NEURAL INFORMATION PROCESSING SYSTEMS

VLM4Bio: A Benchmark Dataset to Evaluate Pretrained Vision-Language Models for Trait Discovery from Biological Images

M. Maruf*, Arka Daw*, Kazi Sajeed Mehrab, Harish Babu Manogaran, Abhilash Neog, Medha Sawhney, Mridul Khurana, James P. Balhoff, Yasin Bakış, Bahadir Altintas, Matthew J Thompson, Elizabeth G Campolongo, Josef C. Uyeda, Hilmar Lapp, Henry L. Bart Jr., Paula M. Mabee, Yu Su, Wei-Lun Chao, Charles Stewart, Tanya Berger-Wolf, Wasila Dahdul, Anuj Karpatne

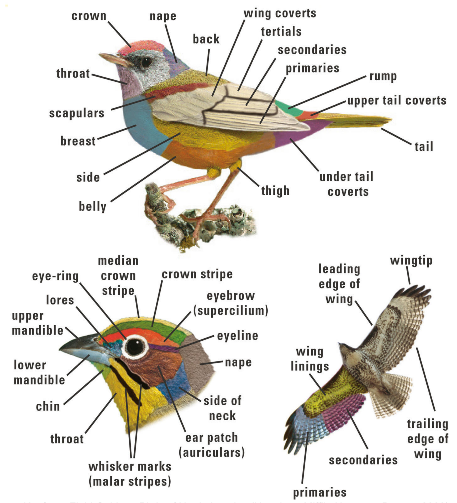
KGML Lab

Discovering Biological Traits from Images

- **Large repositories of organism images are available.**
  - Museum and university library collection.
  - Citizen science data.

- Biologists are interested in discovering biological traits directly from the organism's images.

- **Large Vision-Language Models (VLMs) can solve a diverse range of tasks involving text and images.**



Kaufman Field Guide to Birds of North America (Houghton Mifflin Harcourt, Boston, 2000)

Traits are the externally visible characteristics of an organism that define the species. Traits can be certain regions, colors, patterns, or landmark points of the organism.

Can we use Large Vision-Language Models (VLMs) for trait discovery?

Do pre-trained VLMs contain the necessary scientific knowledge to aid biologists in answering a variety of questions pertinent to the discovery of biological traits from images?

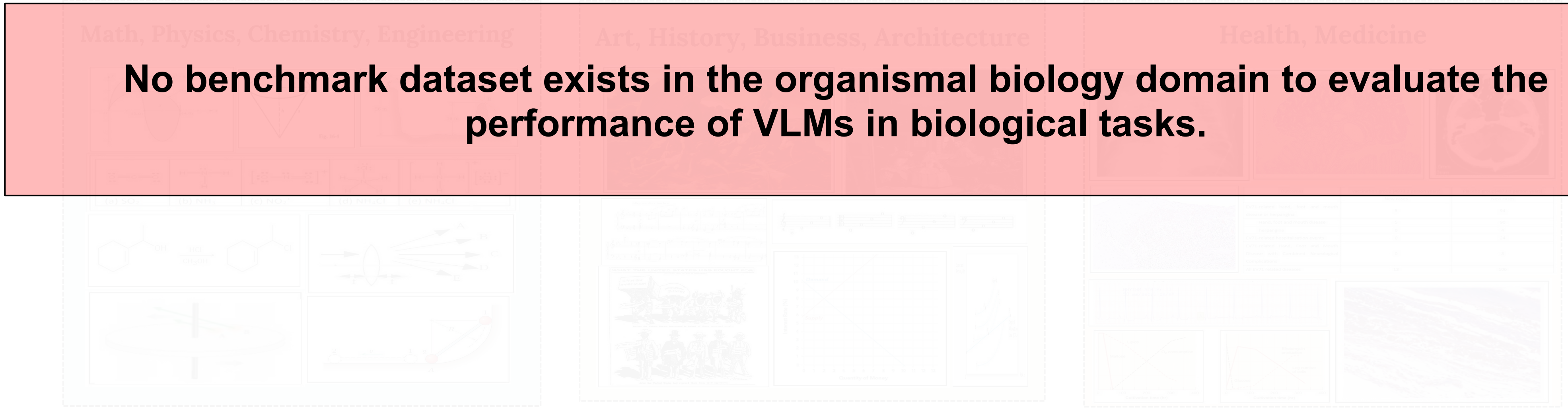
Benchmark datasets for VLMs

- Most of the existing benchmark datasets focus on commonsense knowledge rather than expert knowledge.

- Domain-specific benchmark datasets:
  - MedQA is a collection of VQA problems from medical exams.
  - MathVista mathematical reasoning questions in visual contexts.
  - MMMU covers college-level problems from diverse business, arts, health, medicine, and engineering domains.

