上海科技大学
ShanghaiTech University

HALLUSIONBENCH: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models

University of Maryland, College Park

立志成才报国裕民

Introduction

While LVLMs like GPT-4V(ision) and LLaVA-1.5 excel in various applications, they are hindered by a pronounced language bias. This bias stems from in- stances where knowledge priors conflict with the visual context.

When exploring those LVLMs, we observe that their strong language bias often overshadows visual information, leading to an overreliance on language priors rather than the visual context. To study this phenomenon, we use the term "Language Hallucination," which refers to conclusions drawn without visual input. On the other hand, the vision components within the limited ability in LVLMs can give rise to "Visual Illusion", where visual inputs can be misinterpreted, leading to overconfident yet erroneous assertions by the model.

Visual Question Taxonomy

Our aim is to develop a multimodal image-context reasoning benchmark to investigate the potent language bias inherent in LVLMs, which can sometimes overshadow the visual context. We define the two categories of visual questions: Visual Dependent and Visual Supplement.

The Visual Dependent questions are defined as questions that do not have an affirmative answer without the visual context.

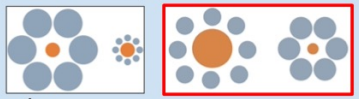
The Visual Supplement questions are questions that can be answered without the visual input; the visual component merely provides supplemental information or corrections.

Hallusionbench

上海科技大学
ShanghaiTech University

Visual Dependent

Illusion



Question:
Is the right orange circle the same size as the left orange circle?
Is the right orange circle larger than the left orange circle?
Is the right orange circle smaller than the left orange circle?

Math

Question:
According to parallel lines theorem, is angle 1 + angle 2 > 180 ?
According to parallel lines theorem, is angle 1 + angle 2 = 180 ?
According to parallel lines theorem, is angle 1 + angle 2 < 180 ?

Poster

Question:
Does the image show "Beijing Roast Duck"?
Does the image show "Guangxi Roast Duck"?

Figure / Other

Question:
Are all the characters in this figure from the manga series One Piece?
Are there any characters in this figure from the manga series Detective Conan?

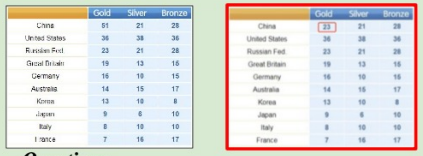
Video

Question:
According to the positive sequence images, does Homer Simpson disappear into the bushes?
According to the positive sequence images, does Homer Simpson come out of the bushes?
Homer Simpson disappears into the bushes. According to the positive sequence, are they in the correct order?
Homer Simpson comes out of the bushes. According to the positive sequence, are they in the correct order?

Visual Supplement

Table

No Visual



Question:
Does China have the most gold medals in 2008 beijing olympic?
Does USA have the most gold medals in 2008 beijing olympic?
Does Russia have the most gold medals in 2008 beijing olympic?

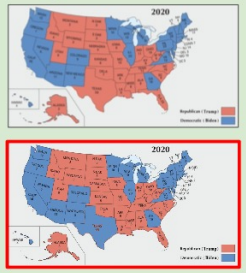
Chart

No Visual

Question:
In 2017, was Tencent the company with the highest revenue from video games, with Sony as the second-highest earner?
In 2017, did Apple generate higher revenue from video games compared to Google?

Map

No Visual



Question:
Based on the map, did the Democratic Party win Texas in the 2020 elections?
Based on the map, did the Republican Party win Texas in the 2020 elections?

OCR

No Visual

$$G \approx 6.67428 \times 10^{-11}\, m^3\, kg^{-1}\, s^{-2}$$

$$G \approx 6.6\ 9428\ \times 10^{-11}\, m^3\, kg^{-1}\, s^{-2}$$

Question:
*According to the image, does the value of Gravity constant 'G' range from 6.66 * 10^-11 to 6.68 * 10^-11?*
*According to the image, does the value of Gravity constant 'G' range from 6.68 * 10^-11 to 6.70 * 10^-11?*

Figure 1. Data samples of HALLUSIONBENCH, which contains diverse topics, visual modalities. Human-edited images are in RED, resulting in different correct answers to the questions.

