


ConTextual: Evaluating Context-Sensitive Text-Rich Visual Reasoning in Large Multimodal Models


Rohan Wadhawan*, Hritik Bansal*, Kai-Wei Chang, Nanyun Peng

TextVQA




Instruction: What word does the license plate say?
Ground Truth: French
GPT4 + OCR + IC Response: California French Protect Our Coast and Ocean

Infographic VQA



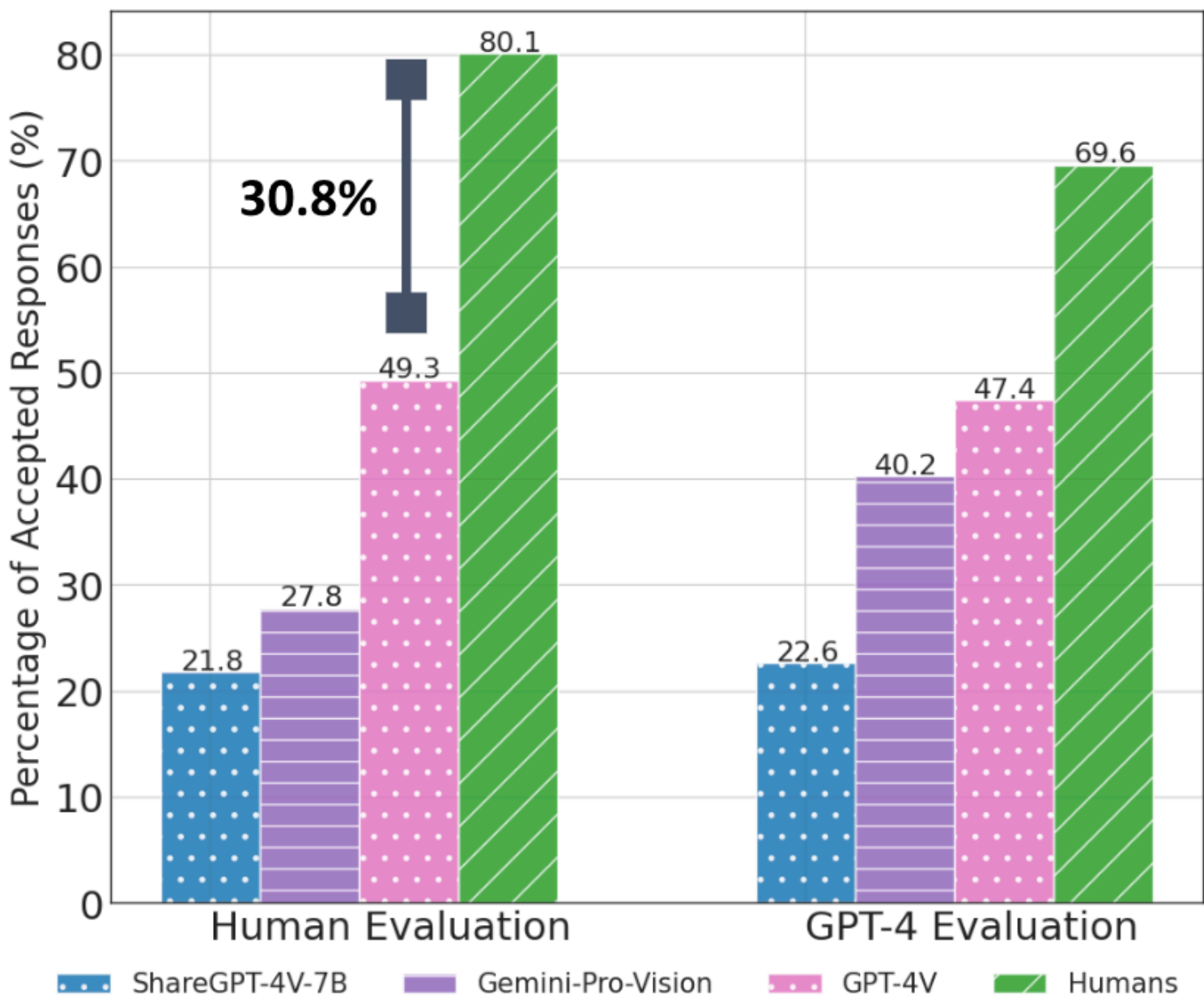
Instruction: What percent of rear-facing convertible seats are used incorrectly?
Ground Truth: 84%
GPT4 + OCR +IC Response: 84%