No benchmark dataset exists in the organismal biology domain to evaluate the performance of VLMs in biological tasks.

- Most of the existing benchmark datasets focus on commonsense knowledge rather than expert knowledge.
  - **VQA, OK-VQA, MSCOCO, and GQA** are generic benchmarks for visual tasks.
  - **MMBench, SEED, and LAMM** evaluate multimodal perception and reasoning abilities.
  - **HaELM** evaluates hallucination.
- Domain-specific benchmark datasets:
  - **MMMU** covers problems from diverse domains such as business, arts, health, medicine, and engineering.



Math, Physics, Chemistry, Engineering
Art, History, Business, Architecture
Health, Medicine

VLM4Bio Dataset

- A benchmark dataset of **469K** question-answer pairs involving **30K** images from three groups of organisms**: fishes, birds, and butterflies,** covering **five** biologically relevant tasks.

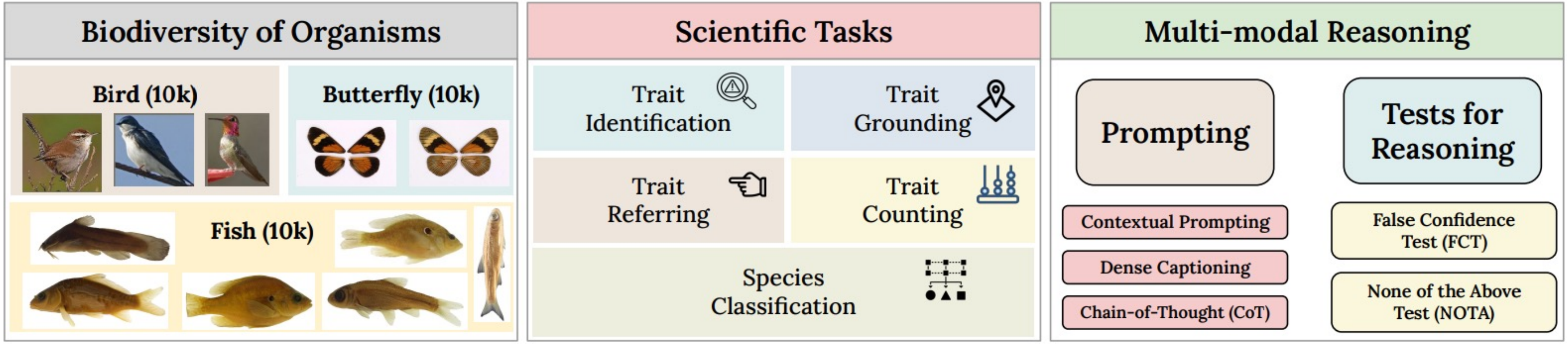
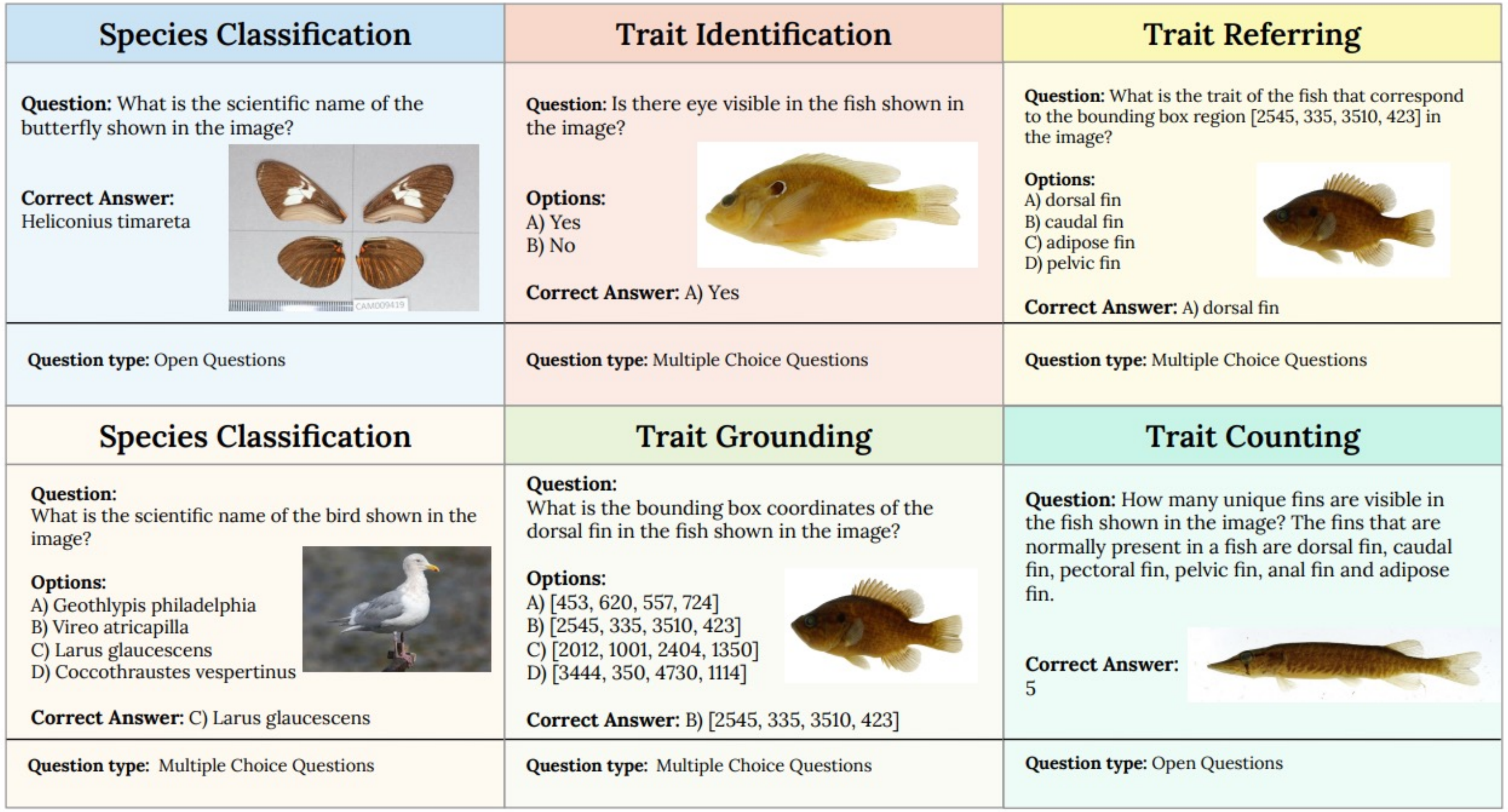