Visual, Question, and Annotation Structures

Notations: Let $(I,q)\in\mathcal{V}\subseteq\mathcal{I}\times\mathbb{Q}$ be the tuple of the image $I\in\mathbb{I}$ and question $q\in\mathbb{Q}$, where ν is the set of valid VQ pairs. Let N be the number of original images obtained from the Internet, and $\mathbb{I}_o = \{I_{(i,0)}\}_{0<i\leq N}$ be the set of those original images. We define $\mathbb{I}_i' = \{I_{(i,j)}\}_{0<j\leq N_i}$ be the set of images modified from $I_{(i,0)}$, and I_0 be an empty image.

The entire images set $\mathbb{I} = \{I_0\}\cup\mathbb{I}_o\cup(\cup_{0<i\leq N}\ \mathbb{I}_i')$

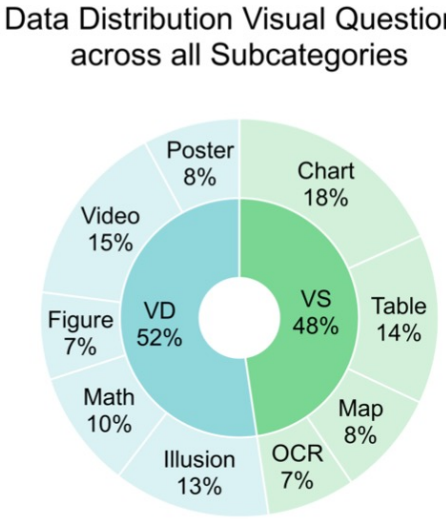
Let $\mathbb{Q}_i = \{q_{(i,k)}\}_{0<k\leq M_i}$ be the set of questions that can be applied to any image in I_i , which is defined differently for Visual Dependent (VD) and Visual Supplement (VS):

$$\mathbb{I}_i = \begin{cases} \{I_{(i,0)}\}\cup\mathbb{I}_i' & \text{for } VD \\ \{I_0, I_{(i,0)}\}\cup\mathbb{I}_i' & \text{for } VS \end{cases}$$

Dataset Statistics

上海科技大学
ShanghaiTech University

		No Visual	Original Visual	Edited Visual	Overall	
Visual Dependent	*Illusion*	-	72	72	144	591
	Math	-	54	54	108	
	Video	-	69	101	170	
	Poster	-	43	46	89	
	Others	-	39	41	80	
Visual Supplement	*Chart*	76	68	62	206	538
	Table	43	43	69	155	
	Map	32	32	32	96	
	OCR	27	27	27	81	
Overall		178	447	504	1129	



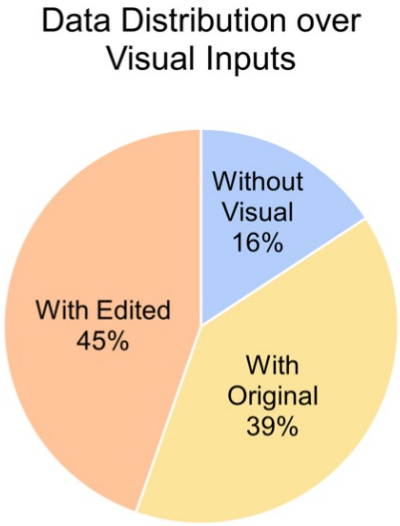
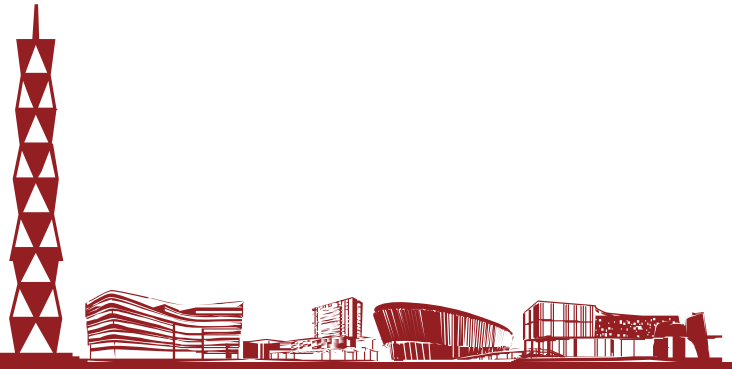


