MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark

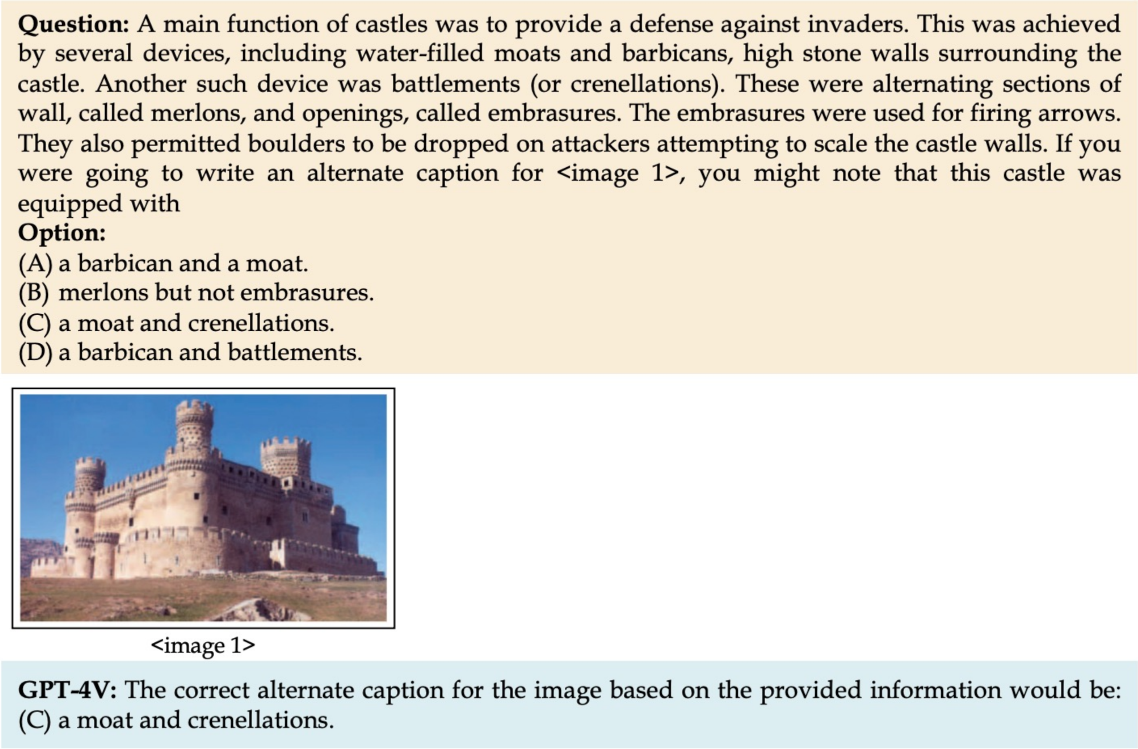
ICML 2024 Oral

Dongping Chen[1*], Ruoxi Chen[2*], Shilin Zhang[1*], Yaochen Wang[1*], Yinuo Liu[1*], Huichi Zhou[1*],
Qihui Zhang[1*], **Yao Wan**[1†], **Pan Zhou**[1†], **Lichao Sun**[3]

Motivation

Evaluating free-form query is difficult

- Traditional benchmarks consist of MCQA, whereas free-form queries are more common in everyday scenarios and are difficult to assess even with a golden answer.



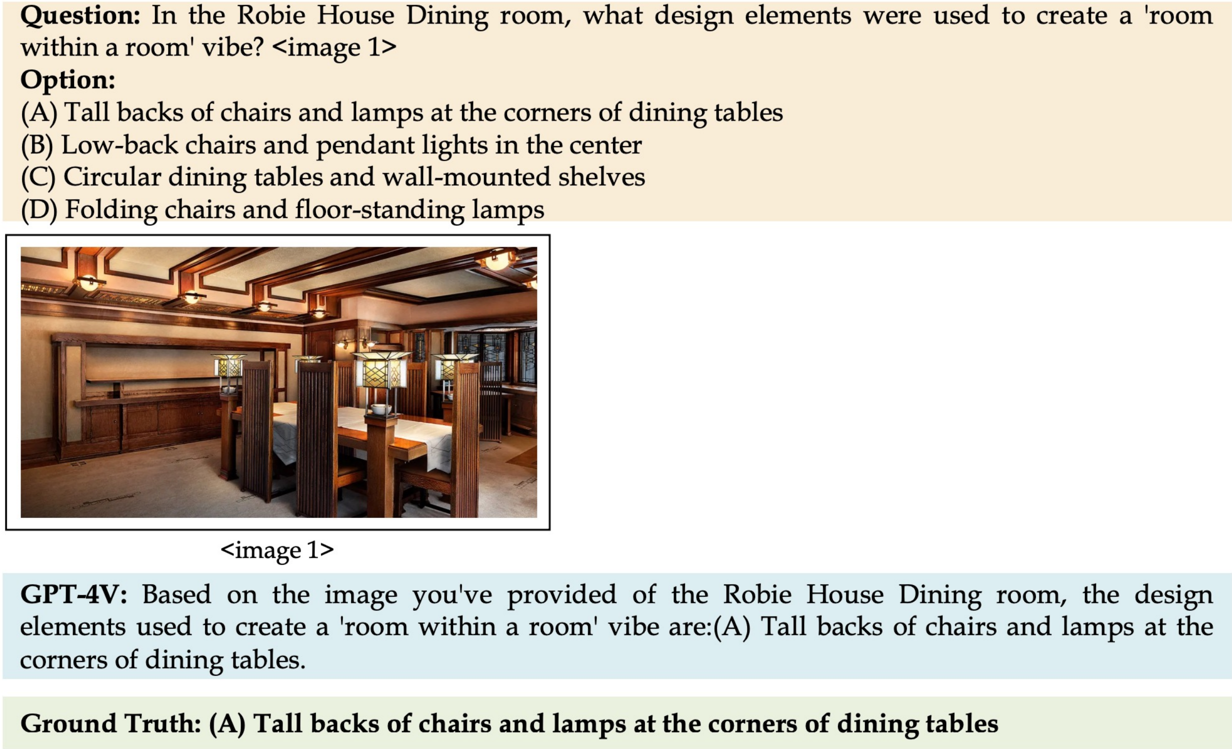
Question: A main function of castles was to provide a defense against invaders. This was achieved by several devices, including water-filled moats and barbicans, high stone walls surrounding the castle. Another such device was battlements (or crenellations). These were alternating sections of wall, called merlons, and openings, called embrasures. The embrasures were used for firing arrows. They also permitted boulders to be dropped on attackers attempting to scale the castle walls. If you were going to write an alternate caption for <image 1>, you might note that this castle was equipped with

Option:

(A) a barbican and a moat.
(B) merlons but not embrasures.
(C) a moat and crenellations.
(D) a barbican and battlements.

<image 1>

GPT-4V: The correct alternate caption for the image based on the provided information would be: (C) a moat and crenellations.



Question: In the Robie House Dining room, what design elements were used to create a 'room within a room' vibe? <image 1>

Option:

(A) Tall backs of chairs and lamps at the corners of dining tables
(B) Low-back chairs and pendant lights in the center
(C) Circular dining tables and wall-mounted shelves
(D) Folding chairs and floor-standing lamps

<image 1>

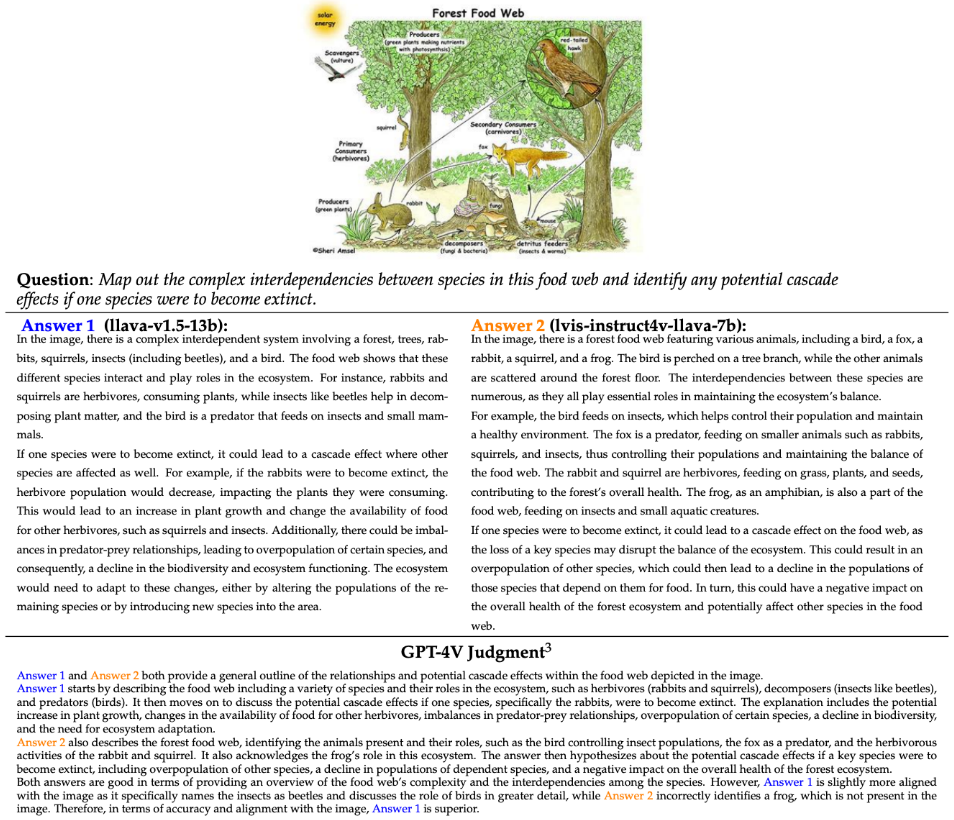
GPT-4V: Based on the image you've provided of the Robie House Dining room, the design elements used to create a 'room within a room' vibe are:(A) Tall backs of chairs and lamps at the corners of dining tables.

