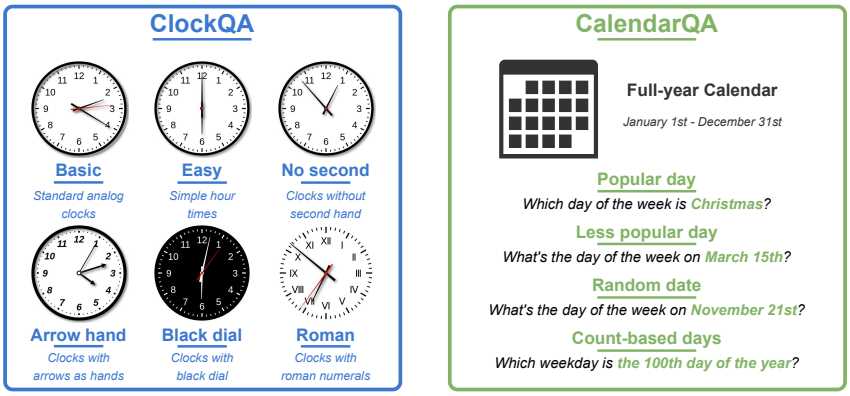


Figure 2: Overview of DateTimeReasoning and its two main subsets: ClockQA and CalendarQA

Model	Clock				Calendar			
	EM↑	MAE↓	Hour Err↓	Min Err↓	Acc↑	P↑	R↑	F1↑
Llama 3.2-Vision	3.23	10825.84	3.02	13.08	11.67	9.18	11.67	10.03
Qwen2-VL-7B	0.0	11167.71	3.06	13.16	18.33	9.51	18.33	12.09
MiniCPM-V-2.6	3.23	11078.79	3.06	10.73	20.0	19.18	20.0	15.61
Gemini-2.0	**22.58**	**6494.37**	**1.82**	**6.4**	31.67	30.32	31.67	29.79
GPT-4o	8.06	8268.42	2.29	11.97	43.33	46.12	43.33	42.12
Claude-3-5-sonnet	6.45	7964.24	2.13	12.32	46.67	42.56	46.67	43.79
GPT-o1	4.84	7954.45	2.21	8.19	**80.0**	**80.39**	**80.0**	**80.04**

Table 1: Performance of each model on Clock (left) and Calendar (right) tasks. Higher values are better (↑); lower values are better (↓).

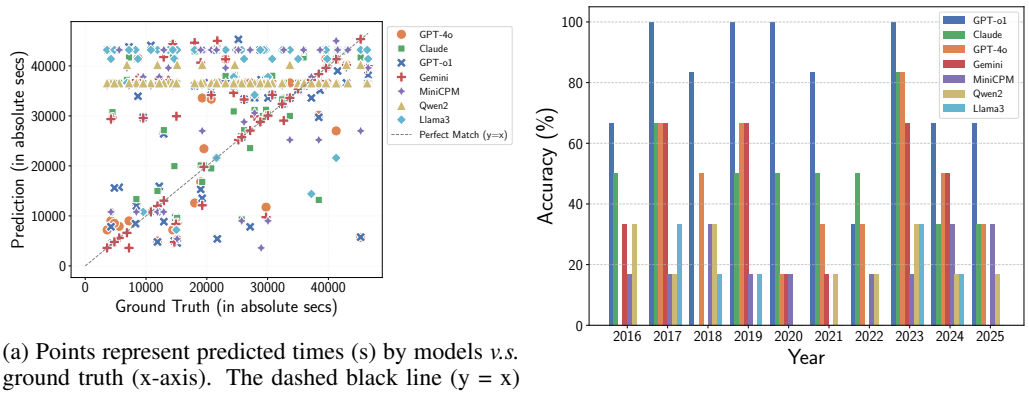
In this paper, we explore how MLLMs handle these temporal tasks. We constructed a focused test set consisting of two subsets: *ClockQA*, which includes diverse analogue clock images across six categories (including variations with Roman numerals, missing second hands, and different dial colours) paired with time-related questions, and *CalendarQA*, which comprises 10 years of calendar images paired with questions ranging from straightforward date lookups (What day of the week is New Year's Day?) to more complex queries (What is the 153rd day of the year?). While our dataset is intentionally small in scale, it is designed to probe specific aspects of temporal reasoning, visual parsing, and date/time inference.

