AI's Struggle to Read Analogue Clocks May Have Deeper Significance

Originally published at https://www.unite.ai/ais-struggle-to-read-analogue-clocks-may-have-deeper-significance/

A new paper from researchers in China and Spain finds that even advanced multimodal AI models such as GPT-4.1 struggle to tell the time from images of analog clocks. Small visual changes in the clocks can cause major interpretation errors, and fine-tuning only helps with familiar examples. The results raise concerns about the reliability of these models when faced with unfamiliar images in real-world tasks.

When humans develop a deep enough understanding of a domain, such as gravity or other basic physical principles, we move beyond specific examples to grasp the underlying abstractions. This allows us to apply that knowledge creatively across contexts and to recognize new instances, even those we have never seen before, by identifying the principle in action.

When a domain carries enough importance, we may even begin to perceive it *where it does not exist*, as with pareidolia, driven by the high cost of failing to recognize a real instance. So strong is this pattern-recognizing survival mechanism that it even disposes us to find a wider range of patterns where there are none.

The earlier and more repetitively a domain is instilled in us, the deeper its grounding and lifelong persistence; and one of the earliest visual datasets that we are exposed to as children comes in the form of teaching-clocks, where printed material or interactive analog clocks are used to teach us how to tell time:

Teaching aids to help children learn to tell time. Source: https://www.youtube.com/watch?v=IBBQXBhSNUs

Though changing fashions in watch design may sometimes challenge us, the resilience of this early domain-mastery is quite impressive, allowing us to discern analogue clock faces even in the face of complex or 'eccentric' design choices:

Some challenging faces in watch couture. Source: https://www.ablogtowatch.com/wait-a-minute-legibility-is-the-most-important-part-of-watch-design/

Humans do not need thousands of examples to learn how clocks work; once the basic concept is grasped, we can recognize it in almost any form, even when distorted or abstracted.

The difficulty that AI models face with this task, by contrast, highlights a deeper issue: their apparent strength may depend more on high-volume exposure than on understanding.

Beyond the Imitation Game?

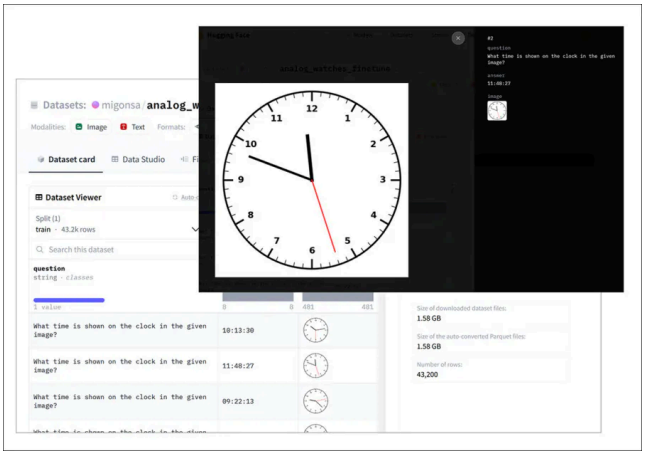
The tension between surface-level performance and genuine 'understanding' has surfaced repeatedly in recent investigations of large models. Last month Zhejiang University and Westlake University re-framed the question in a paper titled *Do PhD-level LLMs Truly Grasp Elementary Addition?* (not the focus of this article), concluding:

> *'Despite impressive benchmarks, models show critical reliance on pattern matching rather than true understanding, evidenced by failures with symbolic representations and violations of basic properties.*
>
> *'Explicit rule provision impairing performance suggests inherent architectural constraints. These insights reveal evaluation gaps and highlight the need for architectures capable of genuine mathematical reasoning beyond pattern recognition.'*

This week the question arises again, now in a collaboration between Nanjing University of Aeronautics and Astronautics and the Universidad Politécnica de Madrid in Spain. Titled *Have Multimodal Large Language Models (MLLMs) Really Learned to Tell the Time on Analog Clocks?*, the new paper explores how well multimodal models understand time-telling.

Though the progress of the research is covered only in broad detail in the paper, the researchers' initial tests established that OpenAI's GPT-4.1 multimodal language model struggled to correctly read the time from a diverse set of clock images, often giving incorrect answers even on simple cases.