Ground Truth: (A) Tall backs of chairs and lamps at the corners of dining tables

MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Issues in Judging multimodal queries

- Some previous works utilize LLM/MLLM-as-a-judge for multimodal domain, but the issues beyond this framework remain under explored.


Forest Food Web
solar energy
Producers (green plants making nutrients with photosynthsis)
Scavengers (vulture)
red-tailed hawk
squirrel
Secondary Consumers (carnivores)
Primary Consumers (herbivores)
fox
Producers (green plants)
rabbit
fungi
mouse
decomposers (fungi & bacteria)
detritus feeders (insects & worms)
©Sheri Amsel

Question: *Map out the complex interdependencies between species in this food web and identify any potential cascade effects if one species were to become extinct.*

Answer 1 (llava-v1.5-13b):

In the image, there is a complex interdependent system involving a forest, trees, rabbits, squirrels, insects (including beetles), and a bird. The food web shows that these different species interact and play roles in the ecosystem. For instance, rabbits and squirrels are herbivores, consuming plants, while insects like beetles help in decomposing plant matter, and the bird is a predator that feeds on insects and small mammals.

If one species were to become extinct, it could lead to a cascade effect where other species are affected as well. For example, if the rabbits were to become extinct, the herbivore population would decrease, impacting the plants they were consuming. This would lead to an increase in plant growth and change the availability of food for other herbivores, such as squirrels and insects. Additionally, there could be imbalances in predator-prey relationships, leading to overpopulation of certain species, and consequently, a decline in the biodiversity and ecosystem functioning. The ecosystem would need to adapt to these changes, either by altering the populations of the remaining species or by introducing new species into the area.

Answer 2 (lvis-instruct4v-llava-7b):

In the image, there is a forest food web featuring various animals, including a bird, a fox, a rabbit, a squirrel, and a frog. The bird is perched on a tree branch, while the other animals are scattered around the forest floor. The interdependencies between these species are numerous, as they all play essential roles in maintaining the ecosystem's balance.

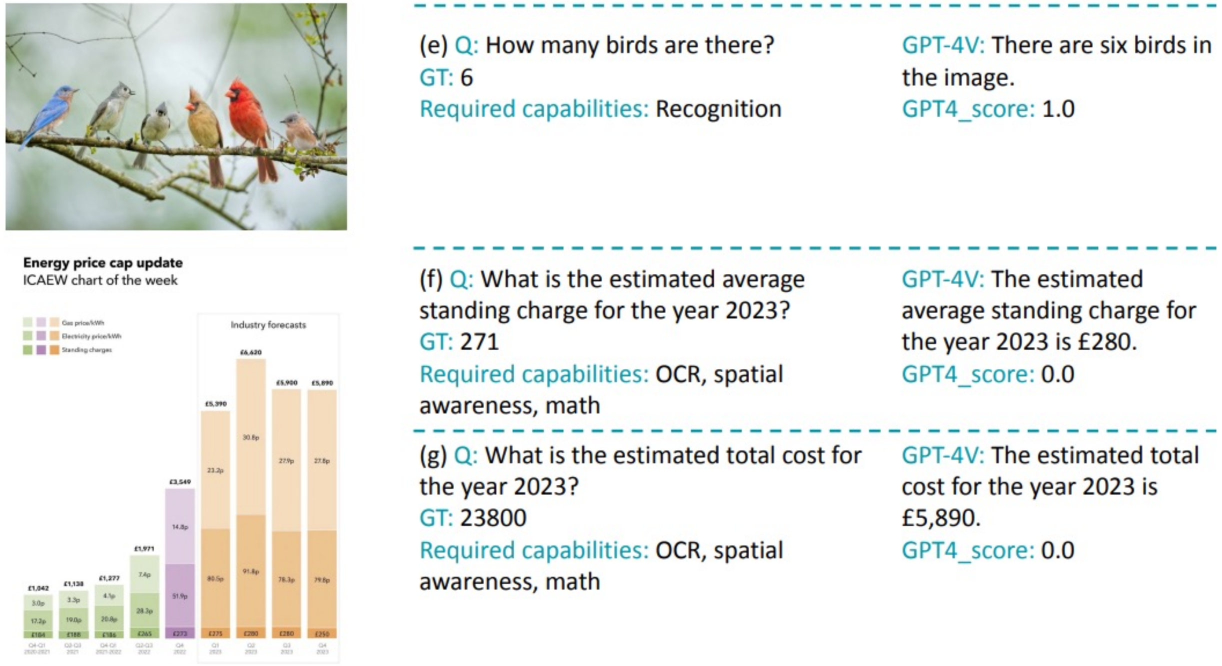
For example, the bird feeds on insects, which helps control their population and maintain a healthy environment. The fox is a predator, feeding on smaller animals such as rabbits, squirrels, and insects, thus controlling their populations and maintaining the balance of the food web. The rabbit and squirrel are herbivores, feeding on grass, plants, and seeds, contributing to the forest's overall health. The frog, as an amphibian, is also a part of the food web, feeding on insects and small aquatic creatures.

If one species were to become extinct, it could lead to a cascade effect on the food web, as the loss of a key species may disrupt the balance of the ecosystem. This could result in an overpopulation of other species, which could then lead to a decline in the populations of those species that depend on them for food. In turn, this could have a negative impact on the overall health of the forest ecosystem and potentially affect other species in the food web.

GPT-4V Judgment[3]

Answer 1 and Answer 2 both provide a general outline of the relationships and potential cascade effects within the food web depicted in the image.
Answer 1 starts by describing the food web including a variety of species and their roles in the ecosystem, such as herbivores (rabbits and squirrels), decomposers (insects like beetles), and predators (birds). It then moves on to discuss the potential cascade effects if one species, specifically the rabbits, were to become extinct. The explanation includes the potential increase in plant growth, changes in the availability of food for other herbivores, imbalances in predator-prey relationships, overpopulation of certain species, a decline in biodiversity, and the need for ecosystem adaptation.
Answer 2 also describes the forest food web, identifying the animals present and their roles, such as the bird controlling insect populations, the fox as a predator, and the herbivorous activities of the rabbit and squirrel. It also acknowledges the frog's role in this ecosystem. The answer then hypothesizes about the potential cascade effects if a key species were to become extinct, including overpopulation of other species, a decline in populations of dependent species, and a negative impact on the overall health of the forest ecosystem.
Both answers are good in terms of providing an overview of the food web's complexity and the interdependencies among the species. However, Answer 1 is slightly more aligned with the image as it specifically names the insects as beetles and discusses the role of birds in greater detail, while Answer 2 incorrectly identifies a frog, which is not present in the image. Therefore, in terms of accuracy and alignment with the image, Answer 1 is superior.

MLLM-Bench, Evaluating Multi-modal LLMs using GPT-4V



(e) Q: How many birds are there?
GT: 6
Required capabilities: Recognition

GPT-4V: There are six birds in the image.
GPT4_score: 1.0

(f) Q: What is the estimated average standing charge for the year 2023?
GT: 271
Required capabilities: OCR, spatial awareness, math

GPT-4V: The estimated average standing charge for the year 2023 is £280.
GPT4_score: 0.0

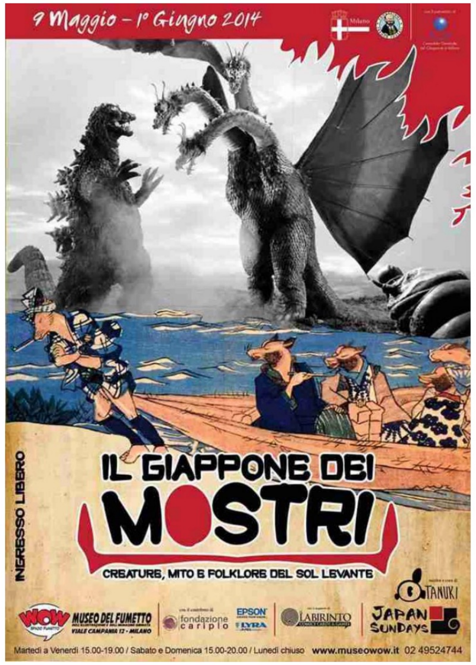
(g) Q: What is the estimated total cost for the year 2023?
GT: 23800
Required capabilities: OCR, spatial awareness, math

GPT-4V: The estimated total cost for the year 2023 is £5,890.
GPT4_score: 0.0

MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities

Towards Human Preference

- In free-form QA, human preference goes beyond mere accuracy in evaluating a problem.