We evaluate multiple state-of-the-art closed and open-source models on these tasks. Our preliminary findings reveal that while some models show promise in clock reading (*e.g.,* Gemini-2.0 demonstrates lower hour and minute errors) or in calendar question-answering (*e.g.,* o1 exhibits high accuracy), persistent challenges remain. Despite the limited scale of our evaluation, error analyses highlight challenges in correctly parsing clock-hand positions and performing arithmetic on dates for calendar tasks. These insights provide valuable direction for future work in the temporal reasoning capabilities of MLLMs.

2 DATASET

We created a small dataset comprising two subsets, *ClockQA* and *CalendarQA*, each containing images paired with question-answer pairs to test the time and date reasoning of multimodal large language models (MLLMs). Figure 2 illustrates the dataset and its two subsets.

ClockQA. Given an image of an analogue clock, a multimodal LLM is asked the following question *"What time is shown on the clock in the given image?"* This requires (1) detecting the clock hand positions (hour, minute, and second) and (2) converting them into time representation. The ClockQA subset contains 62 samples of analogue clocks with varying appearances, requiring precise readings of the hour, minute, and second hands. It includes the following categories: (1) **Basic clocks:** standard analogue clocks; (2) **Black dial clocks:** featuring a darker face for contrast-based parsing;



(a) Points represent predicted times (s) by models *v.s.* ground truth (x-axis). The dashed black line (y = x) represents a perfect model. Models show varying errors from this line.

(b) Year-wise accuracy of the models. Blank bar indicates accuracy as 0% for that year.

Figure 3: Error analysis for ClockQA and CalendarQA.

(3) **No second hand clocks:** simplified version of the task; (4) **Easy clocks:** standard clock set on the hour (*e.g.,* 4:00); (5) **Roman number clocks:** for digit-recognition challenges; and (6) **Arrow hand clocks:** stylized with arrows as the hands for more obvious pointer cues.

CalendarQA. Given a full image of a yearly calendar (January 1–December 31), the model answers questions from one of the categories like *"Which day of the week is New Year's Day?"* or *"What is the 153rd day of the year?"*. The system must combine visual parsing of date cells and textual labels with date arithmetic or reasoning about day offsets. The CalendarQA subset spans 10 full years (January 1 to December 31), each with six questions focusing on four main categories: (1) **Popular days** (*e.g.,* Christmas and New Year's); (2) **Less popular dates** such as the Ides of March (March 15); (3) **Random dates** like November 21; and (4) **Count-based days** referring to the nth day of the year (*e.g.,* the 100th).

Together, these two subsets evaluate an MLLM's ability to parse and reason about time and date information in a multimodal context.

