CLOCKBENCH: VISUAL TIME BENCHMARK WHERE HUMANS BEAT THE CLOCK, LLMS DON'T

ALEK SAFAR (OLEG CHICHIGIN)

alek@clockbench.ai

Sep 2, 2025

Abstract. Introducing ClockBench, an open-ended visual benchmark focused on telling the time with analog clocks. ClockBench includes 180 clocks, 720 total questions, and demonstrates the limitations of current frontier LLMs. Untrained humans reach 89.1% average accuracy in valid time recognition, whereas the top model out of 11 tested LLMs only reaches 13.3%, making this visual reasoning test harder for models than knowledge-intensive benchmarks such as Humanity's Last Exam.

1. Introduction

Over the last few years, we have seen major progress in LLMs across multiple domains, with frontier models quickly saturating popular benchmarks. *MMLU* (Dan Hendrycks et al., 2020) and subsequently *GPQA* (David Rein et al., 2023) are no longer posing a challenge to SOTA models, and even latest benchmarks designed to require both extensive specialized knowledge and strong reasoning capabilities are seeing quick progress. One such example is *Humanity's Last Exam* (Long Phan et al., 2025) that saw an increase from 2.7% for OpenAI GPT-4o to 25.4% for xAI Grok 4, and results getting into 40-50% territory with tool use and other optimizations.

However, we are still seeing models struggle with certain tasks that are often trivial for humans, so benchmarks such as *SimpleBench* (Philip and Hemang, 2024) and *ARC-AGI* (Francois Chollet, 2019) were designed to be simple for an average person, while challenging for LLMs.

ClockBench is inspired by this "easy for humans, hard for AIs" approach and is built upon the insight that reading analog clocks seems to be a challenge for both reasoning and non-reasoning models (Rohit Saxena et al., 2025), by designing a robust dataset that requires high visual precision and reasoning.

2. Methods

2.1. **Dataset.** ClockBench dataset was fully designed from scratch to avoid any potential contamination problems. Full dataset is intentionally kept private (with a small public sample made available) and includes:

- 36 unique clock faces, 5 sample clocks per clock face
- 180 total clocks, with 4 questions per clock, i.e. 720 total questions

Each of the clock faces has a unique combination of features out of the following list of options:

- *Validity*: valid or invalid (impossible) time
- *Dial*: white, black, with a background (multi-colored or image-based)
- *Format*: 12 hours or 24 hours (with two 24 hour versions)
- *Hands*: one (hours), two (hours and minutes) or all three (hours, minutes, seconds)
- *Hands Size*: small or large
- *Numerals*: 24, 12, 4 or no numerals present
- *Numerals Type*: Arabic or Roman
- *Numerals Orientation*: upright or circular
- *Graduations*: with or without graduations
- *Border*: with border, without border, or custom border for each number
- *Mirroring*: regular or mirrored
- *Distortion*: regular or distorted (irregular) shape
- *Date*: with or without date indication
- *Day of the Week*: with or without day of the week indication
- *Month*: with or without month indication

Each of the 180 entries included the following 4 questions, as well as additional prompt for the models to respond in a specific JSON format for consistency:

- What time is it? Is it a valid time that is possible? Always include validity, hours and minutes in your response, only include seconds if there is a seconds hand present. Include date, month and day of the week in your answer if they are explicitly shown. Use 12 or 24 hour format based on what you see in the image.
- What time is it going to be if we were to add (or subtract) X hours, Y minutes and Z seconds to (from) it (X, Y and Z varies per question)? Use 24 hour format. If the original time was invalid, include that in your response.
- What time is it going to be if we were to move the minutes hand by X (X varies between 30, 60 or 90) degrees clockwise (or counterclockwise)? Use 12 or 24 hour format based on the original image. If the original time was invalid, include that in your response.
- If the time in the image is from New York in June, what is the corresponding time in X (X varying between London, Lisbon etc.) time zone? Use 24 hour format. If the original time was invalid, include that in your response.