Figure 1: Overview of our goals and contributions. We analyze the capabilities of 12 state-of-the-art (SOTA) vision-language models (VLMs) in answering scientific questions using images from three groups of organisms: fishes, birds, and butterflies, over five groups of biologically relevant tasks. We also explore the effectiveness of these models for reasoning using various prompting techniques and tests for reasoning hallucination.

Designing Benchmark: Biologically Relevant Tasks



Species Classification

Question: What is the scientific name of the butterfly shown in the image?

Correct Answer: Heliconius timareta

Question type: Open Questions

Trait Identification

Question: Is there eye visible in the fish shown in the image?

Options:
A) Yes
B) No

Correct Answer: A) Yes

Question type: Multiple Choice Questions

Trait Referring

Question: What is the trait of the fish that correspond to the bounding box region [2545, 335, 3510, 423] in the image?

Options:
A) dorsal fin
B) caudal fin
C) adipose fin
D) pelvic fin

Correct Answer: A) dorsal fin

Question type: Multiple Choice Questions

Species Classification

Question:
What is the scientific name of the bird shown in the image?

Options:
A) Geothlypis philadelphia
B) Vireo atricapilla
C) Larus glaucescens
D) Coccothraustes vespertinus

Correct Answer: C) Larus glaucescens

Question type: Multiple Choice Questions

Trait Grounding

Question:
What is the bounding box coordinates of the dorsal fin in the fish shown in the image?

Options:
A) [453, 620, 557, 724]
B) [2545, 335, 3510, 423]
C) [2012, 1001, 2404, 1350]
D) [3444, 350, 4730, 1114]

Correct Answer: B) [2545, 335, 3510, 423]