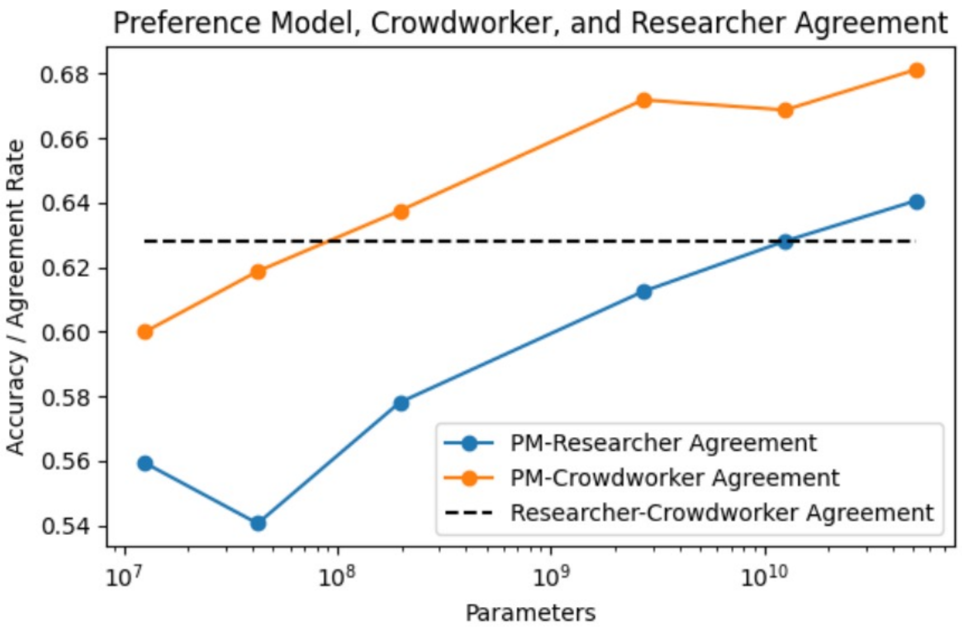

9 Maggio – 1° Giugno 2014
IL GIAPPONE DEI MOSTRI
CREATURE, MITO E FOLKLORE DEL SOL LEVANTE
INGRESSO LIBERO
WOW
MUSEO DEL FUMETTO
VIALE CAMPANIA 12 - MILANO
fondazione cariplo
EPSON
LYRA
LABIRINTO
TANUKI
JAPAN SUNDAYS
Martedì a Venerdì 15.00-19.00 / Sabato e Domenica 15.00-20.00 / Lunedì chiuso www.museowow.it 02 49524744

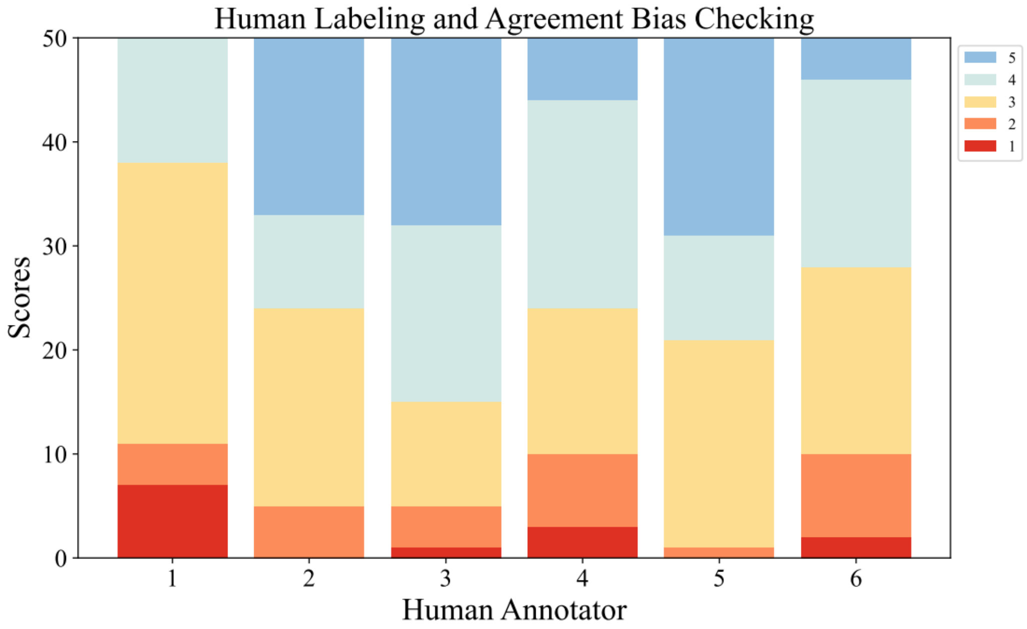
answer1.name: Assistant A
answer1.answer: The year at the top of the magazine is 2014.
answer2.name: Assistant B
answer2.answer: The year at the top of the magazine is 2014. It is prominently displayed in large, bold letters, making it easily readable and the focal point of the image.
answer3.name: Assistant C
answer3.answer: The year at the top of the magazine is 2014. This can be determined by reading the text that says ’9 Maggio – 10 Giugno 2014,’ which indicates that the magazine or event was from May 9 to June 10, 2014. This text is found in the upper section of the image, in a red-colored font, just above the image of the two monsters.
answer4.name: Assistant D
answer4.answer: The year at the top of the magazine is 2014. The top of the magazine says “9 MAGGIO - 1 GIUGNO 2014”. “9 MAGGIO - 1 GIUGNO” means “May 9 - June 1”. “2014” is the year.

Instruction: This is a task of text reading on natural image. Please analyze this figure in detail and answer the following question with reason based on this figure. what year is at the top of the magazine?

Human Preference is not consistent

- Human preference varies among different annotators, even when they are trained with tutorials to perform judgment.





Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback

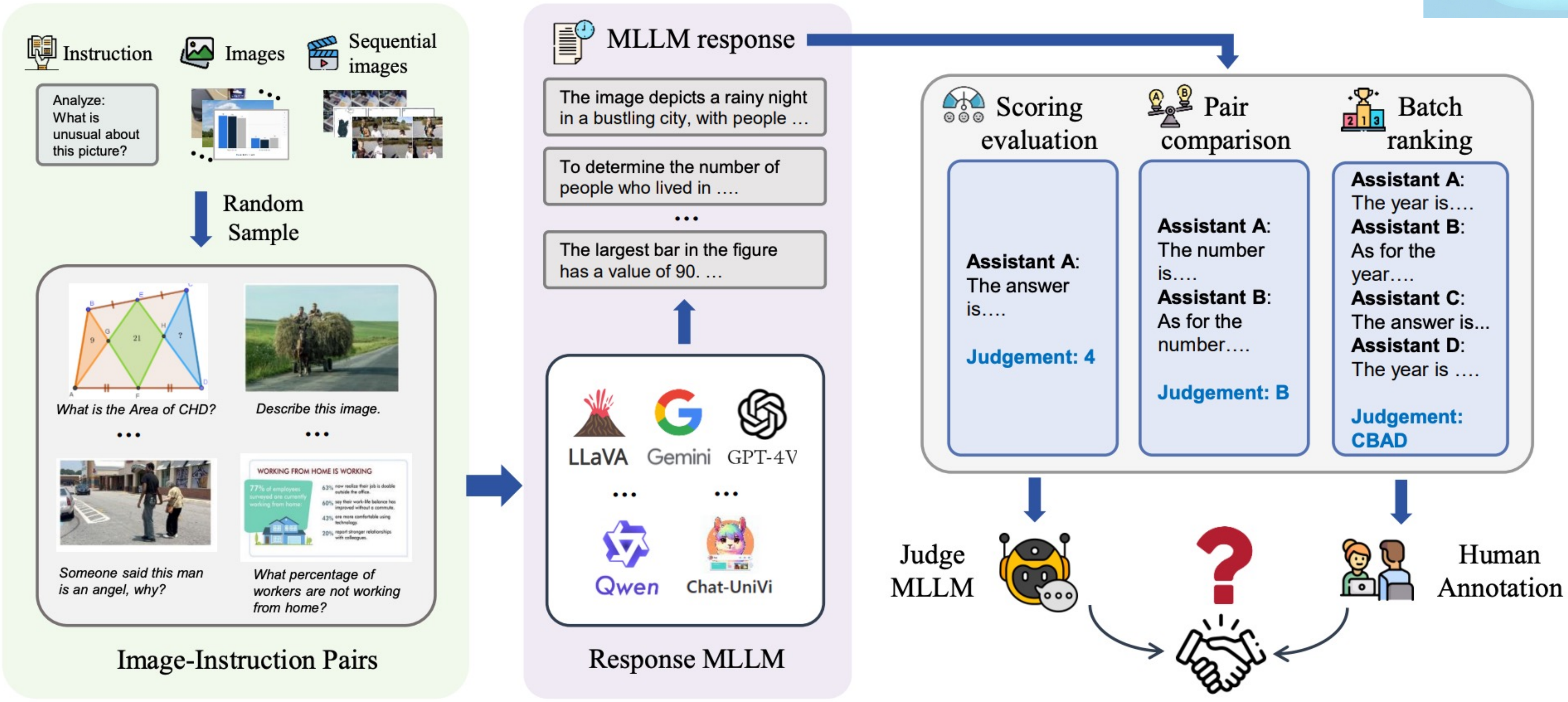
Questions

1. Can MLLMs effectively serve as judges in the multimodal domain?

2.Can MLLMs closely do their Judgment align with human preferences?

Q1: Overall Framework of MLLM-as-a-Judge

Q1: MLLM-as-a-Judge Framework



Instruction

Images

Sequential images

Analyze:
What is unusual about this picture?

Random Sample

What is the Area of CHD?

Describe this image.

…

Someone said this man is an angel, why?

What percentage of workers are not working from home?

Image-Instruction Pairs

MLLM response

The image depicts a rainy night in a bustling city, with people …

To determine the number of people who lived in ….

…

The largest bar in the figure has a value of 90. …

LLaVA

Gemini

GPT-4V

…

Qwen

Chat-UniVi

Response MLLM

Scoring evaluation

Assistant A:
The answer is….

Judgement: 4

Pair comparison

Assistant A:
The number is….
Assistant B:
As for the number….

Judgement: B

Batch ranking

Assistant A:
The year is….
Assistant B:
As for the year….
Assistant C:
The answer is...
Assistant D:
The year is ….