Figure 2. **Statistics of HALLUSIONBENCH:** We show the number of questions in the table (*left*), and the distribution of visual questions across each subcategory of Visual Dependent (VD) and Visual Supplement (VS) (*middle*) and visual input types categorized by no visual, original, and edited images (*right*). HALLUSIONBENCH covers a diverse visual format and nearly half of the images are manually edited.



立志成才报国裕民

Uniqueness of HALLUSIONBENCH

上海科技大学
ShanghaiTech University

Provide more topics, more image types

Include human-edited images to assess the robustness of current LVLMs

Focus on evaluating both language hallucinations and visual illusions

Benchmarks	Visaul Format	# Total QA	# H-Edited QA	# Total Img.	# H-Edited Img.	Control Pair?	Purpose
Lynx-Bench [56]	Image,Video	450	450	450	0	✗	Image&Video QA Evaluation
SciGraphQA [22]	Image	295K	0	657K	0	✗	Scientific Chart QA Evaluation
MathVista [34]	Image	6141	0	5487	0	✗	Math Reasoning Evaluation
MME [14]	Image	1457	1457	1187	0	✗	Comprehensive Evaluation
POPE [23]	Image	3000	0	500	0	✗	Object Hallucination
M-HalDetect [18]	Image	4000	0	4000	0	✗	Object Hallucination
GAVIE [28]	Image	1000	0	1000	0	✗	Object Hallucination
Bingo [10]	Image	370	370	308	N/A	✓	Hallucination, Bias
HALLUSIONBENCH	Image, Video **Image Pairs**	1129	1129	346	**181**	✓	**Visual Illusion, Language Hallucination, Quantitative Analysis and Diagnosis**

Table 1. **Comparison of HALLUSIONBENCH with most recent VL benchmarks:** HALLUSIONBENCH is the **first** and the **only** benchmark that focuses on control-group analysis by carefully editing each image in the database manually. "# H-Edited QA" means Human-edited question-answer pairs. "# H-Edited Img" means Human-edited images. *N/A* denotes that the information is not provided.

Text-Only GPT4-Assisted Evaluation

For each sample, we fill the template with its question, ground truth, and LVLM output. By taking the filled prompt into GPT-4, GPT-4 will generate "correct", "incorrect" or "unclear" for the sample. We utilize GPT-4 to evaluate the outputs of LLMs 3 times and report average scores.

Prompt:

Imagine you are an intelligent teacher. Thoroughly read the question, reference answer, and the prediction answer to ensure a clear understanding of the information provided. Assess the correctness of the predictions. If the prediction answer does not conflict with the reference answer, please generate “correct”. If the prediction answer conflicts with the reference answer, please generate “incorrect”. If the pre- diction answer is unclear about the answer, please generate "unclear".