Level of required precision varied from exact match to a small range in certain cases (such as a range of 2-10 minutes when only hour hand was present).

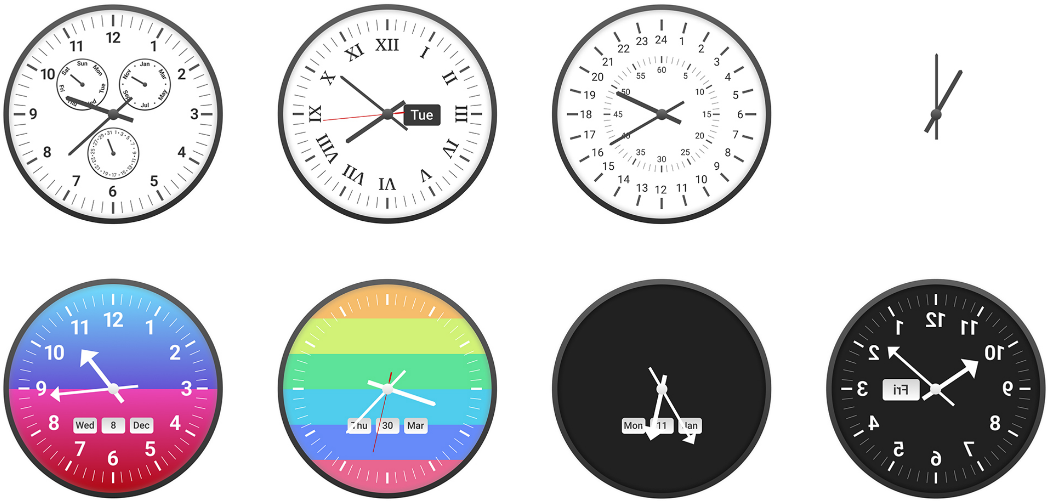


FIGURE 1. Sample clock faces

2.2. **Experiment Setup.**

2.2.1. *LLMs.*

- 11 models capable of visual understanding from 6 labs were tested
- Models completed all 180 entries and 720 total questions in a single pass
- Models were tested via OpenRouter API with no additional parameters; exception was GPT-5, for which the reasoning budget was set to 'high'
- Models were prompted via their APIs with no additional restrictions or additional tooling
- Models were not restricted in terms of total time spent or time spent per question
- Models were asked to provide a structured JSON output; in a rare case of a syntax or formatting issue, given clock would be re-run (i.e. models were not judged based on their JSON formatting capabilities)

2.2.2. *Human Participants.*

- 5 adult unspecialized humans were tested
- Humans completed all 180 entries, focusing on a time recognition question
- Humans were not controlled for IQ or ability
- Humans were provided with the same questions and instructions (including structuring their answers in JSON) as models
- Humans were not restricted in terms of total time spent or time spent per question

3. Results and Discussion

3.1. **Overall Accuracy.** Overall accuracy in recognizing valid clocks for all tested models was significantly lower than an average human baseline:

- Human accuracy was ranging between 88.1% and 93.7% (89.1% average)
- Top model - Google Gemini 2.5 Pro - scored 13.3%