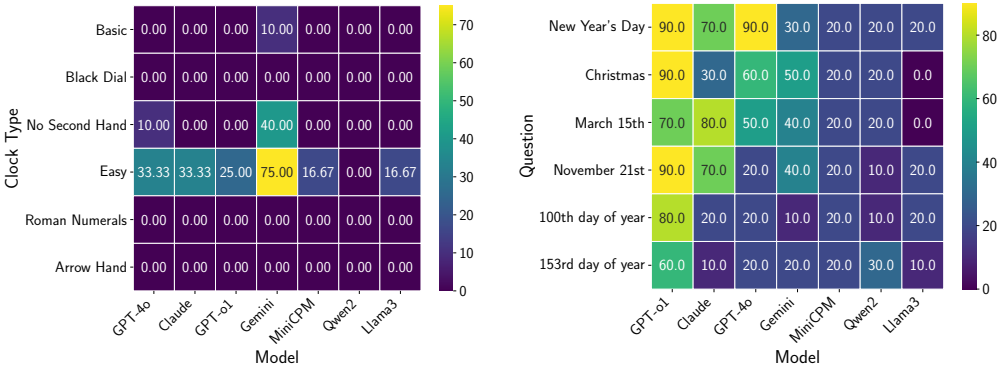
3 Tasks and Experiments

Experimental Setup. We evaluate seven multimodal LLMs in a zero-shot setting. We evaluate closed-source multimodal models, including GPT-4o (OpenAI et al., 2024), GPT-o1 (OpenAI et al., 2024), Gemini 2.0 (Team et al., 2024), and Claude 3.5 Sonnet (Anthropic, 2024). We also evaluate open-source models such as Llama 3.2-11B-Vision-Instruct (Grattafiori et al., 2024), Qwen2-VL-7B-Instruct (Yang et al., 2024), and MiniCPM-V-2.6 (Yao et al., 2024). The experiment details and exact prompt template used are in Appendix D.

ClockQA Metrics. We measure performance using four metrics. **Exact Match (EM)** is the proportion of predicted clock readings that exactly match the ground truth. **MAE (Seconds)** quantifies the average absolute difference between predicted and actual times in seconds, applying a circular 12-hour wraparound (maximum error: 21,600 seconds). We also report **Hour Error and Minute Error**, which compute mean absolute differences for each clock hand, again with a circular wraparound. See Appendix C.1 for more details.

CalendarQA Metrics. We adopt standard classification metrics: **Accuracy (Acc)** to measure correct weekday predictions, and macro-averaged **Precision (P)**, **Recall (R)**, and **F1** across date categories. See Appendix C.2 for more details.

Implementation Details. We conduct experiments on a shared test set of 62 clock samples (across six clock-face variants) and 10 calendar years (with six question types per year). Model prompts followed a consistent format, providing one image and one question. Responses are automatically parsed to extract time or weekday (*i.e.,* removing explanation or conversion of short forms) and evaluated against the reference answer.



(a) Clock category-wise accuracy of the models.

(b) Calendar question-wise accuracy of the models.

Figure 4: ClockQA and CalendarQA question and category-based analysis.

4 Results and Discussion

Table 1 summarizes performance across both tasks. In ClockQA, **Gemini-2.0** achieves the highest EM score (22.58%) and the lowest hour/minute errors, indicating relatively stronger clock understanding compared to other models. However, overall EM scores remain low, underscoring persistent difficulties in clock reading by MLLMS. Conversely, **GPT-o1** excels in CalendarQA with an accuracy of 80%, highlighting robust date arithmetic and reasoning capabilities. Other models lag substantially, indicating that date arithmetic and structured layout parsing remain challenging. Overall performance on both ClockQA and CalendarQA remains poor, except for the high performance of GPT-o1 on CalendarQA. See Appendix E for a sample of generated predictions.

Clock Reading Remains Error-Prone. Across the ClockQA subset, performance was notably weaker than for the calendar questions (see Table 1). Figures 4a and 3a reveal that performance remains poor even on standard dials; some models exhibit bias toward a single "default" time. Roman numerals and stylized clock hands further increase the errors. Removing the second hand did not simplify reasoning, suggesting deep-seated issues with hand detection and angle interpretation.

Calendar Reasoning Analysis By contrast, calendar tasks elicited higher success rates for certain models and question types. GPT-o1 dominates the CalendarQA subset with an 80% overall accuracy (Table 1 and Figure 3b).

Closed-source models like GPT-o1 and Claude-3.5 outshine open-source ones on popular holidays, potentially reflecting memorized patterns in the training data (see Figure 4b). However, accuracy diminishes substantially for lesser-known or arithmetically demanding queries (*e.g.,* 153rd day), indicating that performance does not transfer well to offset-based reasoning. The drop is especially evident among smaller or open-source models (MiniCPM, Qwen2-VL-7B, and Llama3.2-Vision), which exhibit near-random performance on less popular or offset-based queries.

5 Conclusion

In this work, we conduct a preliminary study on understanding and reasoning about time from visual inputs, which remains a significant challenge for multimodal large language models. We build a small dataset to benchmark these models for clock and calendar understanding. The experimental results highlight key shortcomings in the ability of these models to accurately interpret time from analogue clocks and yearly calendars. Our findings suggest that successful temporal reasoning requires a combination of precise visual perception, numerical computation, and structured logical inference that current MLLMs have not yet mastered. This work highlights the need for further research to improve the processing of geometric relationships in clock faces and structured calendar information in MLLMs.