Judgement: CBAD

Judge MLLM

Human Annotation

Step 1：Image-Instruction Pair Collection

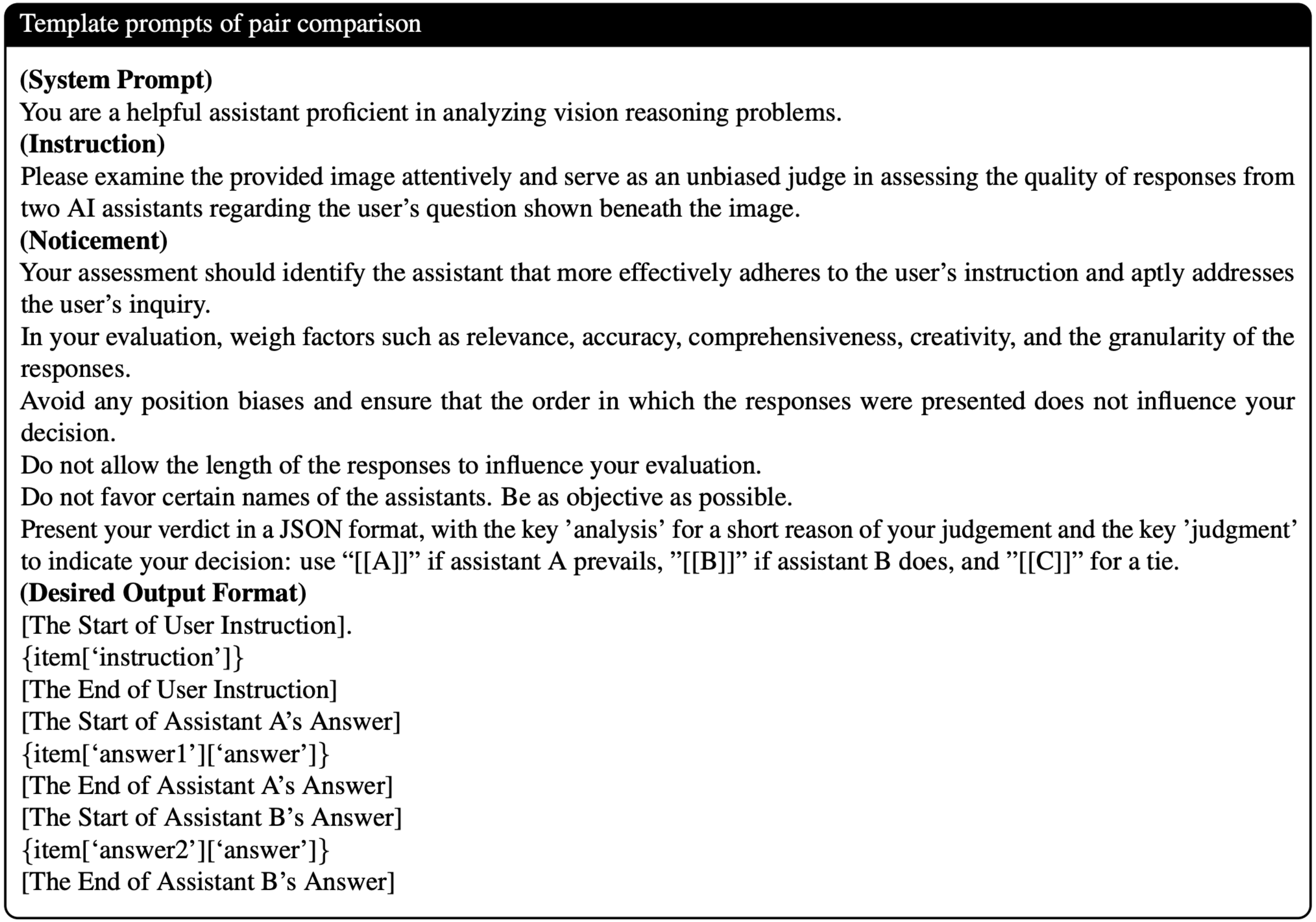
Step 2：MLLM Response Collection

Step 3：MLLM Judge v.s. Human Annotation

Q1: Question Formulation

Input: Text Instruction + Image + one/two/ multiple MLLM’s response

Output: Judgment



Template prompts of pair comparison

(System Prompt)
You are a helpful assistant proficient in analyzing vision reasoning problems.

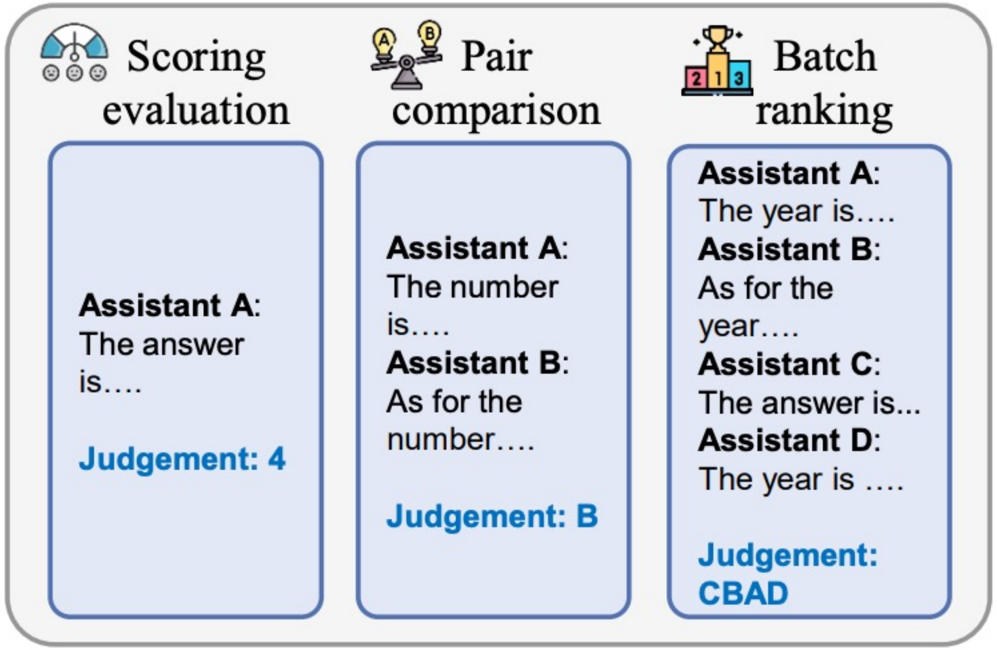
(Instruction)
Please examine the provided image attentively and serve as an unbiased judge in assessing the quality of responses from two AI assistants regarding the user’s question shown beneath the image.

(Noticement)
Your assessment should identify the assistant that more effectively adheres to the user’s instruction and aptly addresses the user’s inquiry.
In your evaluation, weigh factors such as relevance, accuracy, comprehensiveness, creativity, and the granularity of the responses.
Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision.
Do not allow the length of the responses to influence your evaluation.
Do not favor certain names of the assistants. Be as objective as possible.
Present your verdict in a JSON format, with the key ’analysis’ for a short reason of your judgement and the key ’judgment’ to indicate your decision: use “[[A]]” if assistant A prevails, ”[[B]]” if assistant B does, and ”[[C]]” for a tie.

(Desired Output Format)
[The Start of User Instruction].
{item[‘instruction’]}
[The End of User Instruction]
[The Start of Assistant A’s Answer]
{item[‘answer1’][‘answer’]}
[The End of Assistant A’s Answer]
[The Start of Assistant B’s Answer]
{item[‘answer2’][‘answer’]}
[The End of Assistant B’s Answer]

Q1: Three Judging Settings

- Score Evaluation: 1-5 Likert scale
- Pair Comparision: Win/Lose/Tie
- Batch Ranking: Desc/Asc Order



Scoring evaluation

Assistant A:
The answer is….

Judgement: 4

Pair comparison

Assistant A:
The number is….
Assistant B:
As for the number….

Judgement: B

Batch ranking

Assistant A:
The year is….
Assistant B:
As for the year….
Assistant C:
The answer is...
Assistant D:
The year is ….

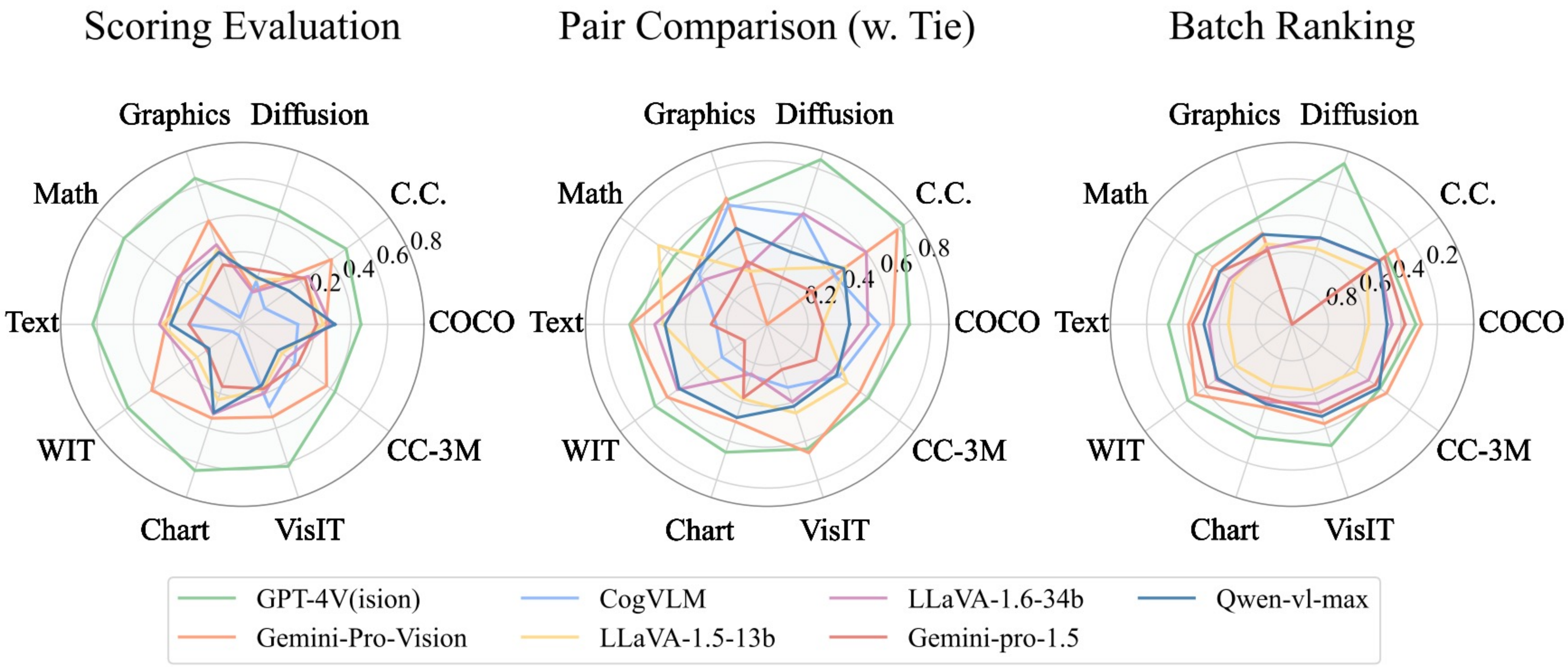
Judgement: CBAD

Q2: How does current MLLMs perform on Judging tasks?