Question type: Multiple Choice Questions

Trait Counting

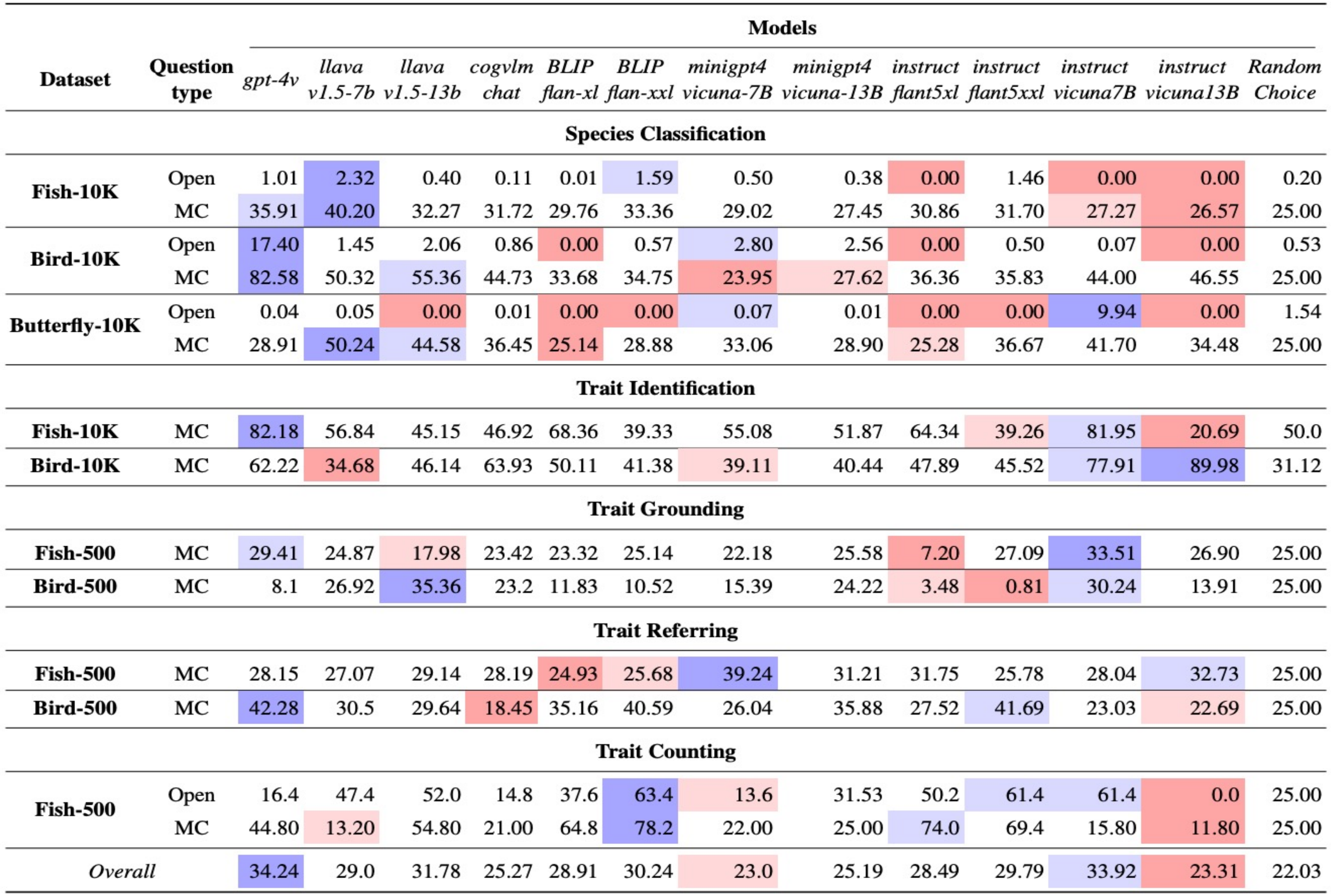
Question: How many unique fins are visible in the fish shown in the image? The fins that are normally present in a fish are dorsal fin, caudal fin, pectoral fin, pelvic fin, anal fin and adipose fin.

Correct Answer: 5

Question type: Open Questions

Figure 2: Illustrative examples of **VLM4Bio** tasks with different question-types.