ACKNOWLEDGMENTS

This work was supported in part by the School of Informatics at the University of Edinburgh. Aryo Pradipta Gema was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. Pasquale Minervini was partially funded by ELIAI (The Edinburgh Laboratory for Integrated Artificial Intelligence), EPSRC (grant no. EP/W002876/1), and a donation from Accenture LLP. This work was also supported by the Edinburgh International Data Facility (EIDF) and the Data-Driven Innovation Programme at the University of Edinburgh.

REFERENCES

Alexey Alexeev, Georgy Kukharev, Yuri Matveev, and Anton Matveev. A highly efficient neural network solution for automated detection of pointer meters with different analog scales operating in different conditions. *Mathematics*, 8(7):1104, 2020.

Anthropic. Claude 3.5 - sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. Accessed: 2024-12-06.

Haojing Bao, Qingchang Tan, Siyuan Liu, and Jianwei Miao. Computer vision measurement of pointer meter readings based on inverse perspective mapping. *Applied Sciences*, 9(18), 2019. ISSN 2076-3417. doi: 10.3390/app9183729. URL https://www.mdpi.com/2076-3417/9/18/3729.

Weidong Cai, Bo Ma, Liu Zhang, and Yongming Han. A pointer meter recognition method based on virtual sample generation technology. *Measurements*, 163:107962, October 2020. doi: 10.1016/j.measurement.2020.107962.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. URL https://arxiv.org/abs/2306.13394.

Isaac R. Galatzer-Levy, Jed McGiffin, David Munday, Xin Liu, Danny Karmon, Ilia Labzovsky, Rivka Moroshko, Amir Zait, and Daniel McDuff. Evidence of cognitive deficits and developmental advances in generative AI: A clock drawing test analysis. *CoRR*, abs/2410.11756, 2024.

Deepanway Ghosal, Vernon Toh Yan Han, Chia Yew Ken, and Soujanya Poria. Are language models puzzle prodigies? algorithmic puzzles unveil serious challenges in multimodal reasoning, 2024. URL https://arxiv.org/abs/2403.03864.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

B. Howells, J. Charles, and R. Cipolla. Real-time analogue gauge transcription on mobile phone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

David T. Jones and Jonathan Graff-Radford. Executive dysfunction and the prefrontal cortex. *CONTINUUM Lifelong Learning in Neurology*, 27(6):1586–1601, December 2021. ISSN 1080-2371. doi: 10.1212/CON.0000000000001009. Publisher Copyright: © Lippincott Williams & Wilkins.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning. In *AI for Math Workshop @ ICML 2024*, 2024a. URL https://openreview.net/forum?id=1AUbiBrOF1.

Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Sreenivas Gollapudi, Dee Guo, and Ahmed Qureshi. ReMI: A dataset for reasoning with multiple images. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=930e8v5ctj.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis and insights from multimodal llm pre-training. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXIX*, pp. 304–323, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73396-3. doi: 10.1007/978-3-031-73397-0_18. URL https://doi.org/10.1007/978-3-031-73397-0_18.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and Ilge Akkaya et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, and Radu Soricut et al. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXII*, pp. 164–182, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73410-6. doi: 10.1007/978-3-031-73411-3_10. URL https://doi.org/10.1007/978-3-031-73411-3_10.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, and Chang Zhou et al. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Charig Yang, Weidi Xie, and Andrew Zisserman. It's about time: Analog clock reading in the wild. In *CVPR*, pp. 2498–2507. IEEE, 2022.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. URL https://arxiv.org/abs/2408.01800.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. MME-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k5VHHgsRbi.

A RELATED WORK

Vision-Language Benchmarks. Existing benchmarks for MLLMs cover various tasks, from college-level subject knowledge to advanced mathematical and multi-step reasoning. Massive Multi-discipline Multimodal Understanding (MMMU, Yue et al., 2024) tests deliberate reasoning across 11.5K multimodal questions drawn from college exams, quizzes, and textbooks, spanning disciplines such as Art, Business, Science, Health, Social Science, and Engineering. MathVista Lu et al. (2024) gauges mathematical reasoning within visual contexts, while GeomVerse Kazemi et al. (2024a) evaluates geometry-based problem-solving in MLLMs. ReMI Kazemi et al. (2024b) focuses on multi-image reasoning, encompassing diverse domains, including math, physics, logic, code, table/chart comprehension, and spatio-temporal tasks. Despite this breadth, none of these datasets specifically targets analogue clock and calendar interpretation.