Q2: Evaluation Metrics and Models

- Metrics
  - Score Evaluation: Pearson Similarity (↑)
  - Pair Comparision: Accuracy (↑)
  - Batch Ranking: Levenshtein distance (↓)
  - Human-in-the-Loop: Human Agreement Rate
- 11 Models:
  - Proprietary: GPT-4V, Gemini-Pro-Vision-1.0/1.0-latest, Qwen-VL-Plus/Max
  - Open-Source: LLaVA-1.5-13b, LLaVA-1.6-7b/13b/34b, Qwen-VL-Chat, CogVLM
- Inference Prompt Design:
  - Analyze-then-Judge: 2 step COT

Q2: Quantitative Results



Q2: Qualitative Results

- Human annotators agree more on MLLM-as-a-Judge in Pairwise setting, while still fall short in Batch Ranking tasks.

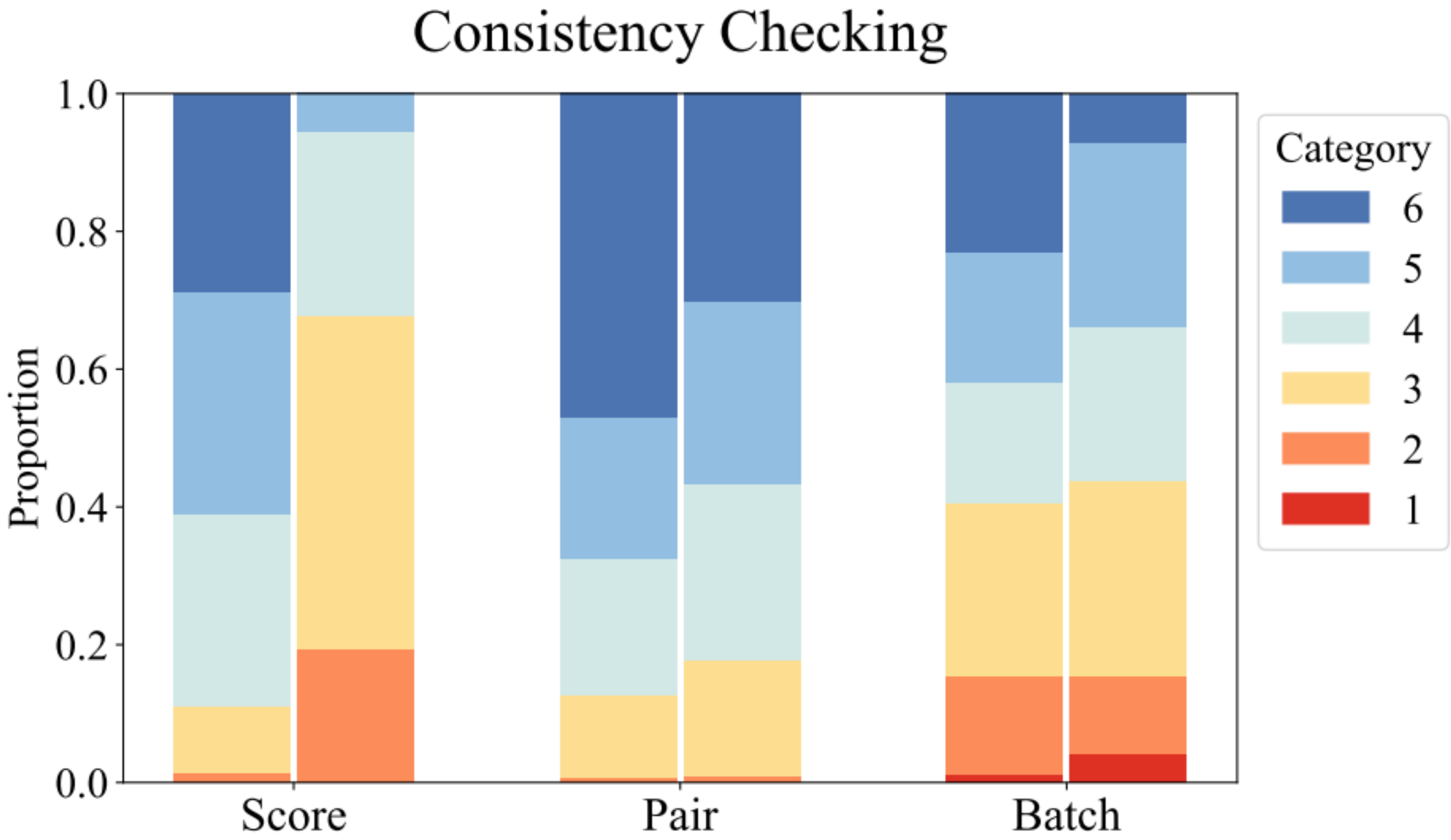
Settings	**MLLM**	COCO	C.C.	Diffusion	Graphics	Math	Text	WIT	Chart	VisIT	CC-3M	Average
Score (↑)	Gemini	0.783	**0.739**	-	0.618	0.536	0.621	**0.749**	0.630	0.712	0.702	0.677
	GPT-4V	**0.799**	0.725	**0.506**	**0.688**	**0.638**	**0.706**	0.714	**0.676**	**0.779**	**0.754**	**0.699**
Pair (↑)	Gemini	0.705	0.833	-	0.733	0.520	0.717	**0.827**	0.620	**0.853**	0.703	0.724
	GPT-4V	**0.821**	**0.926**	**0.873**	**0.794**	**0.618**	**0.752**	0.790	**0.796**	0.797	**0.766**	**0.793**
Batch (↓)	Gemini	0.642	**0.639**	-	0.333	0.330	0.473	0.511	0.315	0.422	**0.554**	0.469
	GPT-4V	**0.663**	**0.639**	**0.912**	**0.536**	**0.475**	**0.615**	**0.641**	**0.640**	**0.622**	0.467	**0.621**

Q3: Notable findings in the MLLM-as-a-Judge process

Q3: Problems in MLLM-as-a-Judge

- Judging Consistency
- Bias: Egocentric Bias, Position Bias, Length Bias
- Hallucination: Detection & Mitigation