ConTextual



Instruction: Provide the cost of hiring the machine type depicted in the image for nine months.
Ground Truth: The machine shown in the image is a front-loading machine, and the rental fee for a period of 9 months amounts to INR 7200.
GPT4 + OCR + IC Response: The cost of hiring the depicted machine type (Top Load) for nine months is 4500 units

Contextual Overall Results and Data Statistics



Statistic	Number
Total number of samples	506
Root verbs in instructions	79
# Visual Contexts	8
- Time	50
- Shopping	50
- Navigation	50
- Abstract	50
- Application Usage	50
- Web Usage	50
- Infographic	50
- Miscellaneous Natural Scenes	156
Average Instruction Length	65
Average Response Length	117

Key Takeaways!

- GPT-4o, the best-performing LMM, struggles the most with time-reading and infographic visual contexts but outperforms humans in abstract contexts (memes, visual quotes, etc.).
- Open models perform poorly in most visual contexts except abstract and natural scenes, owing to the lack of diverse visual contexts in their training data.
- Both proprietary and open models have limited capacity for fine-grained visual perception, which is often required for context-sensitive reasoning in text-rich scenes.

Can Large Multimodal Models help you navigate a map, interpret charts, budget an outfit, organise your daily schedule and offer intuitive guidance in apps/websites?

To excel in these tasks, LMMs must understand the context in which the embedded text interacts with visual elements in text-rich images 🤔

GPT-4V, lags behind humans by 30.8% 😬




Paper, Data, Code, and Leaderboard are available publicly!




Data Example

Navigation



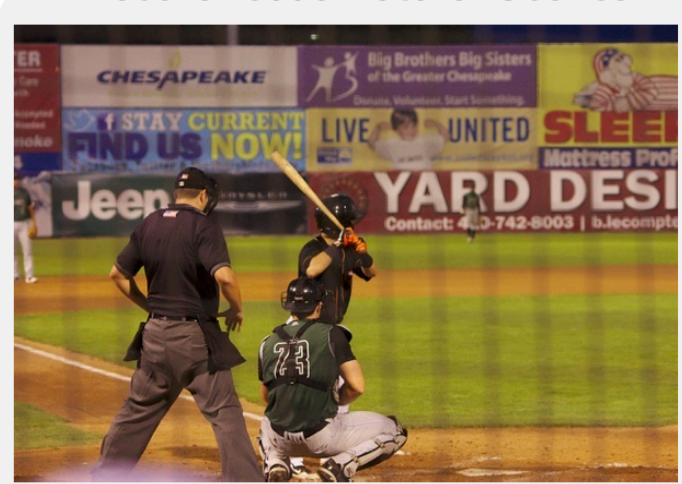
Group gates by direction.

Shopping



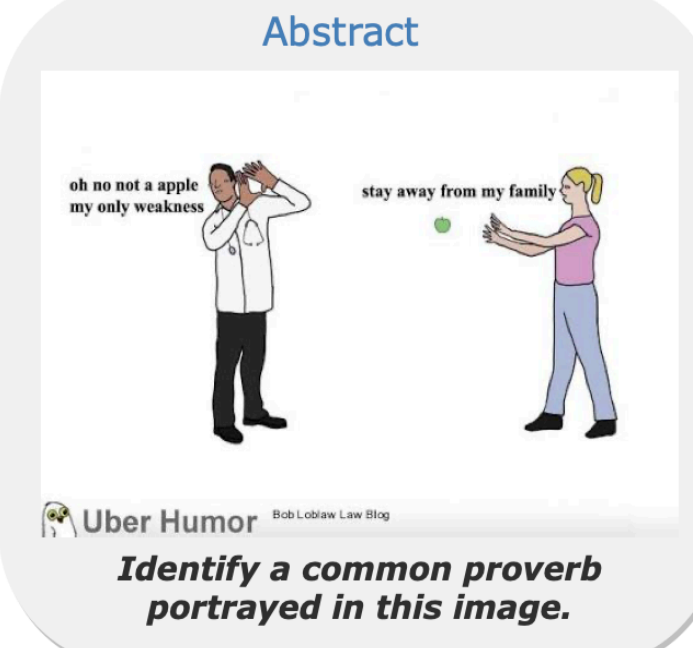
Calculate the total cost of assembling an outfit consisting of a white dress, heeled sandals, and sunglasses.