Analogue Clock and Dial Reading. Reading analogue clocks is a complex cognitive task that engages multiple mental processes Galatzer-Levy et al. (2024). It involves several key cognitive components: visuospatial skills for understanding spatial relationships between clock elements, executive functioning for planning and reasoning, working memory to maintain mental representations of time concepts, and sustained attention to process the information accurately Galatzer-Levy et al. (2024); Jones & Graff-Radford (2021). Yang et al. (2022) provide a comprehensive framework for reading analogue clocks in natural images and videos, introducing a synthetic data generation pipeline which can produce a wide variety of clock images reflecting the challenges encountered in real-world scenes. Recently, Deitke et al. (2024) introduced a large-scale synthetic dataset of analogue clocks comprising images rendered from different watch models. However, the dataset's generation pipeline is not publicly available, making it difficult to reproduce. Ghosal et al. (2024) propose a puzzle-based task that includes clock puzzles to identify when an event occurred or will occur, given a starting time and an elapsed or future duration. Another line of work focuses on automatic dial or gauge meter readings. The solutions proposed for dial reading rely on neural models Alexeev et al. (2020), projective transforms Bao et al. (2019), and virtual dataset generators Cai et al. (2020); Howells et al. (2021), which produce accurate results for gauges with known shape and style.

B ANALYSIS OF EASY CLOCK PREDICTIONS

Time	Number of Models	Models with wrong prediction
1:00	6	All except Gemini
2:00	7	All
3:00	4	Llama3, Qwen2, MiniCPM, GPT-o1
4:00	7	All
5:00	7	All
6:00	3	Qwen2, MiniCPM, GPT-o1
7:00	6	All except Gemini
8:00	5	Llama3, Qwen2, MiniCPM, GPT-4o, GPT-o1
9:00	4	Llama3, Qwen2, MiniCPM, Claude, GPT-o1
10:00	4	Llama3, Qwen2, Claude, GPT-4o
11:00	6	All except Gemini
12:00	1	Qwen2

Table 2: Number of models with incorrect predictions at each hour (12-hour format).

Table 2 provides model performance across different times of the day for the easy clock category. The results indicate that certain times pose greater challenges for MLLMs, with times such as 2:00, 4:00, and 5:00 being misclassified by all models. Errors on this simpler task highlight gaps in accurate clock reading.

C METRICS

C.1 CLOCKQA METRICS

We evaluate the clock-reading performance with the following metrics:

Exact Match (EM). The proportion of predictions that exactly match the ground truth time:

$$\text{Exact Match Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\Big[T_{\text{true},i} = T_{\text{pred},i}\Big]. \tag{1}$$

MAE (Seconds). The mean absolute error in seconds, with a 12-hour circular wraparound (*i.e.,* we measure the shorter way around the clock face with a maximum error of 21,600):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\min\Big(\big|T_{\text{true},i} - T_{\text{pred},i}\big|, ; 43200 - \big|T_{\text{true},i} - T_{\text{pred},i}\big|\Big). \tag{2}$$

Hour and Minute Errors. Average absolute differences for hour and minute hands, each with a circular wraparound:

$$\text{MAE}_{\text{hours}} = \frac{1}{n}\sum_{i=1}^{n}\min\Big(|H_{\text{true},i} - H_{\text{pred},i}|,\ 12 - |H_{\text{true},i} - H_{\text{pred},i}|\Big), \tag{3}$$

$$\text{MAE}_{\text{minutes}} = \frac{1}{n}\sum_{i=1}^{n}\min\Big(|M_{\text{true},i} - M_{\text{pred},i}|,\ 60 - |M_{\text{true},i} - M_{\text{pred},i}|\Big), \tag{4}$$

C.2 CALENDARQA METRICS

For calendar-based reasoning, we employ standard classification metrics:

Accuracy (Acc). The fraction of correct predictions for the day of the week.

