Input files

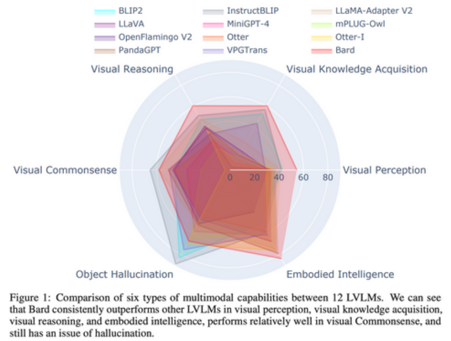


Figure 1: Comparison of six types of multimodal capabilities between 12 LVLMs. We can see that Bard consistently outperforms other LVLMs in visual perception, visual knowledge acquisition, visual reasoning, and embodied intelligence, performs relatively well in visual Commonsense, and still has an issue of hallucination.

AgentX_181.jpg

Query

Which model performs the best in Visual Knowledge Acquisition, and what is the total number of distinct models shown in the radar chart?

Reasoning steps: 5

			Tools used
STEP 1	TASK	Describe the image to understand the content and context.	SCENE DESCRIBER
	THOUGHT	Understanding the image structure helps guide the following steps to identify which model performs best and count the number of distinct models.	
STEP 2	TASK	Identify the model performing best in Visual Knowledge Acquisition.	LOCATE OBJECTBYTEXT
	THOUGHT	To find the best-performing model, I need to locate the highest value in the 'Visual Knowledge Acquisition' axis among all models.	
STEP 3	TASK	Overlay output to identify the best performer clearly.	OVERLAY TEXT
	THOUGHT	Using coordinates, I can locate the peak performance on the specific axis and identify the top model.	
STEP 4	TASK	Count the total number of distinct models represented in the chart.	OCR
	THOUGHT	Recognizing and listing out each model name ensures accurate counting.	
STEP 5	TASK	Compute the total number of distinct models.	CALCULATOR
	THOUGHT	Calculating the length of the list from OCR results gives the number of distinct models.	

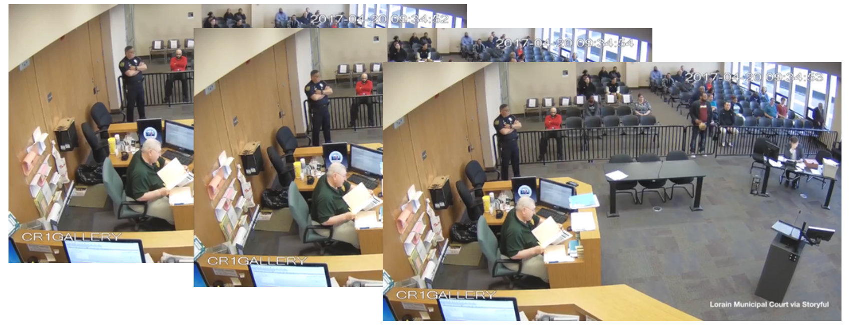
GT Answer

{'best_model': 'Bard', 'total_models': 12}

Justification

Bard was identified as the model performing best in Visual Knowledge Acquisition based on its position in the radar chart. The OCR tool was used to enumerate and count all distinct models, resulting in a total of 12.

Input files



AgentX_629.mp4

Query

What place is shown in the video, and how many people are performing duties?

Reasoning steps: 3

			Tools used
STEP 1	TASK	Identify the scene using the video.	SCENE DESCRIBER
	THOUGHT	SceneDescriber provides a general overview of the setting to identify that it's a courtroom.	
STEP 2	TASK	Determine the location of the video scene from text in the image.	OCR
	THOUGHT	The OCR tool scanned the image for text and identified the location as Lorain Municipal Court.	
STEP 3	TASK	Count the number of people performing duties in the video.	OBJECT COUNTER
	THOUGHT	The ObjectCounter tool detects how many people are present who appear to be performing duties in the courtroom scene.	

GT Answer

Lorain Municipal Court, 3

Justification

Using the SceneDescriber and OCR tools, I identified the scene as Lorain Municipal Court. The ObjectCounter tool confirmed three people performing duties: a judge, security officer, and lawyer, while the rest appear to be seated as the audience.