Visual, Question, and Annotation Structures

$$b_{\mathcal{M}}(I,q) = \begin{cases} GPT(\mathcal{M}(I,q), y(I,q)) & \text{if} GPT(\mathcal{M}, y) \leq 1 \\ 1 & \text{else if } I = I_0 \\ 0 & \text{otherwise} \end{cases},$$

All accuracy: $aAcc = \frac{\sum_{(I,q)\in\mathcal{V}} b_{\mathcal{M}}(I,q)}{|\mathcal{V}|}$

Figure Accuracy: $fAcc = \frac{\sum_{i,j} 1(\Lambda_{q\in\mathbb{Q}_i} b_{\mathcal{M}}(I_{(i,j)},q))}{|\mathbb{I}|}$

Question Pair Accuracy: $qAcc = \frac{\sum_{i,k} 1(\Lambda_{I\in\mathbb{I}_i} b_{\mathcal{M}}(I,q_{(i,k)}))}{|\mathbb{Q}|}$

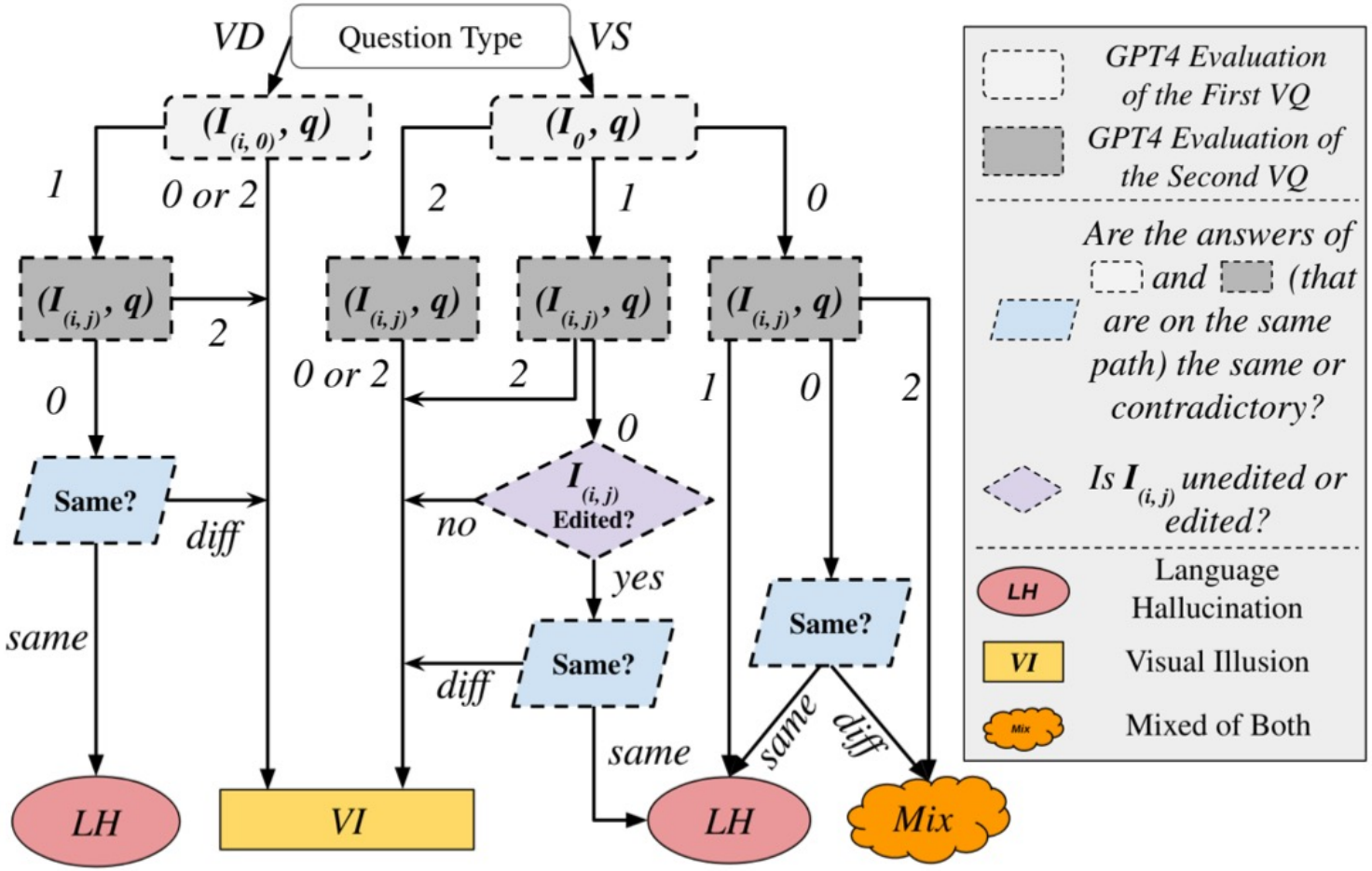
Yes / No Bias Test

Some models tend to respond with "yes" in most cases. No further analysis is necessary if the model has a very strong bias or tendency to answer one way regardless of the actual question

$$d_y = \frac{\sum_{(I,q)\in\mathcal{V}} \left[\mathbb{1}\left(\mathcal{M}(I,q) = \text{“yes”}\right) - \mathbb{1}\left(y(I,q) = \text{“yes”}\right)\right]}{|\mathcal{V}|}$$

$$r_{fp} = \frac{\sum_{(I,q)\in\mathcal{W}} \mathbb{1}\left(\mathcal{M}(I,q) = \text{“yes"}\right)}{|\mathcal{W}|},$$

Diagnostic Test



Model

We conduct massive experiments on HALLUSIONBENCH to evaluate a total of 15 LVLMs, including GPT-4V, LLaVA-1.5, Gemini Pro Vision, Claude 3, MiniGPT4, MiniGPT5, GiT, InstructBLIP, Qwen-VL, mPLUG-Owl-v1, mPLUG-Owl-v2, LRV-Instruction, BLIP2, BLIP2-T5, and Open-Flamingo.