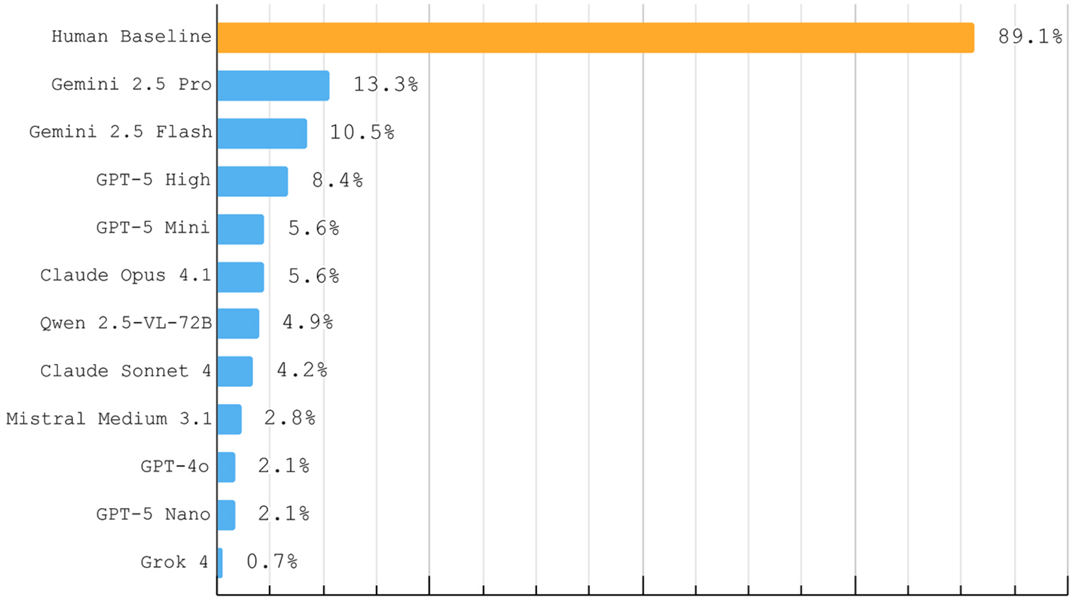


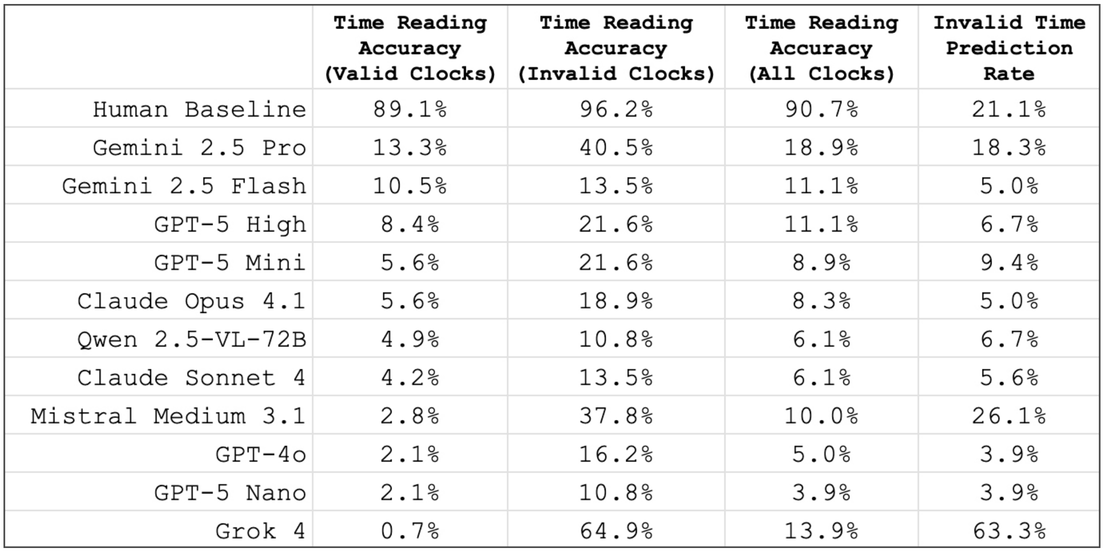
Figure 2. Time reading accuracy (valid clocks)

Between different models performance followed general trend of larger reasoning models outperforming smaller and non-reasoning models, with a few notable observations:

- Google models tend to outperform other models in their categories
- Anthropic models tend to lag behind other models in their categories
- Grok 4 performed unexpectedly poorly for a model of its size and general capability

3.2. **Accuracy on Valid vs Invalid Clocks.** In the original dataset 37 out of 180 clocks had invalid (impossible) times. Both humans and models alike had a higher success rate at detecting invalid times:

- Human difference was small - 96.2% accuracy on invalid vs 89.1% on valid
- Models difference was substantial - on average models were 349% more accurate on invalid clocks, with every model showing better performance
- Gemini 2.5 Pro remained the top performer with 40.5% accuracy
- Grok 4 was an outlier with the highest accuracy of identifying invalid clocks at 64.9%, but it also labeled 63.3% of the whole dataset as invalid, i.e. the result appears to be random



	Time Reading Accuracy (Valid Clocks)	Time Reading Accuracy (Invalid Clocks)	Time Reading Accuracy (All Clocks)	Invalid Time Prediction Rate
Human Baseline	89.1%	96.2%	90.7%	21.1%
Gemini 2.5 Pro	13.3%	40.5%	18.9%	18.3%
Gemini 2.5 Flash	10.5%	13.5%	11.1%	5.0%
GPT-5 High	8.4%	21.6%	11.1%	6.7%
GPT-5 Mini	5.6%	21.6%	8.9%	9.4%
Claude Opus 4.1	5.6%	18.9%	8.3%	5.0%
Qwen 2.5-VL-72B	4.9%	10.8%	6.1%	6.7%
Claude Sonnet 4	4.2%	13.5%	6.1%	5.6%
Mistral Medium 3.1	2.8%	37.8%	10.0%	26.1%
GPT-4o	2.1%	16.2%	5.0%	3.9%
GPT-5 Nano	2.1%	10.8%	3.9%	3.9%
Grok 4	0.7%	64.9%	13.9%	63.3%

FIGURE 3. Time reading accuracy (valid and invalid clocks)

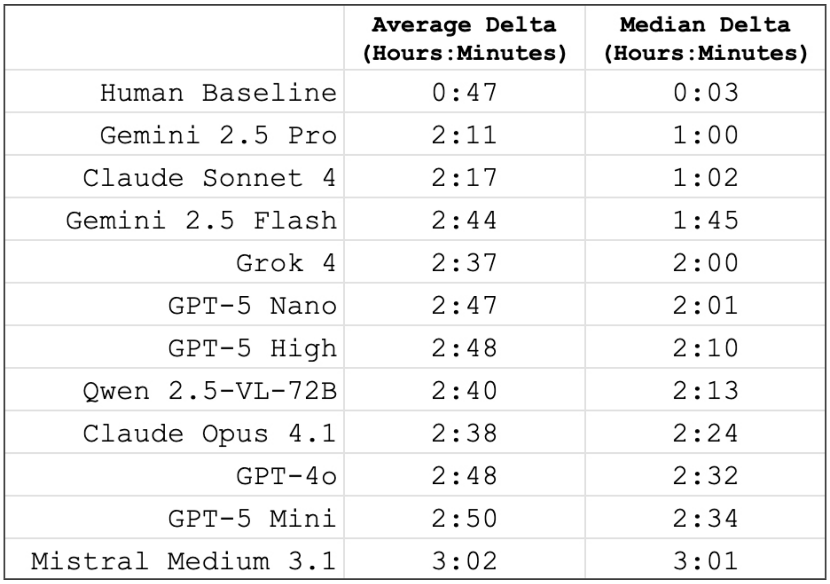
3.3. **Distribution of Correct Answers.** Models had a significant overlap in terms of clocks that they were able to read correctly:

- **61.7%** of all clocks were not read correctly by any model
- **38.3%** of all clocks were read correctly by at least 1 model
- **22.8%** of all clocks were read correctly by at least 2 models
- **13.9%** of all clocks were read correctly by at least 3 models
- **8.9%** of all clocks were read correctly by 4 or more models

Overall, the distribution and validity data points towards clustering of correct model answers around a subset of clocks.