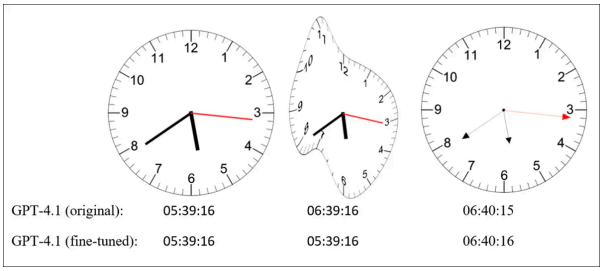
This points to a possible gap in the model's training data, raising the need for a more balanced dataset, to test whether the model can actually learn the underlying concept. Therefore the authors curated a synthetic dataset of analog clocks, evenly covering every possible time, and avoiding the usual biases found in internet images:



An example from the researchers' synthetic analog clock dataset, used to fine-tune a GPT model in the new work. Source: https://huggingface.co/datasets/migonsa/analog_watches_finetune

Before fine-tuning on the new dataset, GPT-4.1 consistently failed to read these clocks. After some exposure to the new collection, however, its performance improved – but only when the new images looked like ones it had already seen.

When the shape of the clock or the style of the hands changed, accuracy fell sharply; even small tweaks, such as thinner hands or arrowheads (rightmost image below), were enough to throw it off; and GPT-4.1 struggled additionally to interpret Dali-esque 'melting clocks':



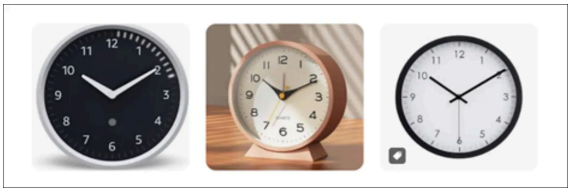
Clock images with standard design (left), distorted shape (middle), and modified hands (right), alongside the times returned by GPT-4.1 before and after fine-tuning. Source: https://arxiv.org/pdf/2505.10862

The authors deduce that current models such as GPT-4.1 may therefore be learning clock-reading mainly through *visual pattern matching*, rather than any deeper concept of time, asserting:

> *'[GPT 4.1] fails when the clock is deformed or when the hands are changed to be thinner and to have an arrowhead. The Mean Absolute Error (MAE) in the time estimate over 150 random times was 232.48s for the initial clocks, 1380.69s when the shape is deformed and 3726.93s when hands are changed.*
>
> *'These results suggest that the MLLM has not learned to tell the time but rather memorized patterns.'*

Enough Time

Most training datasets rely on scraped web images, which tend to repeat certain times – especially 10:10, a popular setting in watch advertisements:



From the new paper, an example of the prevalence of the 'ten past ten' time in analog clock images.

As a result of this limited range of times depicted, the model may see only a narrow range of possible clock configurations, limiting its ability to generalize beyond those repetitive patterns.

Regarding why models fail to correctly interpret the distorted clocks, the paper states:

> *'Although GPT-4.1 performs exceptionally well with standard clock images, it is surprising that modifying the clock hands by making them thinner and adding arrowheads leads to a significant drop in its accuracy.*
>
> *'Intuitively, one might expect that the more visually complex change – a distorted dial –would have a greater impact on performance, yet this modification seems to have a relatively smaller effect.*
>
> *'This raises a question: how do MLLMs interpret clocks, and why do they fail? One possibility is that thinner hands impair the model's ability to perceive direction, weakening its understanding of spatial orientation.*
>
> *'Alternatively, there could be other factors that cause confusion when the model attempts to combine the hour, minute, and second hands into an accurate time reading.'*

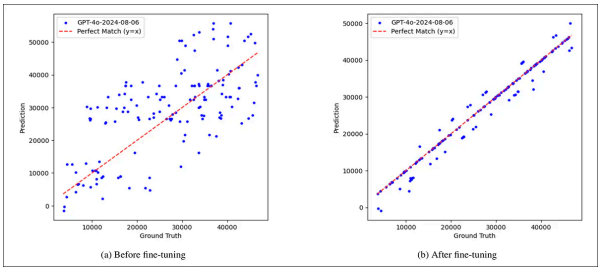
The authors contend that identifying the root cause of these failures is key to advancing multimodal models: if the issue lies in how the model perceives spatial direction, fine-tuning may offer a simple fix; but if the problem stems from a broader difficulty in integrating multiple visual cues, it points to a more fundamental weakness in how these systems process information.

Fine-Tuning Tests

To test whether the model's failures could be overcome with exposure, GPT-4.1 was fine-tuned on the aforementioned and comprehensive synthetic dataset. Before fine-tuning, its predictions were widely scattered, with significant errors across all types of clocks. After fine-tuning on the collection, accuracy improved sharply on standard clock faces, and, to a lesser extent, on distorted ones.

However, clocks with modified hands, such as thinner shapes or arrowheads, continued to produce large errors.

Two distinct failure modes emerged: on normal and distorted clocks, the model typically misjudged the direction of the hands; but on clocks with altered *hand styles*, it often confused the function of each hand, mistaking *hour* for *minute* or *minute* for *second*.