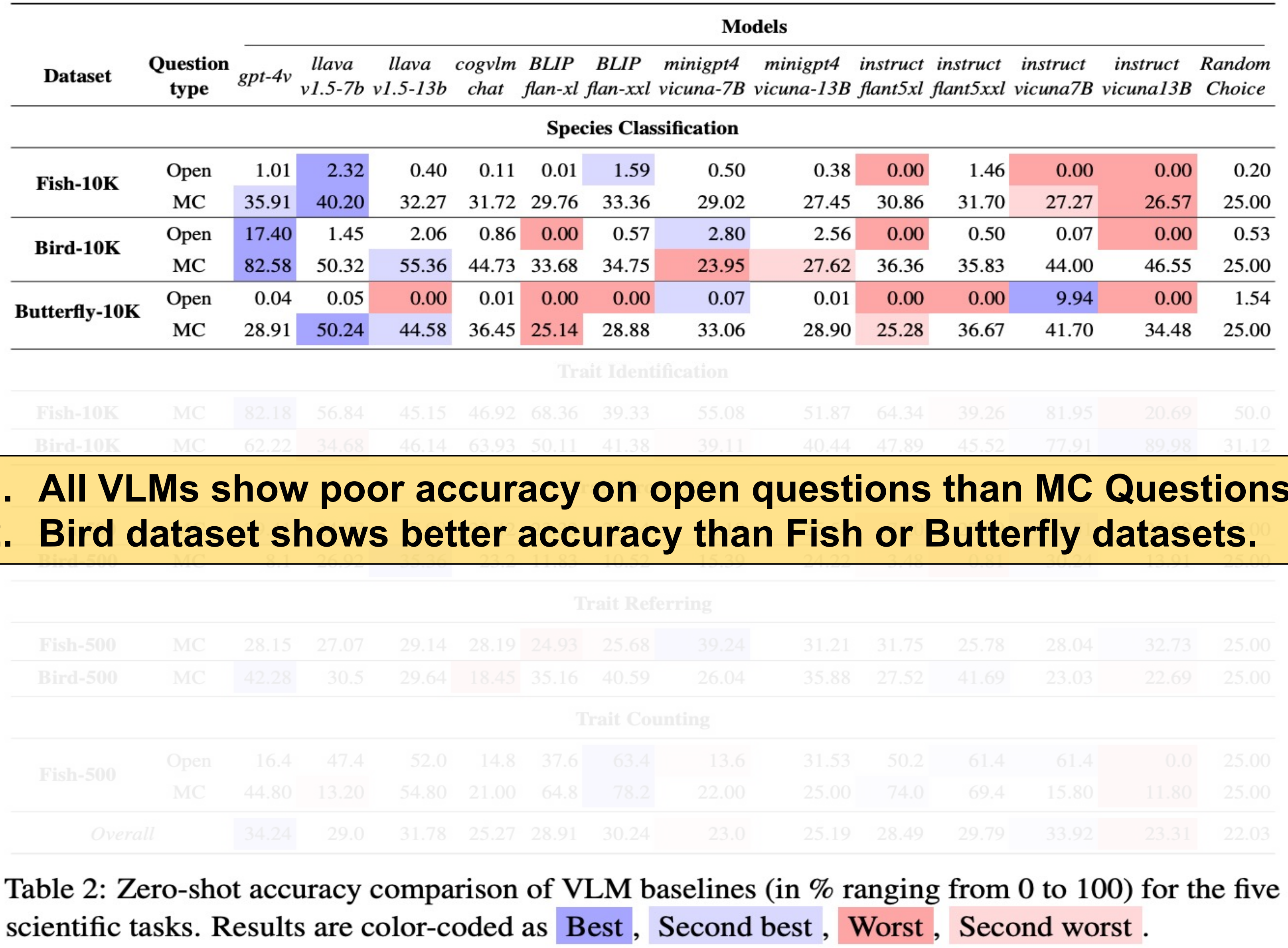
Zero-shot Evaluation



Dataset	Question type	*gpt-4v*	*llava v1.5-7b*	*llava v1.5-13b*	*cogvlm chat*	*BLIP flan-xl*	*BLIP flan-xxl*	*minigpt4 vicuna-7B*	*minigpt4 vicuna-13B*	*instruct flant5xl*	*instruct flant5xxl*	*instruct vicuna7B*	*instruct vicuna13B*	*Random Choice*
		Models												
Species Classification														
Fish-10K	Open	1.01	2.32	0.40	0.11	0.01	1.59	0.50	0.38	0.00	1.46	0.00	0.00	0.20
	MC	35.91	40.20	32.27	31.72	29.76	33.36	29.02	27.45	30.86	31.70	27.27	26.57	25.00
Bird-10K	Open	17.40	1.45	2.06	0.86	0.00	0.57	2.80	2.56	0.00	0.50	0.07	0.00	0.53
	MC	82.58	50.32	55.36	44.73	33.68	34.75	23.95	27.62	36.36	35.83	44.00	46.55	25.00
Butterfly-10K	Open	0.04	0.05	0.00	0.01	0.00	0.00	0.07	0.01	0.00	0.00	9.94	0.00	1.54
	MC	28.91	50.24	44.58	36.45	25.14	28.88	33.06	28.90	25.28	36.67	41.70	34.48	25.00
Trait Identification														
Fish-10K	MC	82.18	56.84	45.15	46.92	68.36	39.33	55.08	51.87	64.34	39.26	81.95	20.69	50.0
Bird-10K	MC	62.22	34.68	46.14	63.93	50.11	41.38	39.11	40.44	47.89	45.52	77.91	89.98	31.12
Trait Grounding														
Fish-500	MC	29.41	24.87	17.98	23.42	23.32	25.14	22.18	25.58	7.20	27.09	33.51	26.90	25.00
Bird-500	MC	8.1	26.92	35.36	23.2	11.83	10.52	15.39	24.22	3.48	0.81	30.24	13.91	25.00
Trait Referring														
Fish-500	MC	28.15	27.07	29.14	28.19	24.93	25.68	39.24	31.21	31.75	25.78	28.04	32.73	25.00
Bird-500	MC	42.28	30.5	29.64	18.45	35.16	40.59	26.04	35.88	27.52	41.69	23.03	22.69	25.00
Trait Counting														
Fish-500	Open	16.4	47.4	52.0	14.8	37.6	63.4	13.6	31.53	50.2	61.4	61.4	0.0	25.00
	MC	44.80	13.20	54.80	21.00	64.8	78.2	22.00	25.00	74.0	69.4	15.80	11.80	25.00
Overall		34.24	29.0	31.78	25.27	28.91	30.24	23.0	25.19	28.49	29.79	33.92	23.31	22.03