3.4. **Error Sizes for Incorrect Answers.** In cases where models were unable to read time correctly, they tended to make larger mistakes compared to humans:

- Circular 12 or 24-hour wrap-around was used depending on the original clock format to determine the error sizes (time deltas between given answer and correct answer)
- Human median error size was 3 minutes, with a large percentage of human mistakes being minor misreadings
- Model median error size was 1 hour for the best performing model, exceeding 3 hours for the worst performing model
- Average model error sizes were clustered between roughly 2 and 3 hours
- Given that the majority of clocks were in 12 hour format (165 out of 180), error size of 3 hours would be roughly equivalent a random guess



	Average Delta (Hours:Minutes)	Median Delta (Hours:Minutes)
Human Baseline	0:47	0:03
Gemini 2.5 Pro	2:11	1:00
Claude Sonnet 4	2:17	1:02
Gemini 2.5 Flash	2:44	1:45
Grok 4	2:37	2:00
GPT-5 Nano	2:47	2:01
GPT-5 High	2:48	2:10
Qwen 2.5-VL-72B	2:40	2:13
Claude Opus 4.1	2:38	2:24
GPT-4o	2:48	2:32
GPT-5 Mini	2:50	2:34
Mistral Medium 3.1	3:02	3:01

Figure 4. Time deltas (error sizes) for incorrect answers

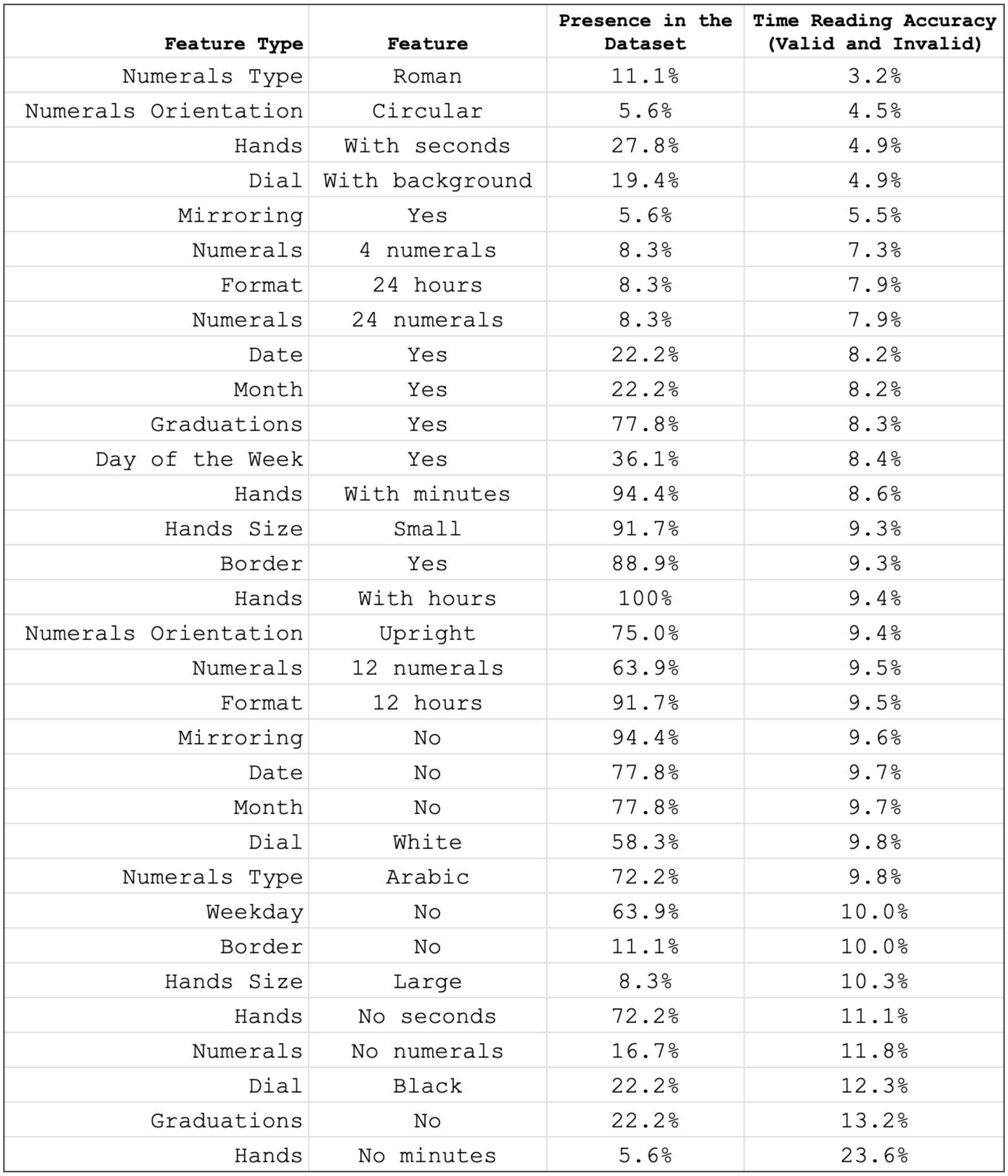
3.5. **Distribution of Accuracy by Feature.** Model accuracy varied significantly depending on the features of a given clock.