Precision (P), Recall (R), & F1. Macro-averaged scores across different date categories.

D EXPERIMENT DETAILS AND PROMPT TEMPLATE

For closed-source models, we used the default settings specified by their respective platforms. Open-source models were evaluated with a beam size of 4 and greedy sampling to ensure reproducibility. We used the following prompts for the models and also applied standard model-specific chat templates when available.

Prompt Template for ClockQA

<Question> + Only output the time.

Example: What time is shown on the clock in the given image? Only output the time.

E SAMPLE OF CLOCKQA AND CALENDARQA PREDICTIONS

Sample images of clocks and calendars from the dataset along with model predictions.

Prompt Template for CalendarQA

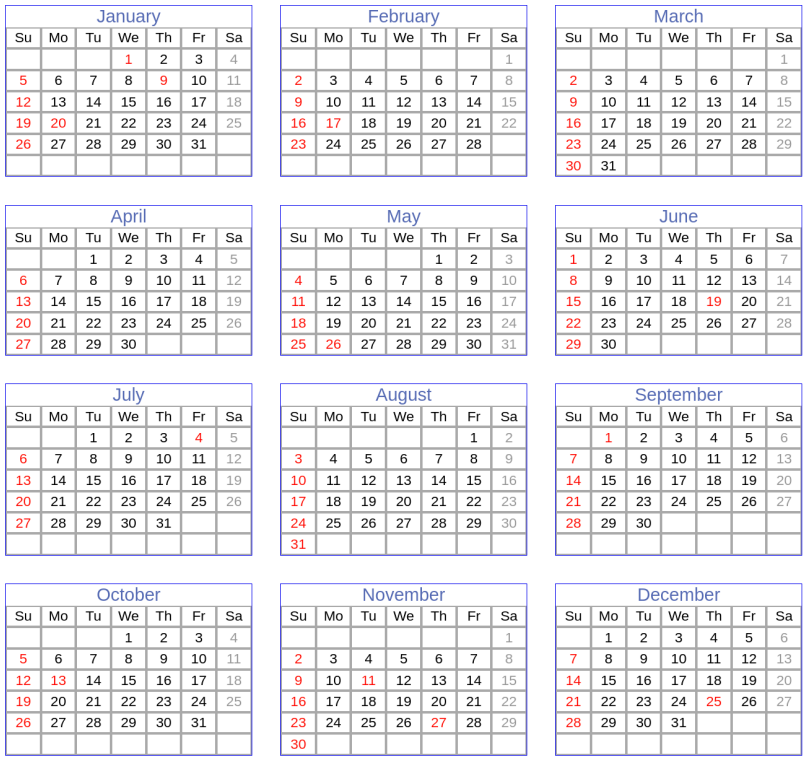
<Question> + in the given calendar image? Only output the day without explanation.

Example: Which day of the week is New Year's Day in the given calendar image? Only output the day without explanation.

Image	Clock Type	Ground Truth	Model Predictions
	Basic Dial	**10:40:15**	GPT-4o: 8:22:15 Claude: 3:40:00 GPT-o1: 8:15:15 Gemini: 10:39:15 MiniCPM: 7:00:00 Qwen2: 10:10:10 Llama3: 10:03:30
	Easy (Exact Hour)	**7:00:00**	GPT-4o: 10:10:00 Claude: 11:58:30 GPT-o1: 12:35:00 Gemini: 7:00:00 MiniCPM: 12:00:00 Qwen2: 10:10:10 Llama3: 12:00:00
	No Second hand	**12:53:00**	GPT-4o: 10:10 Claude: 10:10 GPT-o1: 10:10 Gemini: 10:59 MiniCPM: 10:09 Qwen2: 10:10 Llama3: 12:00
	Arrow hand	**6:49:58**	GPT-4o: 10:08:46 Claude: 10:35:00 GPT-o1: 9:50:00 Gemini: 10:30:00 MiniCPM: 10:00 Qwen2: 11:11:11 Llama3: 10:10:00
	Black Dial	**2:24:38**	GPT-4o: 11:31:45 Claude: 10:10:30 GPT-o1: 9:25:40 Gemini: 10:25:39 MiniCPM: 10:09 Qwen2: 10:10:10 Llama3: 11:30:00
	Roman Numbers	**8:15:44**	GPT-4o: 3:15:45 Claude: 8:40:30 GPT-o1: 10:15:45 Gemini: 2:42:25 MiniCPM: 10:09:09 Qwen2: 10:10:10 Llama3: 12:00:00

Table 3: Clock image samples of different categories with model predictions.