Table 2: Zero-shot accuracy comparison of VLM baselines (in % ranging from 0 to 100) for the five scientific tasks. Results are color-coded as Best, Second best, Worst, Second worst.

Zero-shot Evaluation



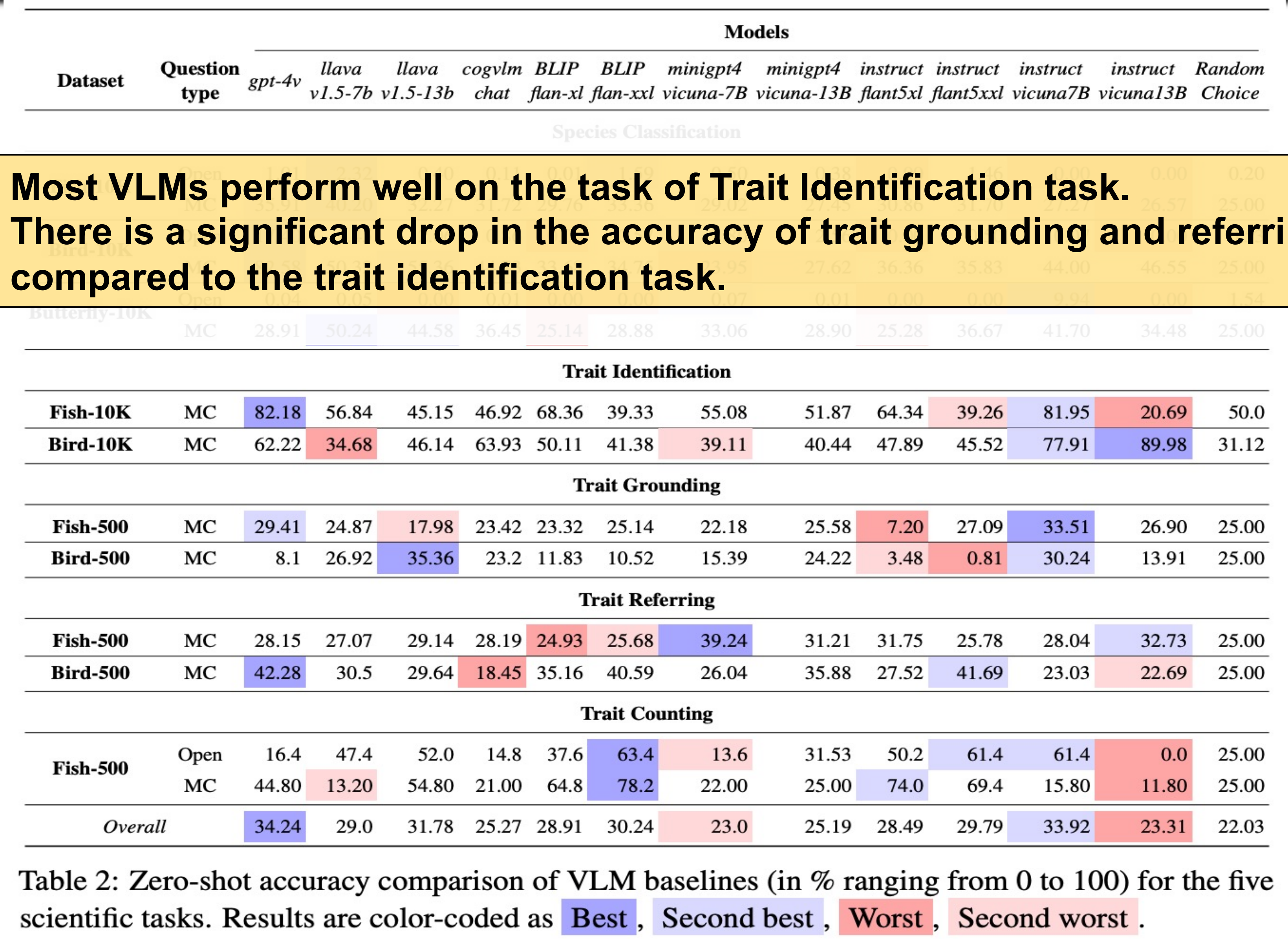
Dataset	Question type	Models												
		gpt-4v	*llava v1.5-7b*	*llava v1.5-13b*	*cogvlm chat*	*BLIP flan-xl*	*BLIP flan-xxl*	*minigpt4 vicuna-7B*	*minigpt4 vicuna-13B*	*instruct flant5xl*	*instruct flant5xxl*	*instruct vicuna7B*	*instruct vicuna13B*	*Random Choice*
Species Classification														
Fish-10K	Open	1.01	2.32	0.40	0.11	0.01	1.59	0.50	0.38	0.00	1.46	0.00	0.00	0.20
	MC	35.91	40.20	32.27	31.72	29.76	33.36	29.02	27.45	30.86	31.70	27.27	26.57	25.00
Bird-10K	Open	17.40	1.45	2.06	0.86	0.00	0.57	2.80	2.56	0.00	0.50	0.07	0.00	0.53
	MC	82.58	50.32	55.36	44.73	33.68	34.75	23.95	27.62	36.36	35.83	44.00	46.55	25.00
Butterfly-10K	Open	0.04	0.05	0.00	0.01	0.00	0.00	0.07	0.01	0.00	0.00	9.94	0.00	1.54
	MC	28.91	50.24	44.58	36.45	25.14	28.88	33.06	28.90	25.28	36.67	41.70	34.48	25.00
Trait Identification														
Fish-10K	MC	82.18	56.84	45.15	46.92	68.36	39.33	55.08	51.87	64.34	39.26	81.95	20.69	50.0
Bird-10K	MC	62.22	34.68	46.14	63.93	50.11	41.38	39.11	40.44	47.89	45.52	77.91	89.98	31.12
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	25.00
Bird-500	MC	8.1	26.92	35.36	23.2	11.83	10.52	15.39	24.22	3.48	0.81	30.24	13.91	25.00
Trait Referring														
Fish-500	MC	28.15	27.07	29.14	28.19	24.93	25.68	39.24	31.21	31.75	25.78	28.04	32.73	25.00
Bird-500	MC	42.28	30.5	29.64	18.45	35.16	40.59	26.04	35.88	27.52	41.69	23.03	22.69	25.00
Trait Counting														
Fish-500	Open	16.4	47.4	52.0	14.8	37.6	63.4	13.6	31.53	50.2	61.4	61.4	0.0	25.00
	MC	44.80	13.20	54.80	21.00	64.8	78.2	22.00	25.00	74.0	69.4	15.80	11.80	25.00
Overall		34.24	29.0	31.78	25.27	28.91	30.24	23.0	25.19	28.49	29.79	33.92	23.31	22.03