Miscellaneous Natural Scenes



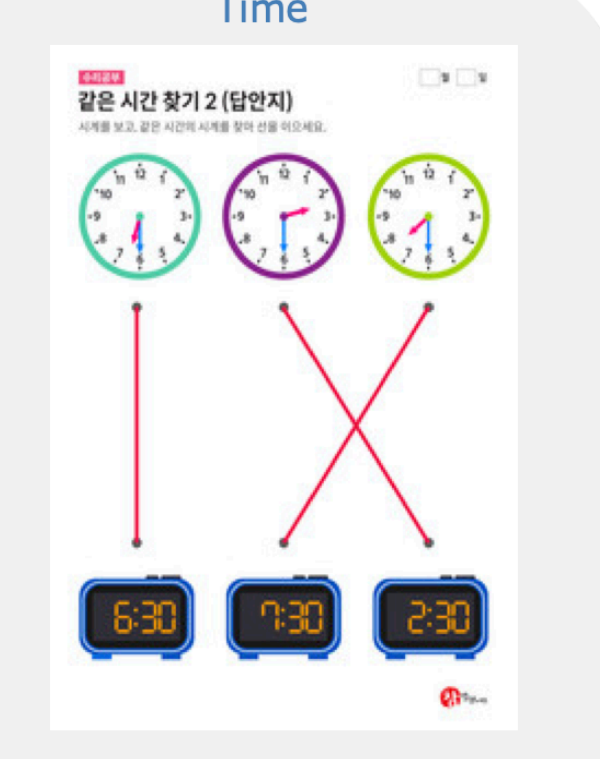
Determine the word obfuscated by the outfielder.

Abstract



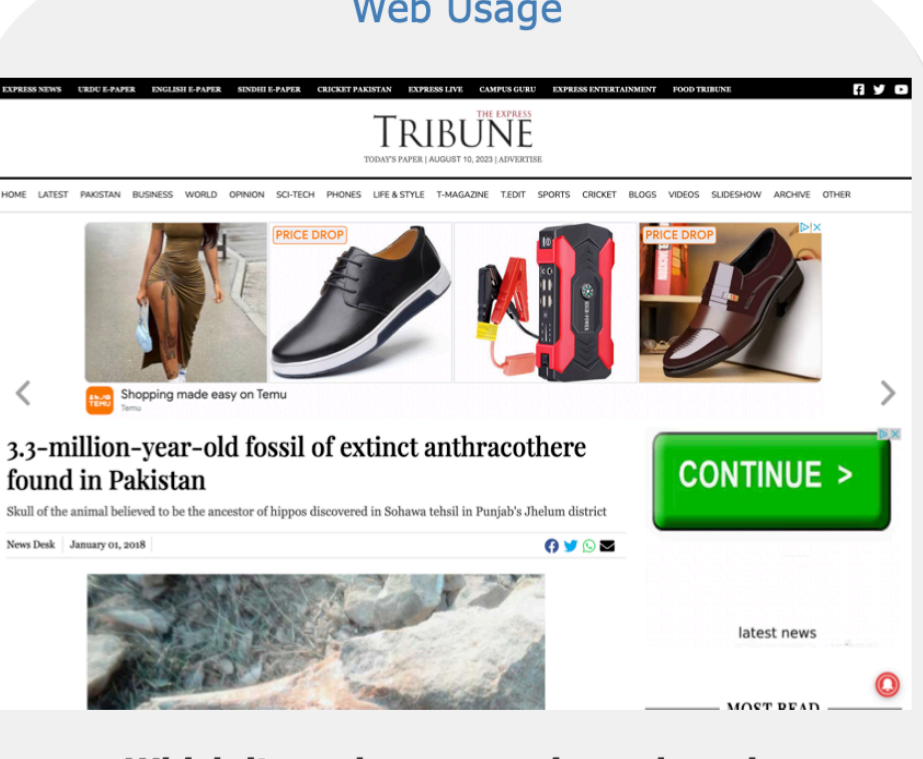
Identify a common proverb portrayed in this image.