Calendar Image – 2025



Question: Which day of the week is New Year's Day in the given calendar image?

Ground Truth: Wednesday

Model Predictions:
GPT-4o: Wednesday
Claude: Wednesday
GPT-o1: Wednesday
Gemini: Tuesday
MiniCPM: Monday
Qwen2: Monday
Llama3: Sunday

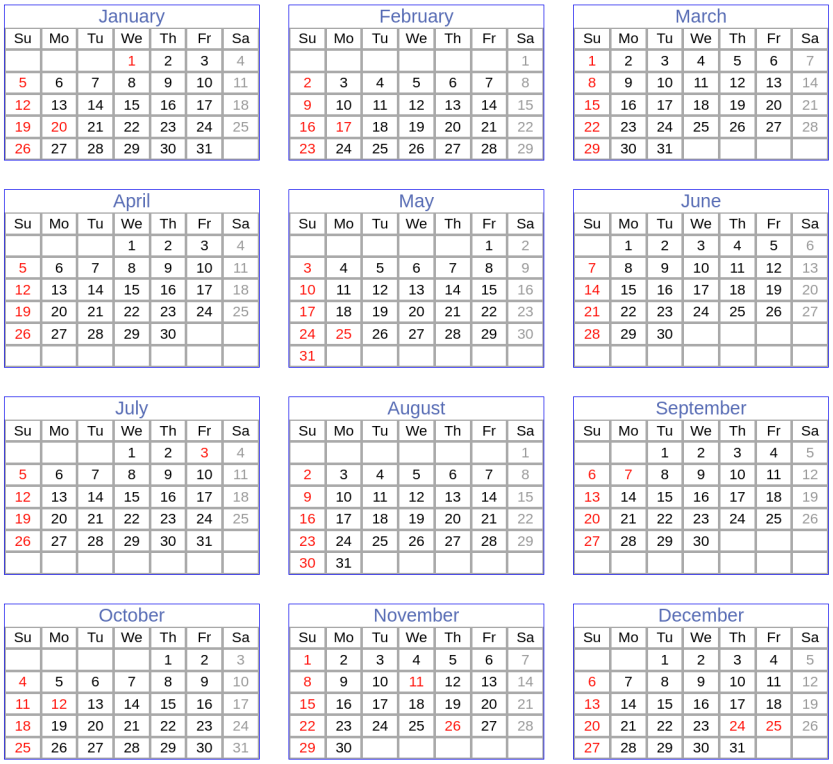
Question: Which weekday corresponds to the 100th day of the year (Assume January 1st is day 1.) in the given calendar image?

Ground Truth: Thursday

Model Predictions:
GPT-4o: Tuesday
Claude: Monday
GPT-o1: Thursday
Gemini: Tuesday
MiniCPM: Thursday
Qwen2: Monday
Llama3: Monday

Table 4: Sample calendar image of the year 2025 with model predictions.

Calendar Image – 2020



Question: What's the day of the week on November 21st in the given calendar image?

Ground Truth: Saturday

Model Predictions:
GPT-4o: Friday
Claude: Tuesday
GPT-o1: Saturday
Gemini: Thursday
MiniCPM: Friday
Qwen2: Wednesday
Llama3: Wednesday

Question: Which day of the week is Christmas in the given calendar image?

Ground Truth: Friday

Model Predictions:
GPT-4o: Thursday
Claude: Friday
GPT-o1: Friday
Gemini: Sunday
MiniCPM: Sunday
Qwen2: Monday
Llama3: December 25

Table 5: Sample calendar image of the year 2019 with model predictions.