We also include Random Chance (i.e. randomly choose Yes or No) as a baseline.

Results

上海科技大学
ShanghaiTech University

Method	# Parameter	Evaluation	Question Pair Accuracy (*qAcc*) ↑	Figure Accuracy (*fAcc*) ↑	Easy Accuracy (Easy *aAcc*) ↑	Hard Accuracy (Hard *aAcc*) ↑	All Accuracy (*aAcc*) ↑
GPT4V [1] (Oct 2023)	-	Human	31.42	44.22	79.56	38.37	67.58
		GPT4-Assisted	**28.79**	**39.88**	**75.60**	37.67	**65.28**
LLaVA-1.5 [31]	13B	Human	9.45	25.43	50.77	29.07	47.12
		GPT4-Assisted	10.55	**24.86**	**49.67**	29.77	46.94
Claude 3 [38]	-	GPT4-Assisted	**21.76**	**28.61**	**55.16**	**41.40**	**56.86**
Gemini Pro Vision [39] (Dec 2023)	-	GPT4-Assisted	7.69	8.67	35.60	30.23	36.85
BLIP2-T5 [21]	12.1B	GPT4-Assisted	**15.16**	20.52	45.49	**43.49**	**48.09**
Qwen-VL [6]	9.6B	GPT4-Assisted	5.93	6.65	31.43	24.88	39.15
Open-Flamingo [3]	9B	GPT4-Assisted	6.37	11.27	39.56	27.21	38.44
MiniGPT5 [62]	8.2B	GPT4-Assisted	10.55	9.83	36.04	28.37	40.30
MiniGPT4 [63]	8.2B	GPT4-Assisted	8.79	10.12	31.87	27.67	35.78
InstructBLIP [11]	8.2B	GPT4-Assisted	9.45	10.11	35.60	**45.12**	45.26
BLIP2 [21]	8.2B	GPT4-Assisted	5.05	12.43	33.85	40.70	40.48
mPLUG_Owl-v2 [51]	8.2B	GPT4-Assisted	13.85	19.94	44.84	39.07	47.30
mPLUG_Owl-v1 [50]	7.2B	GPT4-Assisted	9.45	10.40	39.34	29.77	43.93
LRV_Instruction [28]	7.2B	GPT4-Assisted	8.79	13.01	39.78	27.44	42.78
GIT [44]	0.8B	GPT4-Assisted	5.27	6.36	26.81	31.86	34.37
Random Chance	-	GPT4-Assisted	15.60	18.21	39.12	39.06	45.96