Time



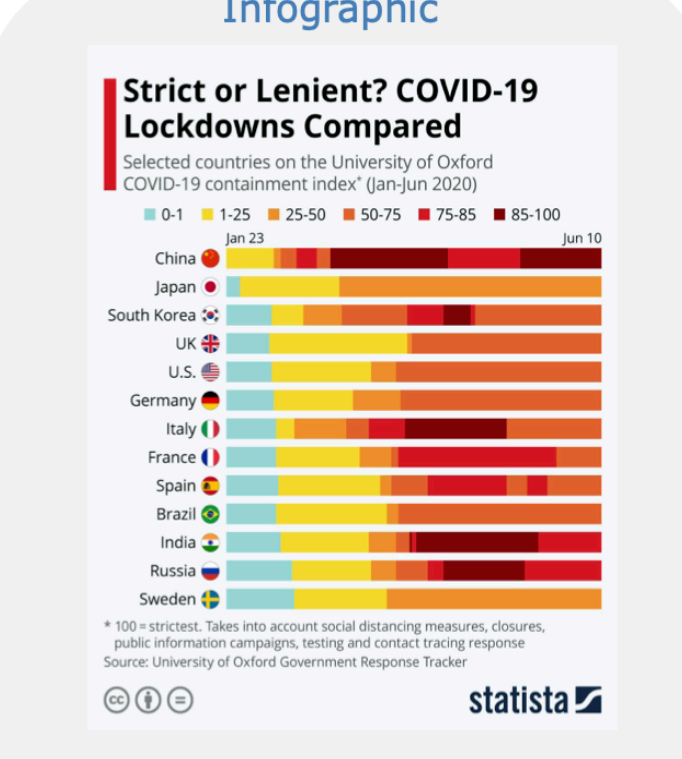
Comment whether the analog clocks are correctly matched to the digital clocks.

Web Usage



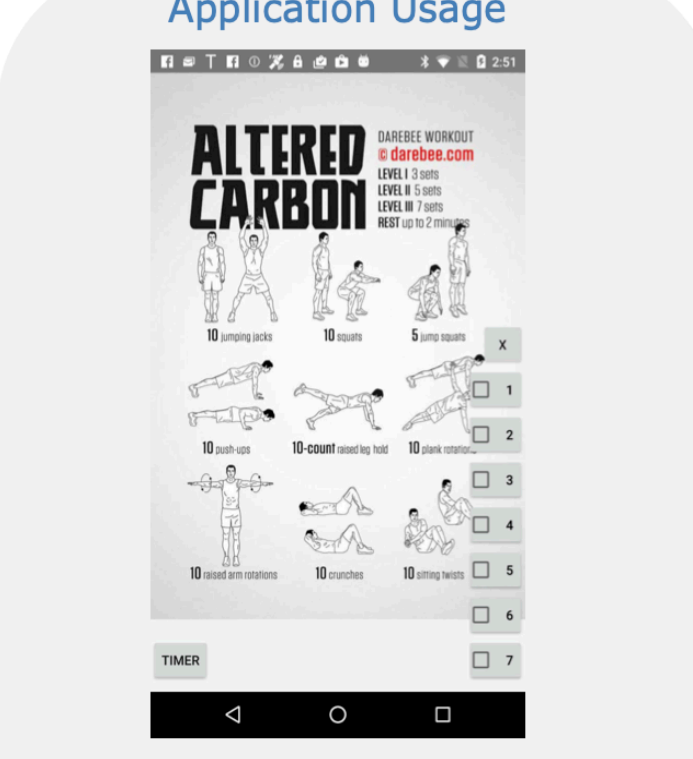
Which items have experienced a price drop?

Infographic



Which countries have had a non-zero containment index that lies between 95-100?

Application Usage



List the exercises where the corresponding illustration showcases a single movement.