3.5.1. *Most Challenging Features.*

- Overall, models were worst at reading less common clocks with more complexity and higher degree of required precision
- Roman numerals and circular numerals orientation were the hardest
- Presence of second hand added to the precision requirement, and multi-colored or image-based backgrounds might have impaired overall visual understanding by adding clutter
- Mirroring proved to be challenging, but not completely impossible - models were able in certain cases correctly identify and untangle it
- Uncommon clock formats - 24 hour clocks, or clocks only showing 4 numerals (3, 6, 9 and 12), as well as additional variables (date, day of the week or month) were also associated with slightly lower accuracy

3.5.2. *Least Challenging Features.*

- Overall, models were best at reading common clocks with less complexity and lower precision thresholds
- Clocks with a single hour hand were easiest to read, given that during grading they had the largest allowed margin of error
- Similarly, clocks with no graduations, no numerals and no second hand required less precision and were easier to get right
- Most common and simple formats - black or white dials with no additional details, large visible hands and regular Arabic numerals were also associated with higher accuracy



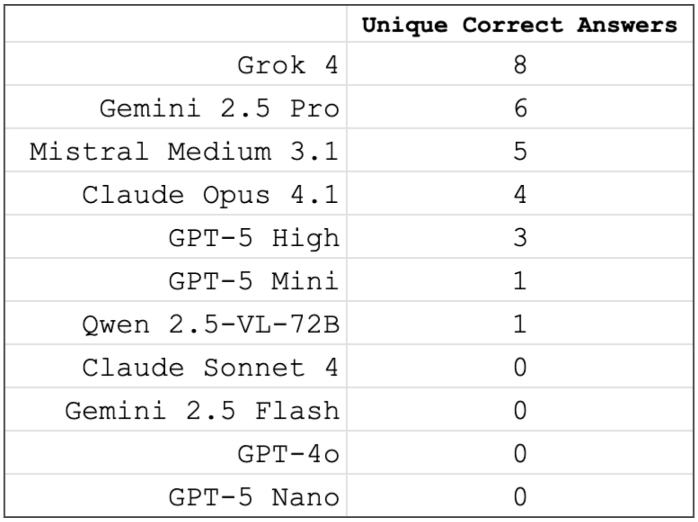
Feature Type	Feature	Presence in the Dataset	Time Reading Accuracy (Valid and Invalid)
Numerals Type	Roman	11.1%	3.2%
Numerals Orientation	Circular	5.6%	4.5%
Hands	With seconds	27.8%	4.9%
Dial	With background	19.4%	4.9%
Mirroring	Yes	5.6%	5.5%
Numerals	4 numerals	8.3%	7.3%
Format	24 hours	8.3%	7.9%
Numerals	24 numerals	8.3%	7.9%
Date	Yes	22.2%	8.2%
Month	Yes	22.2%	8.2%
Graduations	Yes	77.8%	8.3%
Day of the Week	Yes	36.1%	8.4%
Hands	With minutes	94.4%	8.6%
Hands Size	Small	91.7%	9.3%
Border	Yes	88.9%	9.3%
Hands	With hours	100%	9.4%
Numerals Orientation	Upright	75.0%	9.4%
Numerals	12 numerals	63.9%	9.5%
Format	12 hours	91.7%	9.5%
Mirroring	No	94.4%	9.6%
Date	No	77.8%	9.7%
Month	No	77.8%	9.7%
Dial	White	58.3%	9.8%
Numerals Type	Arabic	72.2%	9.8%
Weekday	No	63.9%	10.0%
Border	No	11.1%	10.0%
Hands Size	Large	8.3%	10.3%
Hands	No seconds	72.2%	11.1%
Numerals	No numerals	16.7%	11.8%
Dial	Black	22.2%	12.3%
Graduations	No	22.2%	13.2%
Hands	No minutes	5.6%	23.6%