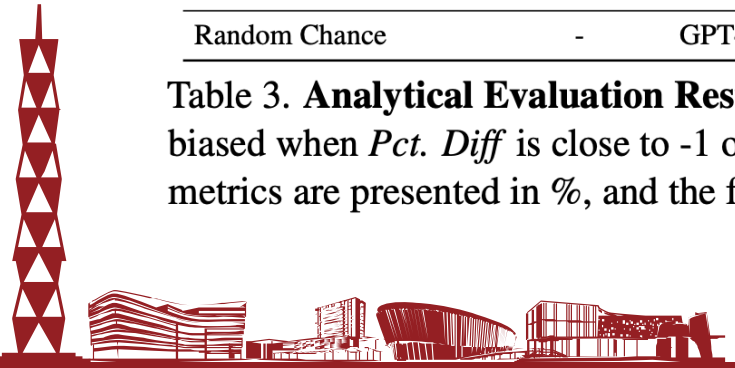
Table 2. **Correctness Leaderboard on HALLUSIONBENCH with various LVLMs:** All the numbers are presented in % and the full score is 100%. Hard questions refer to the edited images. We highlight the Top 3 models with the GPT4-assisted evaluation.

Results

上海科技大学
ShanghaiTech University

Method	# Parameter	Evaluation	Yes/No Bias		Consistency			Language and Vision Diagnosis		
			Pct. Diff (∼ 0)	FP Ratio (∼ 0.5)	Correct ↑	Inconsistent ↓	Wrong ↑	Language Hallucination	Visual Illusion	Mixed
GPT4V [1] (Oct 2023)	-	Human	0.066	0.60	44.22	32.66	23.12	21.86	46.17	31.97
		GPT4-Assisted	**0.058**	0.58	**39.88**	**38.15**	**21.97**	22.19	45.66	32.14
LLaVA-1.5 [31]	13B	Human	0.27	0.76	25.43	42.49	32.08	25.63	51.42	22.95
		GPT4-Assisted	0.26	0.75	**24.86**	**45.38**	29.77	26.71	51.09	22.20
Claude 3 [38]	-	GPT4-Assisted	0.063	**0.57**	**28.61**	**49.42**	**21.97**	19.10	59.14	21.77
Gemini Pro Vision [39] (Dec 2023)	-	GPT4-Assisted	**-0.02**	**0.48**	8.67	56.94	34.39	25.95	49.37	24.68
BLIP2-T5 [21]	12.1B	GPT4-Assisted	0.08	0.58	20.52	59.54	19.94	41.64	40.44	17.92
Qwen-VL [6]	9.6B	GPT4-Assisted	0.12	0.60	6.65	50.29	43.06	0.87	88.06	11.06
Open-Flamingo [3]	9B	GPT4-Assisted	0.33	0.77	11.27	59.83	28.90	30.07	48.06	21.87
MiniGPT5 [62]	8.2B	GPT4-Assisted	0.28	0.71	9.83	56.36	33.82	10.09	73.44	16.47
MiniGPT4 [63]	8.2B	GPT4-Assisted	0.19	0.65	10.12	57.80	32.08	23.59	56.55	19.86
InstructBLIP [11]	8.2B	GPT4-Assisted	-0.13	0.38	10.12	68.50	**21.39**	29.29	54.53	16.18
BLIP2 [21]	8.2B	GPT4-Assisted	0.18	0.65	12.43	63.01	24.57	39.14	43.45	17.41
mPLUG_Owl-v2 [51]	8.2B	GPT4-Assisted	0.25	0.77	19.94	58.09	21.97	28.24	50.42	21.34
mPLUG_Owl-v1 [50]	7.2B	GPT4-Assisted	0.32	0.79	10.40	60.12	29.48	3.95	78.36	17.69
LRV_Instruction [28]	7.2B	GPT4-Assisted	0.26	0.73	13.01	53.47	33.53	4.49	76.47	19.04
GIT [44]	0.8B	GPT4-Assisted	**0.04**	**0.53**	6.36	53.76	39.88	30.90	58.30	10.80
Random Chance	-	GPT4-Assisted	0.08	0.57	18.20	57.51	24.28	-	-	-