Q3: Judging Consistency

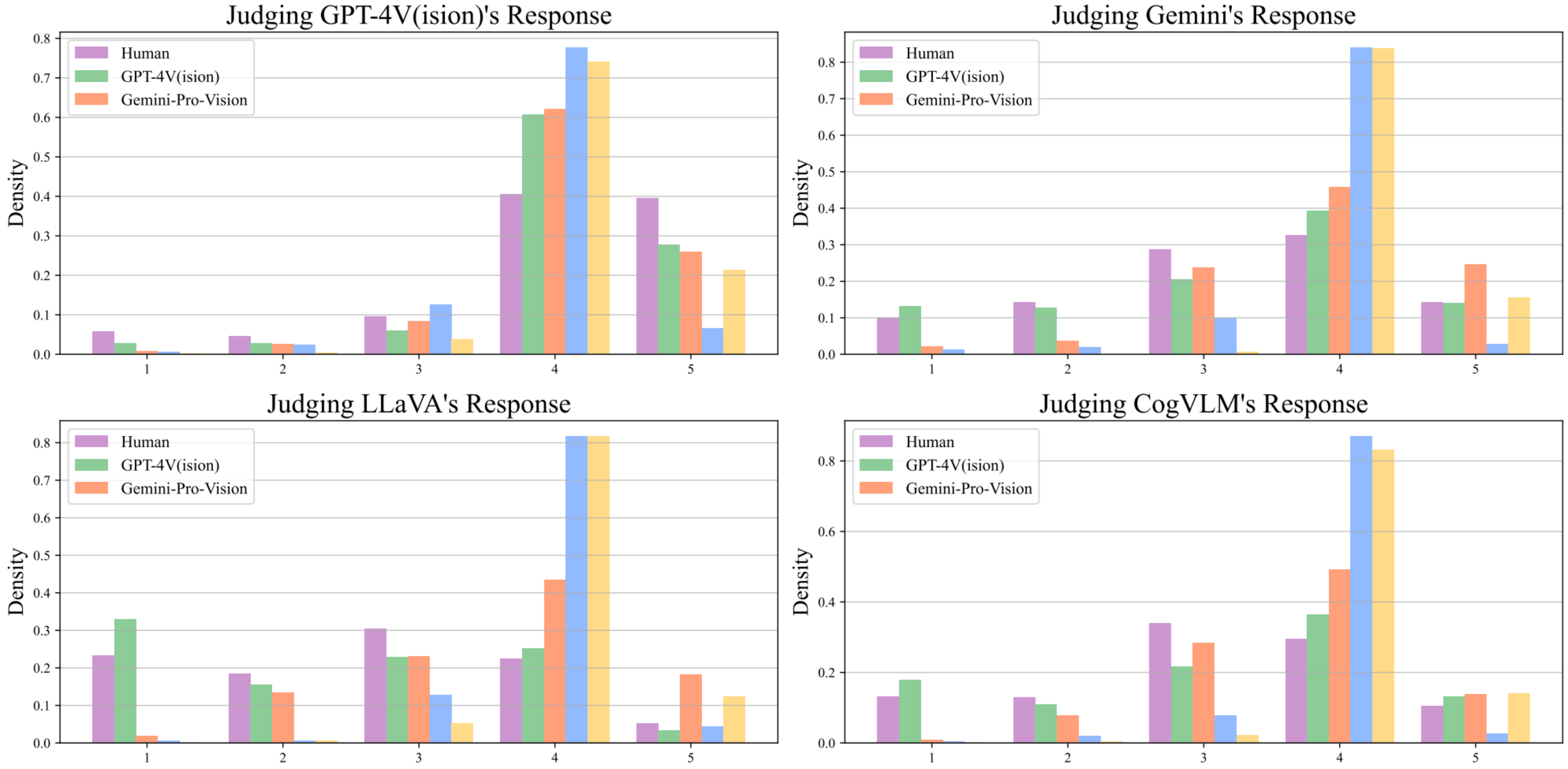


Q3: Multi-steps CoT **Do Not** Enhance Performance

Settings	MLLM	COCO	C.C.	Diffusion	Graphics	Math	Text	WIT	Chart	VisIT	CC-3M	Ave.
Score (↑)	GPT-4V	**0.454**	**0.507**	**0.458**	**0.645**	**0.606**	**0.624**	**0.579**	**0.645**	**0.620**	**0.431**	**0.557**
	GPT-4V (+CoT)	0.246	0.165	0.192	0.385	0.397	0.400	0.298	0.443	0.423	0.038	0.299
	Gemini	0.262	0.408	-	0.400	0.228	0.222	0.418	0.343	0.336	0.374	0.299
	Gemini (+CoT)	0.127	0.068	0.117	0.220	0.132	0.182	0.105	0.140	0.222	0.128	0.144
Pair w. Tie (↑)	GPT-4V	**0.696**	**0.824**	**0.847**	**0.639**	**0.564**	**0.673**	**0.679**	**0.657**	0.640	**0.612**	**0.683**
	GPT-4V (+CoT)	0.507	0.657	0.561	0.601	0.515	0.580	0.489	0.521	**0.646**	0.553	0.563
	Gemini	0.616	0.787	-	0.650	0.436	0.664	0.605	0.500	0.660	0.560	0.609
	Gemini (+CoT)	0.233	0.239	0.420	0.207	0.284	0.329	0.352	0.357	0.247	0.239	0.291
Pair w.o. Tie (↑)	GPT-4V	**0.804**	**0.870**	**0.922**	**0.807**	**0.801**	**0.805**	**0.734**	**0.849**	**0.761**	**0.703**	**0.806**
	GPT-4V (+CoT)	0.673	0.821	0.845	0.707	0.738	0.787	0.548	0.756	0.753	0.654	0.728
	Gemini	0.717	0.840	-	0.770	0.678	0.793	0.688	0.658	0.711	0.652	0.723
	Gemini (+CoT)	0.267	0.275	0.573	0.264	0.414	0.424	0.427	0.511	0.299	0.319	0.377
Batch (↓)	GPT-4V	0.323	0.344	**0.092**	**0.401**	**0.367**	**0.341**	**0.302**	**0.364**	**0.313**	0.407	**0.325**
	GPT-4V (+CoT)	0.428	0.416	-	0.427	0.434	0.401	0.366	0.406	0.422	0.472	0.419
	Gemini	**0.287**	**0.299**	-	0.473	0.462	0.430	0.344	0.520	0.426	**0.357**	0.400
	Gemini (+CoT)	0.441	0.481	0.542	0.595	0.494	0.533	0.483	0.569	0.486	0.463	0.509

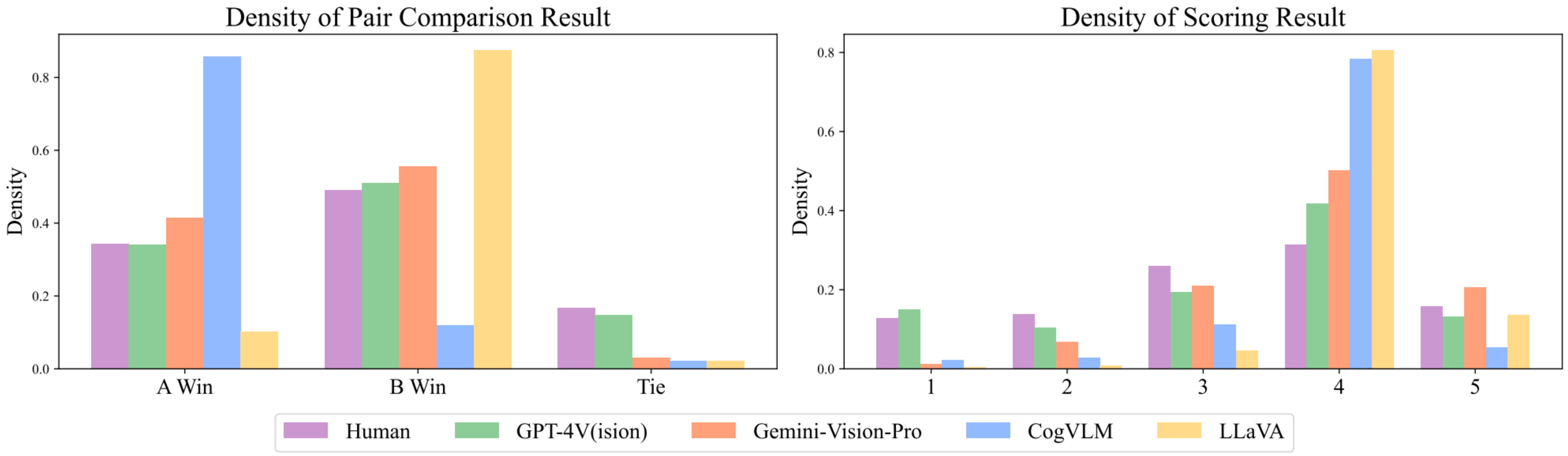
Q3: Egocentric Bias

- GPT-4V and Gemini-Pro both have a slight degree of Egocentricity.



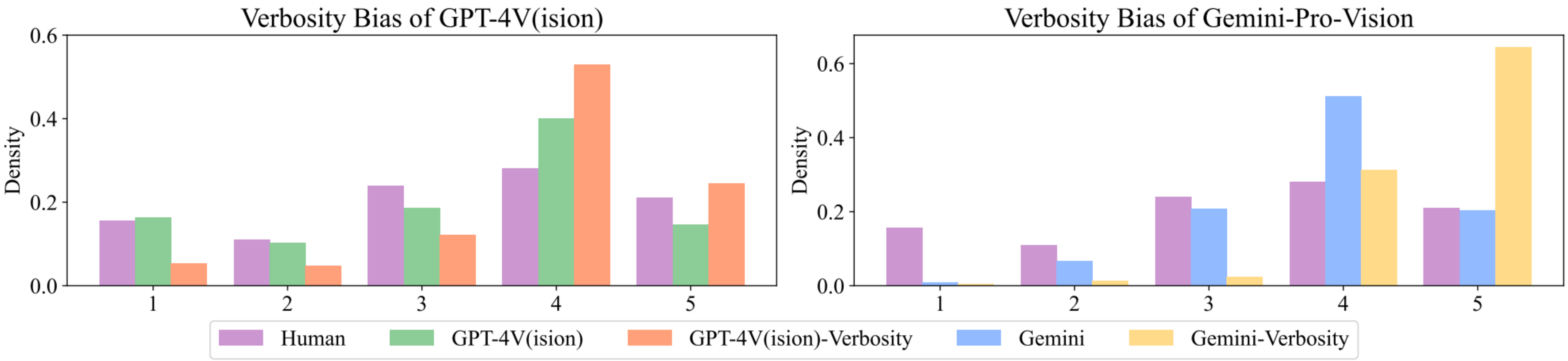
Q3: Position Bias

- Judge MLLMs favor response of specific positions.



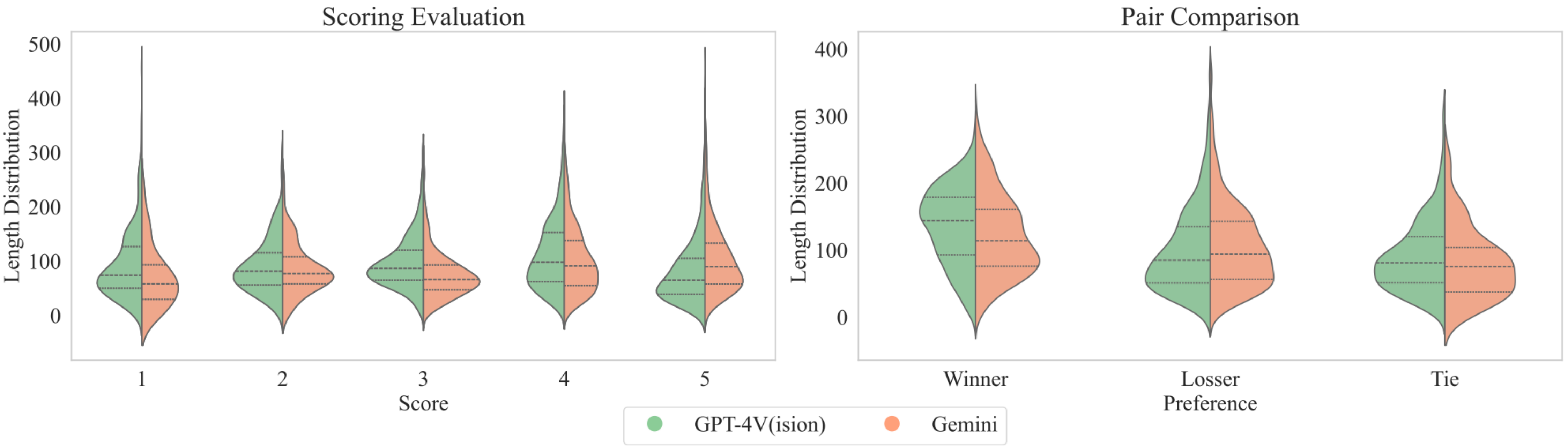
Q3: Length Bias

- Both GPT-4V and Gemini assign higher scores to longer content.