Table 2: Zero-shot accuracy comparison of VLM baselines (in % ranging from 0 to 100) for the five scientific tasks. Results are color-coded as Best, Second best, Worst, Second worst.

1. **All VLMs show poor accuracy on open questions than MC Questions.**
2. **Bird dataset shows better accuracy than Fish or Butterfly datasets.**

Zero-shot Evaluation

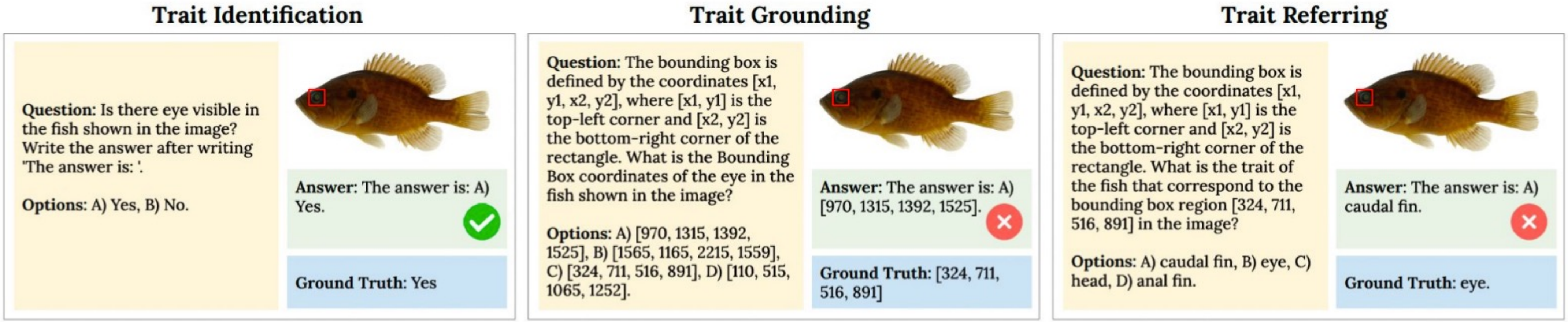
1. **Most VLMs perform well on the task of Trait Identification task.**
2. **There is a significant drop in the accuracy of trait grounding and referring tasks compared to the trait identification task.**