Figure 5. Time reading accuracy by feature (valid and invalid clocks)

3.6. **Models with the Greatest Number of Unique Correct Answers.** Models had a significant overlap in terms of clocks that they were able to read:

- Unique correct answers are cases where exactly one model answered a given clock correctly (for valid or invalid clocks)
- Smaller models tended to overlap with larger models or each other - with several models providing no correct answers unique to them

- Grok 4 technically provided the greatest number of unique correct answers (particularly on invalid clocks) due to labeling the majority of all clocks as invalid
- Most capable model overall - Gemini 2.5 Pro - also provided the greatest number of unique correct answers on valid clocks
- Mistral Medium 3.1 and Claude Opus 4.1 showed most relative divergent thinking, providing greatest number of unique answers out of the total correct answers

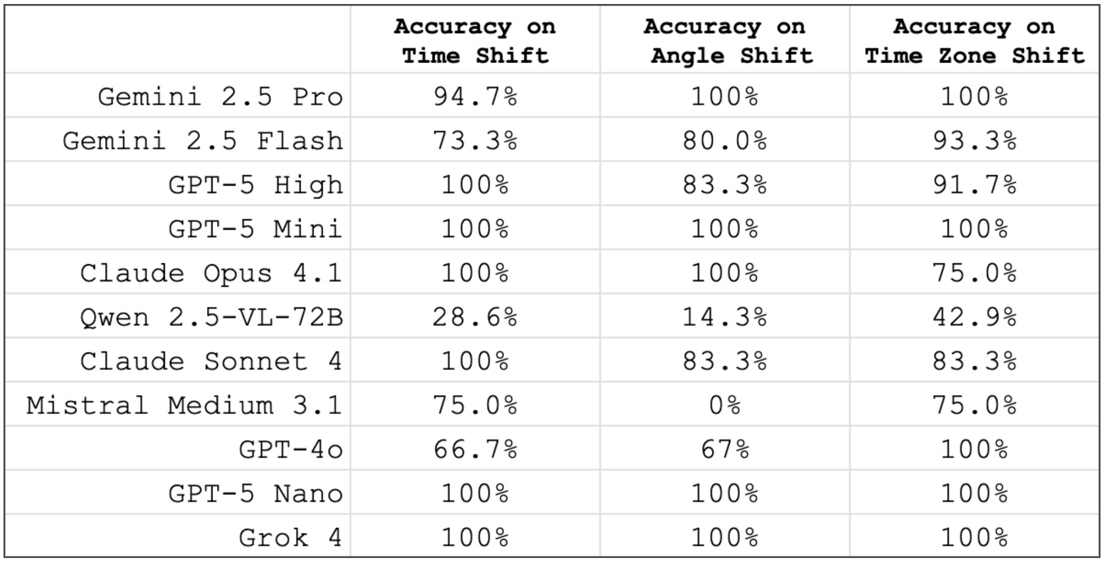


	Unique Correct Answers
Grok 4	8
Gemini 2.5 Pro	6
Mistral Medium 3.1	5
Claude Opus 4.1	4
GPT-5 High	3
GPT-5 Mini	1
Qwen 2.5-VL-72B	1
Claude Sonnet 4	0
Gemini 2.5 Flash	0
GPT-4o	0
GPT-5 Nano	0