Table 3. **Analytical Evaluation Results on HALLUSIONBENCH with various LVLMs:** *Pct. Diff* ranges from [-1, 1]. The model is more biased when *Pct. Diff* is close to -1 or 1. *FP Ratio* ranges from [0, 1]. The model is more robust when *FP Ratio* is close to 0.5. All the other metrics are presented in %, and the full score is 100%. We highlight the Top 3 models with the GPT4-assisted evaluation.



立志成才报国裕民

Results

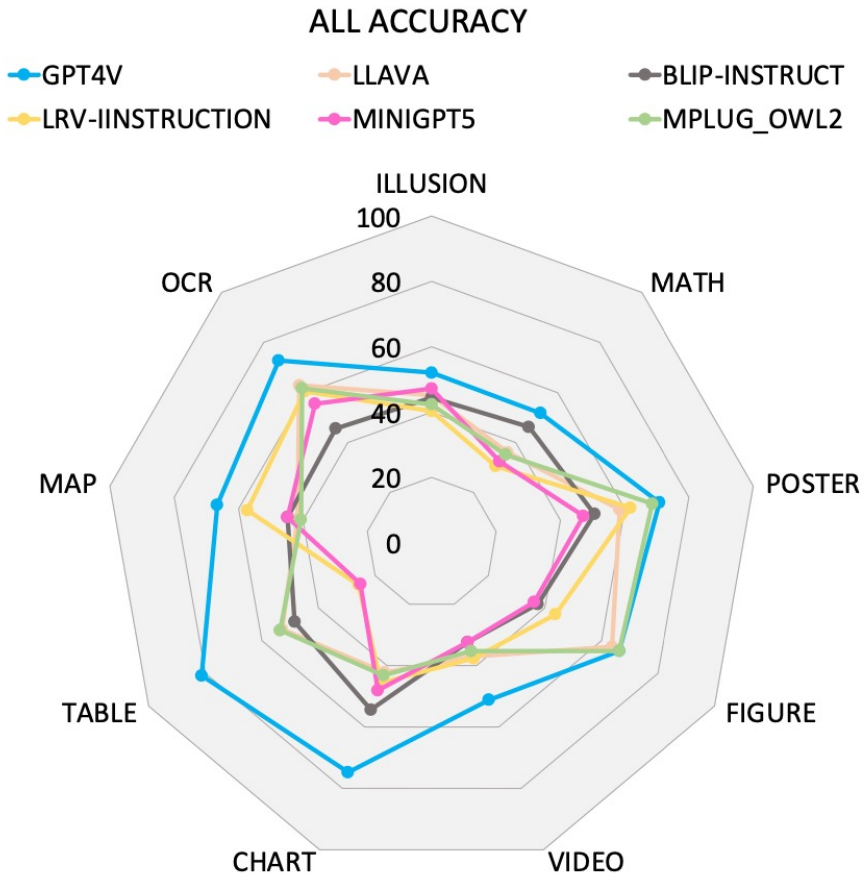


Figure 4. **Accuracies on each subcategories:** We show six prominent LVLMs on HALLUSIONBENCH across different types.