A comparison illustrating the model's initial weakness, and the partial gains achieved through fine-tuning, showing predicted vs. actual time, in seconds, for 150 randomly selected clocks. On the left, before fine-tuning, GPT-4.1's predictions are scattered and often far from the correct values, indicated by the red diagonal line. On the right, after fine-tuning on a balanced synthetic dataset, the predictions align much more closely with the ground truth, although some errors remain.

This suggests that the model had learned to associate visual features like hand thickness with specific roles, and struggled when these cues changed.

The limited improvement on unfamiliar designs raises further doubts about whether a model of this kind learns the abstract concept of time-telling, or merely refines its pattern-matching.

Hand Signs

So, although fine-tuning improved GPT-4.1's performance on conventional analog clocks, it had far less impact on clocks with thinner hands or arrowhead shapes, raising the possibility that the model's failures stemmed less from abstract reasoning and more from confusion over which hand was which.

To test whether accuracy might improve if that confusion were removed, a new analysis was conducted on the model's predictions for the 'modified-hand' dataset. The outputs were divided into two groups: cases where GPT-4.1 correctly recognized the hour, minute, and second hands; and cases where it did not.

The predictions were evaluated for Mean Absolute Error (MAE) before and after fine-tuning, and the results compared to those from standard clocks; angular error was also measured for each hand using dial position as a baseline:

Hand type	Error type	Number	MAE (Before fine-tuning)	MAE (After fine-tuning)
Modified	Confused	45	10088.9s	7937.1s
	Not confused	99	882.2s	595.8s
Normal	Confused	0	N.A.	N.A.
	Not confused	150	232.5s	33.8s

Error comparison for clocks with and without hand-role confusion in the modified-hand dataset before and after fine-tuning.

Confusing the roles of the clock hands led to the largest errors. When GPT-4.1 mistook the hour hand for the minute hand or vice versa, the resulting time estimates were often far off. In contrast, errors caused by misjudging the direction of a correctly identified hand were smaller. Among the three hands, the hour hand showed the highest angular error before fine-tuning, while the second hand showed the lowest.

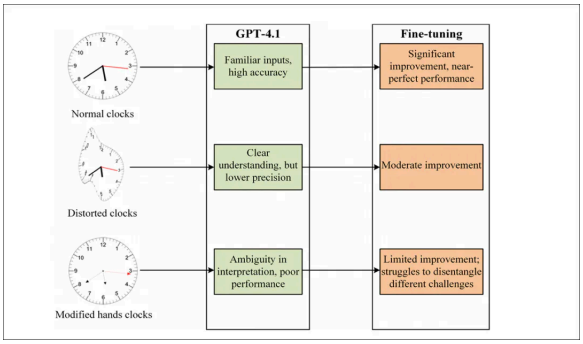
Hand type	Error type	Hand Role	MAE (Before fine-tuning)	MAE (After fine-tuning)
Modified	Confused	Hour	84.0°	66.7°
		Minute	80.9°	58.7°
		Second	24.9°	8.4°
	Not confused	Hour	7.3°	4.5°
		Minute	5.8°	6.7°
		Second	4.8°	1.0°
Normal	Not confused	Hour	2.4°	0°
		Minute	1.6°	1.0°
		Second	1.4°	1.9°

Angular error by hand type for predictions with and without hand-role confusion, before and after fine-tuning, in the modified-hand dataset.

To focus on directional errors alone, the analysis was limited to cases where the model correctly identified each hand's function. If the model had internalized a general concept of time-telling, its performance on these examples should have matched its accuracy on standard clocks. It did not, and accuracy remained noticeably worse.

To examine whether hand *shape* interfered with the model's sense of direction, a second experiment was run: two new datasets were created, each containing sixty synthetic clocks with only an hour hand, pointing to a different minute mark. One set used the original hand design, and the other the altered version. The model was asked to name the tick mark that the hand was pointing to.

Results showed a slight drop in accuracy with the modified hands, but not enough to account for the model's broader failures. A *single unfamiliar visual feature* appeared capable of disrupting the model's overall interpretation, even in tasks it had previously performed well.



Overview of GPT-4.1's performance before and after fine-tuning across standard, distorted, and modified-hand clocks, highlighting uneven gains and persistent weaknesses.

Conclusion

While the paper's focus may seem trivial at first glance, it does not especially matter if vision-language models ever learn to read analog clocks at 100% accuracy. What gives the work weight is its focus on a deeper recurring question: whether saturating models with more (and more diverse) data can lead to the kind of domain understanding humans acquire through abstraction and generalization; or whether the only viable path is to flood the domain with enough examples to anticipate every likely variation at inference.

Either route raises doubts about what current architectures are truly capable of learning.

First published Monday, May 19, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More