Figure 6. Number of unique correct answers per model

3.7. **Accuracy on Follow-Up Questions.** Models showed strong capabilities in manipulating time in responses to the follow-up questions:

- Accuracy here was calculated as fraction of correct answers to follow-up questions out of valid clocks that were read correctly
- Best performing models were able to answer questions about adding or subtracting time, shifting hands by a certain degree or shifting to a different time zone with high accuracy, at times reaching 100%
- Overall larger reasoning models directionally outperformed smaller non-reasoning models, but the difference was smaller compared to the ability to read time accurately in the first place
- Qwen2.5-VL 72B was an outlier with a particularly poor ability to do any time manipulations
- Mistral Medium 3.1 was the only model that failed in one of categories entirely (angle shift), but this could be a function of a small sample size, given that the model had only 4 correct answers



	Accuracy on Time Shift	Accuracy on Angle Shift	Accuracy on Time Zone Shift
Gemini 2.5 Pro	94.7%	100%	100%
Gemini 2.5 Flash	73.3%	80.0%	93.3%
GPT-5 High	100%	83.3%	91.7%
GPT-5 Mini	100%	100%	100%
Claude Opus 4.1	100%	100%	75.0%
Qwen 2.5-VL-72B	28.6%	14.3%	42.9%
Claude Sonnet 4	100%	83.3%	83.3%
Mistral Medium 3.1	75.0%	0%	75.0%
GPT-4o	66.7%	67%	100%
GPT-5 Nano	100%	100%	100%
Grok 4	100%	100%	100%

Figure 7. Accuracy on follow-up questions

4. Limitations and Future Work

Key limitation that should be called out is the sample size. ClockBench is intended as an ongoing benchmark, and for the future iterations it would be beneficial to both continue adding to the pool of human participants and run a large number of cycles per model to further improve the accuracy and decrease potential volatility.

5. Conclusion

As this work shows that LLMs are still limited in their ability to read analog clocks compared to humans, an interesting question to ask is why. We have seen SOTA models showing strong reasoning skills, mathematical ability, and visual understanding on multiple benchmarks, but they do not seem to fully translate into reading analog clocks for now.

One hypothesis might be that reading analog clocks sets a high bar for doing reasoning within the visual space (as opposed to text space). I.e. reading random uncommon clocks might be both:

- Not represented in the training set enough for the models to simply memorize features and times
- Therefore potentially requires models to make connections and determine the relation of clock elements to each other via reasoning
- While at the same time clocks might be challenging objects to translate their representation fully into the text space to perform reasoning there

An optimistic viewpoint might be that despite the challenges above, best performing models seem to be capable of doing some complex visual reasoning (even if to a limited extent), consistently raising above random noise in reading time and median error size.

More research is likely needed to understand if these capabilities can be obtained by scaling existing paradigms, or a novel approach is required - similar to how test-time compute was key to unlocking progress in multiple domains.

References

1. Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhardt. Measuring Massive Multitask Language Understanding. arXiv:2009.03300. https://doi.org/10.48550/arXiv.2009.03300
2. David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. arXiv:2311.12022. https://doi.org/10.48550/arXiv.2311.12022
3. Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Summer Yue, Alexandr Wang, Dan Hendrycks, Center for AI Safety, Scale AI. Humanity's Last Exam. arXiv:2501.14249. https://doi.org/10.48550/arXiv.2501.14249
4. Philip and Hemang. SimpleBench: The Text Benchmark in which Unspecialized Human Performance Exceeds that of Current Frontier Models. https://simple-bench.com/
5. François Chollet. On the Measure of Intelligence. arXiv:1911.01547. https://doi.org/10.48550/arXiv.1911.01547
6. Rohit Saxena, Aryo Pradipta Gema, Pasquale Minervini. Lost in Time: Clock and Calendar Understanding Challenges in Multimodal LLMs. arXiv:2502.05092. https://doi.org/10.48550/arXiv.2502.05092