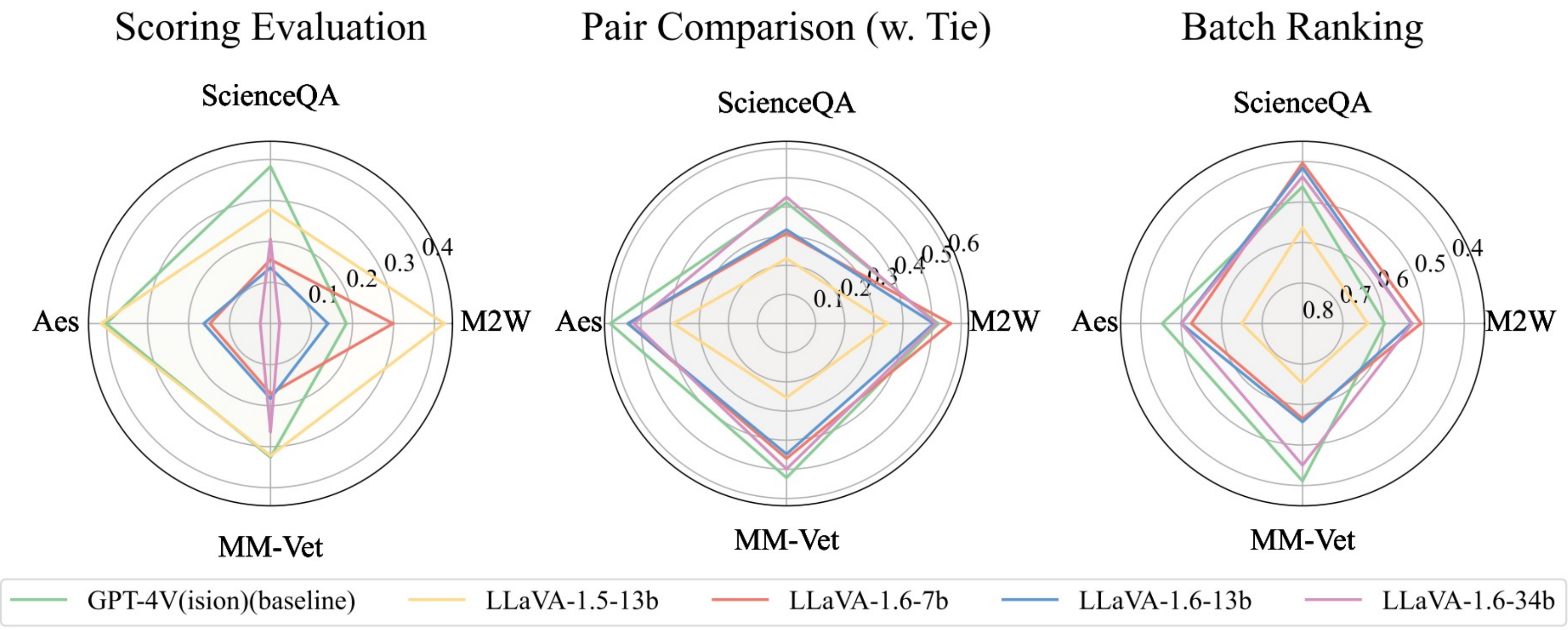
Q3: Length Bias

- GPT-4V favor more to longer response in Pair Comparison.



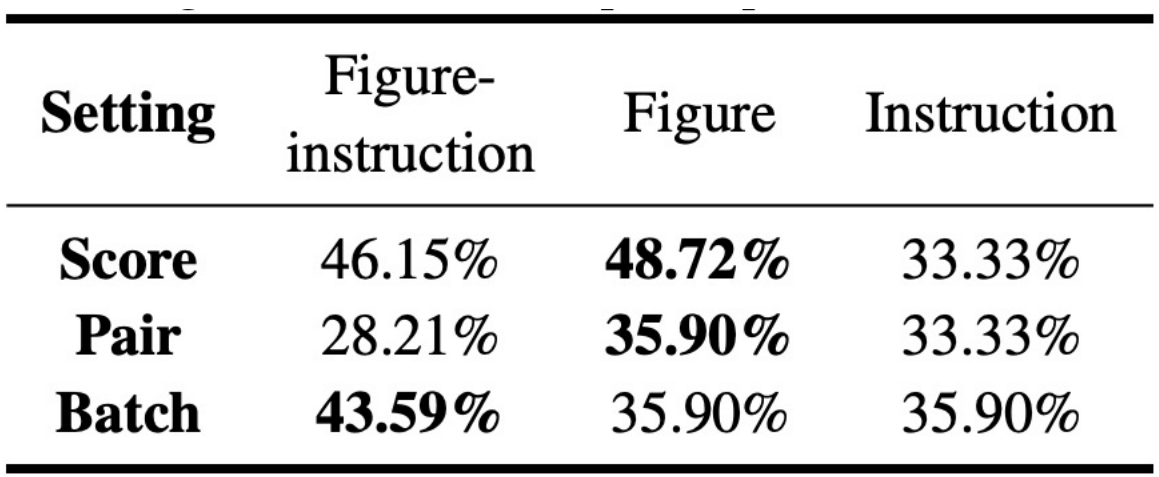
Q3: Scaling Law for MLLM-as-a-Judge

- Model Family: Llava-1.6-7b/13b/34b
- In Score evaluation, LLaVA-1.6-34b slightly outperform others in Math, Chart tasks, showing a relatively strong scaling law.



Q3: Hallucination Detection and Mitigation

- We observe a higher frequency of hallucinations in Batch Ranking, compared to Pair Comparison and Scoring Evaluation.
- Multi-step CoT approach mitigate hallucination.



Setting	Figure-instruction	Figure	Instruction
Score	46.15%	**48.72%**	33.33%
Pair	28.21%	**35.90%**	33.33%
Batch	**43.59%**	35.90%	35.90%

Q3: Can LLM judge multimodal queries?

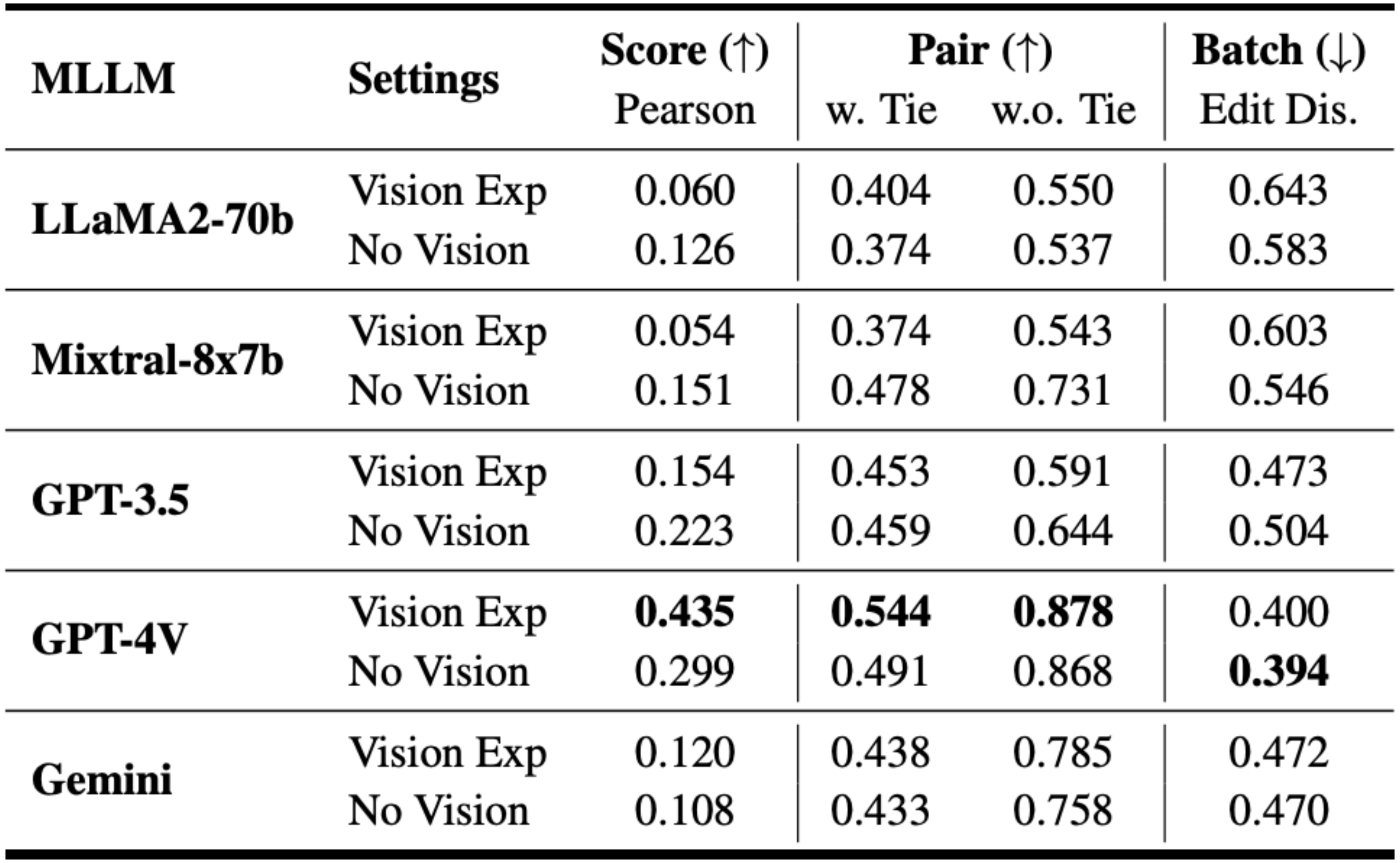
- Caption Model: GPT-4V

- Judge Model
  - LLMs: LLaMA-70b, Mixtral8x7b-v0.1 and GPT-3.5
  - MLLMs: GPT-4V, Gemini-Vision-Pro

- Two Setting: w./w.o. image caption

Q3: Can LLM judge multimodal queries?