Leaderboard (GPT-4 eval)

#	Model	Method	Source	Date	ALL	Time	Shop.	Nav.	Abs.	App.	Web.	Info.	Misc. NS.
-	Human Performance	-	Link	2024-01-24	69.6	64.0	64.0	73.5	75.5	64.0	58.0	72.0	78.0
1	GPT-4o 🏆	LMM	Link	2024-05-18	62.8	32.0	70.0	60.0	98.0	72.0	62.0	48.0	64.7
2	GPT-4o-mini-2024-07-18 🏆	LMM	Link	2024-07-18	61.7	22.0	62.0	62.0	98.0	72.0	64.0	42.0	67.3
3	Claude-3.5-Sonnet-2024-06-20 🏆	LMM	Link	2024-07-18	57.5	22.0	52.0	66.0	96.0	68.0	64.0	44.0	56.7
4	Gemini-1.5-Flash-Preview-0514	LMM	Link	2024-05-18	56.0	30.0	51.0	52.1	84.0	63.0	63.2	42.8	61.7
5	Gemini-1.5-Pro-Preview-0514	LMM	Link	2024-05-18	52.4	24.0	46.9	39.6	84.0	45.8	59.2	43.8	64.0
6	GPT-4V(sion)	LMM	Link	2024-01-24	47.4	18.0	54.0	48.0	100.0	48.0	42.0	28.0	48.0
7	Gemini-Pro-Vision	LMM	Link	2024-01-24	40.2	16.0	32.7	28.6	65.3	44.9	43.8	20.0	52.8
8	Claude-3-Opus-2024-02-29	LMM	Link	2024-03-05	38.1	18.0	32.0	34.0	68.0	44.0	38.0	18.0	44.7
9	LLaVA-Next-34B	LMM	Link	2024-03-05	36.8	10.0	36.0	30.6	66.0	36.0	28.0	12.0	51.3
10	LLaVA-Next-13B	LMM	Link	2024-03-05	30.3	0.0	28.6	32.0	60.0	18.0	32.0	10.0	40.4
11	ShareGPT-4V-7B	LMM	Link	2024-01-24	22.6	0.0	16.0	20.0	28.6	20.0	20.0	14.0	37.7
12	GPT-4 w/ Layout-aware OCR + Caption	LLM	Link	2024-01-24	22.2	6.0	16.0	24.0	57.1	14.0	18.0	8.0	27.3
13	Qwen-VL	LMM	Link	2024-01-24	21.8	4.0	20.0	24.0	53.1	6.0	18.0	14.0	27.3
14	LLaVA-1.5B-13B	LMM	Link	2024-01-24	20.8	4.0	10.0	18.0	44.9	16.0	26.0	4.0	29.7
15	mPLUG-Owl-v2-7B	LMM	Link	2024-01-24	18.6	4.0	8.0	24.0	32.7	20.0	10.0	12.0	26.0
16	GPT-4 w/ Layout-aware OCR	LLM	Link	2024-01-24	18.2	8.0	20.0	18.0	34.7	10.0	16.0	16.0	20.7
17	GPT-4 w/ OCR*	LLM	Link	2024-01-24	15.9	4.0	10.0	14.0	30.6	8.0	16.0	28.6	16.9
18	LLaVAR-13B	LMM	Link	2024-01-24	14.9	10.0	16.0	6.0	44.9	8.0	10.0	6.0	16.7
19	BLIVA	LMM	Link	2024-01-24	10.3	2.0	4.0	14.0	24.5	4.0	8.0	4.0	14.7
20	InstructBLIP-Vicuna-7B	LMM	Link	2024-01-24	9.7	2.0	4.0	16.0	20.0	6.0	12.0	2.1	12.0
21	Idefics-9B	LMM	Link	2024-01-24	7.7	4.0	2.0	12.0	12.0	0.0	6.0	2.0	13.3

Qualitative Analysis



Instruction: Pick out the airlines immediately in front and behind the one featuring a flying bird in yellow and blue.

Reference Response: The airline right in front of the airline with a bird logo is SriLankan Airlines, and the one situated behind it is Air France.

Human Response: Sri Lankan and Air France.

GPT-4V Response: The airline with a flying bird on its fuselage is "SriLankan Airlines." Directly in front of it, the airplane features the livery of "Lufthansa," and right behind it is the airplane with a dark blue tail and text "Condor" representing Condor Airlines.



Instruction: Identify the object on the dial that shows the time 10:10

Reference Response: The clock at 10:10 features an octopus on its dial.

GPT-4V Response: The clock with the image of an octopus on the face is showing the time 10:10.

LLaVA-Next-34B Response: The object on the dial that shows the time 10:10 is the hour hand and the minute hand

GPT-4 w/ Layout-aware OCR + Caption Response: The instruction does not provide enough specific information to identify the object on the dial that shows the time 10:10