- The performance of LLMs in multimodal judging tasks varies with or without image captions.



MLLM	**Settings**	**Score (↑)**	**Pair (↑)**		**Batch (↓)**
		Pearson	w. Tie	w.o. Tie	Edit Dis.
LLaMA2-70b	Vision Exp	0.060	0.404	0.550	0.643
	No Vision	0.126	0.374	0.537	0.583
Mixtral-8x7b	Vision Exp	0.054	0.374	0.543	0.603
	No Vision	0.151	0.478	0.731	0.546
GPT-3.5	Vision Exp	0.154	0.453	0.591	0.473
	No Vision	0.223	0.459	0.644	0.504
GPT-4V	Vision Exp	**0.435**	**0.544**	**0.878**	0.400
	No Vision	0.299	0.491	0.868	**0.394**
Gemini	Vision Exp	0.120	0.438	0.785	0.472
	No Vision	0.108	0.433	0.758	0.470

Q4: Future Direction & Follow-up Works

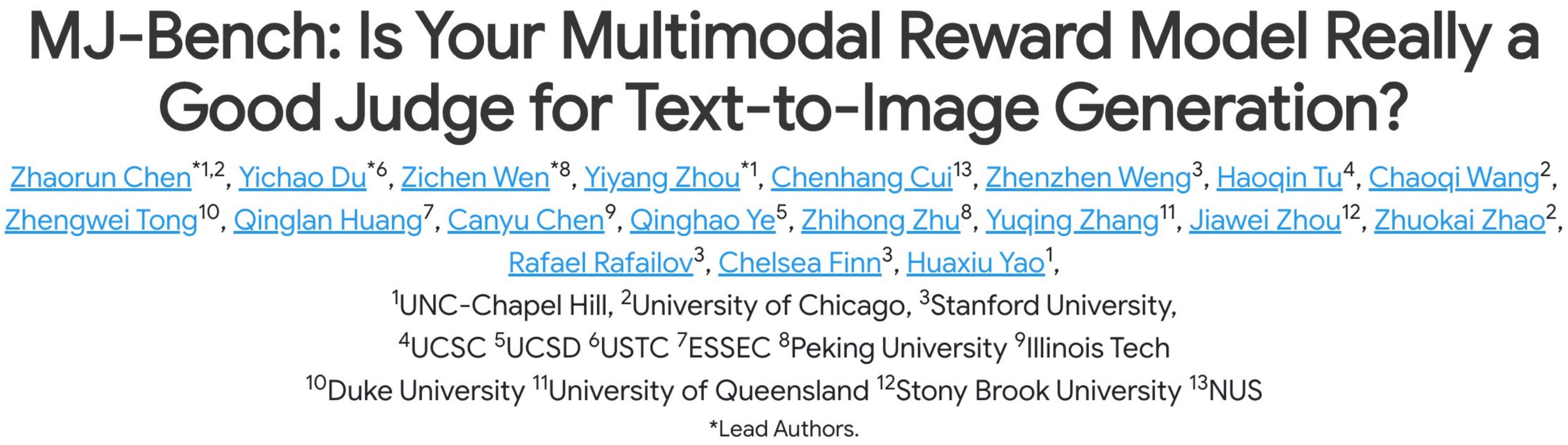
Q4: Future Directions

- Multimodal RLHF

- Exploring the upper bound of MLLM-as-a-Judge
  - Scaling Law: more powerful LLM backbone
  - Human Preference Alignment in Judging tasks
  - Human-in-the-Loop Framework

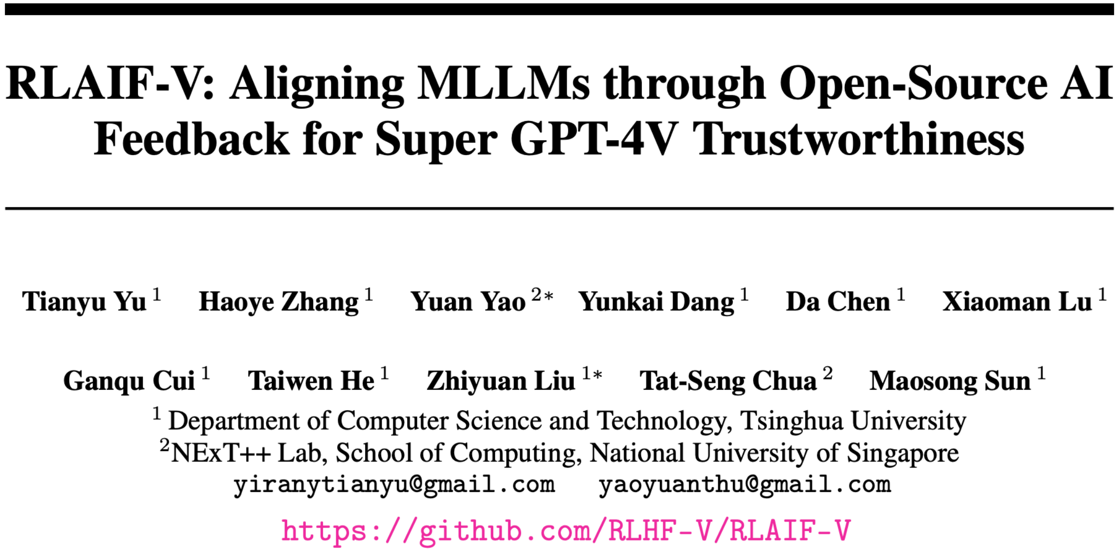
- MLLM-as-a-Judge serving as a reward model

Q4: Follow-up works



MJ-Bench: Is Your Multimodal Reward Model Really a Good Judge for Text-to-Image Generation?

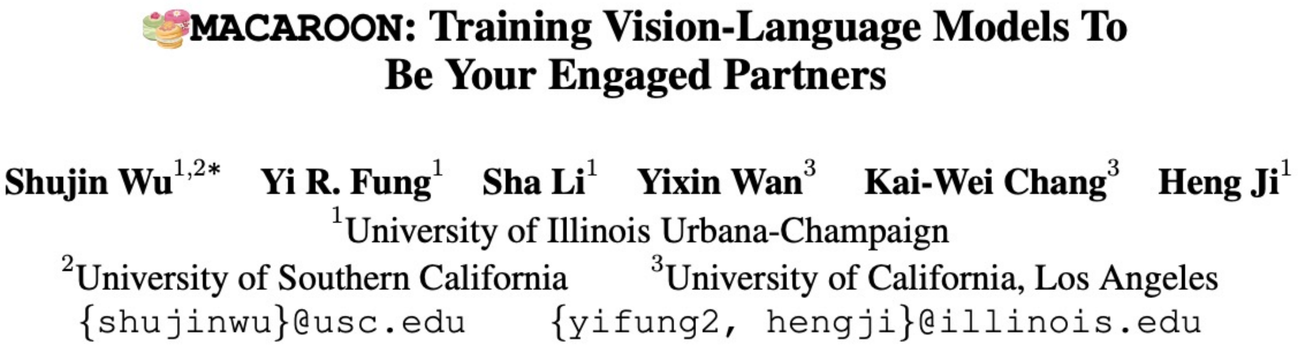
Zhaorun Chen[*1,2], Yichao Du[*6], Zichen Wen[*8], Yiyang Zhou[*1], Chenhang Cui[13], Zhenzhen Weng[3], Haoqin Tu[4], Chaoqi Wang[2],
Zhengwei Tong[10], Qinglan Huang[7], Canyu Chen[9], Qinghao Ye[5], Zhihong Zhu[8], Yuqing Zhang[11], Jiawei Zhou[12], Zhuokai Zhao[2],
Rafael Rafailov[3], Chelsea Finn[3], Huaxiu Yao[1],
[1]UNC-Chapel Hill, [2]University of Chicago, [3]Stanford University,
[4]UCSC [5]UCSD [6]USTC [7]ESSEC [8]Peking University [9]Illinois Tech
[10]Duke University [11]University of Queensland [12]Stony Brook University [13]NUS
*Lead Authors.



RLAIF-V: Aligning MLLMs through Open-Source AI Feedback for Super GPT-4V Trustworthiness

Tianyu Yu[1] Haoye Zhang[1] Yuan Yao[2*] Yunkai Dang[1] Da Chen[1] Xiaoman Lu[1]
Ganqu Cui[1] Taiwen He[1] Zhiyuan Liu[1*] Tat-Seng Chua[2] Maosong Sun[1]
[1] Department of Computer Science and Technology, Tsinghua University
[2]NExT++ Lab, School of Computing, National University of Singapore
yiranytianyu@gmail.com yaoyuanthu@gmail.com

https://github.com/RLHF-V/RLAIF-V



MACAROON: Training Vision-Language Models To Be Your Engaged Partners

Shujin Wu[1,2*] Yi R. Fung[1] Sha Li[1] Yixin Wan[3] Kai-Wei Chang[3] Heng Ji[1]
[1]University of Illinois Urbana-Champaign
[2]University of Southern California [3]University of California, Los Angeles
{shujinwu}@usc.edu {yifung2, hengji}@illinois.edu

Take-Aways

Take-Aways

✓We evaluate the judgment performance of 11 MLLMs across 14 datasets under three settings.

✓**First,** while MLLMs demonstrate proficiency in aligning with human preferences in Pair Comparison tasks, they require further improvement in Score Evaluation and Batch Ranking, particularly in reasoning tasks.

✓**Secondly,** GPT-4V consistently outperforms other models in all tasks and settings, across various data types.

✓**Finally,** MLLMs exhibit hallucinations, biases, and inconsistencies in judgments.

Thank you for attending!

Dongping Chen *is looking for a PhD position in 25 Fall!*

Year 3 Undergraduate

Contact Info.

MLLM-as-a-Judge Team