Dataset	Question type	Models												
		gpt-4v	*llava v1.5-7b*	*llava v1.5-13b*	*cogvlm chat*	*BLIP flan-xl*	*BLIP flan-xxl*	*minigpt4 vicuna-7B*	*minigpt4 vicuna-13B*	*instruct flant5xl*	*instruct flant5xxl*	*instruct vicuna7B*	*instruct vicuna13B*	*Random Choice*
Species Classification														
[illegible]	Open	1.01	2.32	0.10	0.11	0.01	1.59	0.50	0.38	[illegible]	1.46	0.00	0.00	0.20
	MC	35.91	40.20	32.27	31.72	29.76	33.36	29.02	27.45	30.86	31.70	27.27	26.57	25.00
Bird-10K	Open	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	MC	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	23.95	27.62	36.36	35.83	44.00	46.55	25.00
Butterfly-10K	Open	0.04	0.05	0.00	0.01	0.00	0.00	0.07	0.01	0.00	0.00	9.94	0.00	1.54
	MC	28.91	50.24	44.58	36.45	25.14	28.88	33.06	28.90	25.28	36.67	41.70	34.48	25.00
Trait Identification														
Fish-10K	MC	82.18	56.84	45.15	46.92	68.36	39.33	55.08	51.87	64.34	39.26	81.95	20.69	50.0
Bird-10K	MC	62.22	34.68	46.14	63.93	50.11	41.38	39.11	40.44	47.89	45.52	77.91	89.98	31.12
Trait Grounding														
Fish-500	MC	29.41	24.87	17.98	23.42	23.32	25.14	22.18	25.58	7.20	27.09	33.51	26.90	25.00
Bird-500	MC	8.1	26.92	35.36	23.2	11.83	10.52	15.39	24.22	3.48	0.81	30.24	13.91	25.00
Trait Referring														
Fish-500	MC	28.15	27.07	29.14	28.19	24.93	25.68	39.24	31.21	31.75	25.78	28.04	32.73	25.00
Bird-500	MC	42.28	30.5	29.64	18.45	35.16	40.59	26.04	35.88	27.52	41.69	23.03	22.69	25.00
Trait Counting														
Fish-500	Open	16.4	47.4	52.0	14.8	37.6	63.4	13.6	31.53	50.2	61.4	61.4	0.0	25.00
	MC	44.80	13.20	54.80	21.00	64.8	78.2	22.00	25.00	74.0	69.4	15.80	11.80	25.00
Overall		34.24	29.0	31.78	25.27	28.91	30.24	23.0	25.19	28.49	29.79	33.92	23.31	22.03

Table 2: Zero-shot accuracy comparison of VLM baselines (in % ranging from 0 to 100) for the five scientific tasks. Results are color-coded as Best, Second best, Worst, Second worst.

Trait Identification vs. Trait Detection



Trait Identification

Question: Is there eye visible in the fish shown in the image? Write the answer after writing 'The answer is: '.

Options: A) Yes, B) No.

Answer: The answer is: A) Yes. ✓

Ground Truth: Yes

Trait Grounding

Question: The bounding box is defined by the coordinates [x1, y1, x2, y2], where [x1, y1] is the top-left corner and [x2, y2] is the bottom-right corner of the rectangle. What is the Bounding Box coordinates of the eye in the fish shown in the image?

Options: A) [970, 1315, 1392, 1525], B) [1565, 1165, 2215, 1559], C) [324, 711, 516, 891], D) [110, 515, 1065, 1252].

Answer: The answer is: A) [970, 1315, 1392, 1525]. ✗

Ground Truth: [324, 711, 516, 891]

Trait Referring

Question: The bounding box is defined by the coordinates [x1, y1, x2, y2], where [x1, y1] is the top-left corner and [x2, y2] is the bottom-right corner of the rectangle. What is the trait of the fish that correspond to the bounding box region [324, 711, 516, 891] in the image?

Options: A) caudal fin, B) eye, C) head, D) anal fin.

Answer: The answer is: A) caudal fin. ✗

Ground Truth: eye.

Figure 3: Examples of correct and incorrect predictions of GPT-4V for trait identification, trait grounding, and trait-referring tasks related to the "eye". For visualization assistance, a red-colored bounding box is added around the "eye" in the image.

Effects of Prompting on VLM Performance

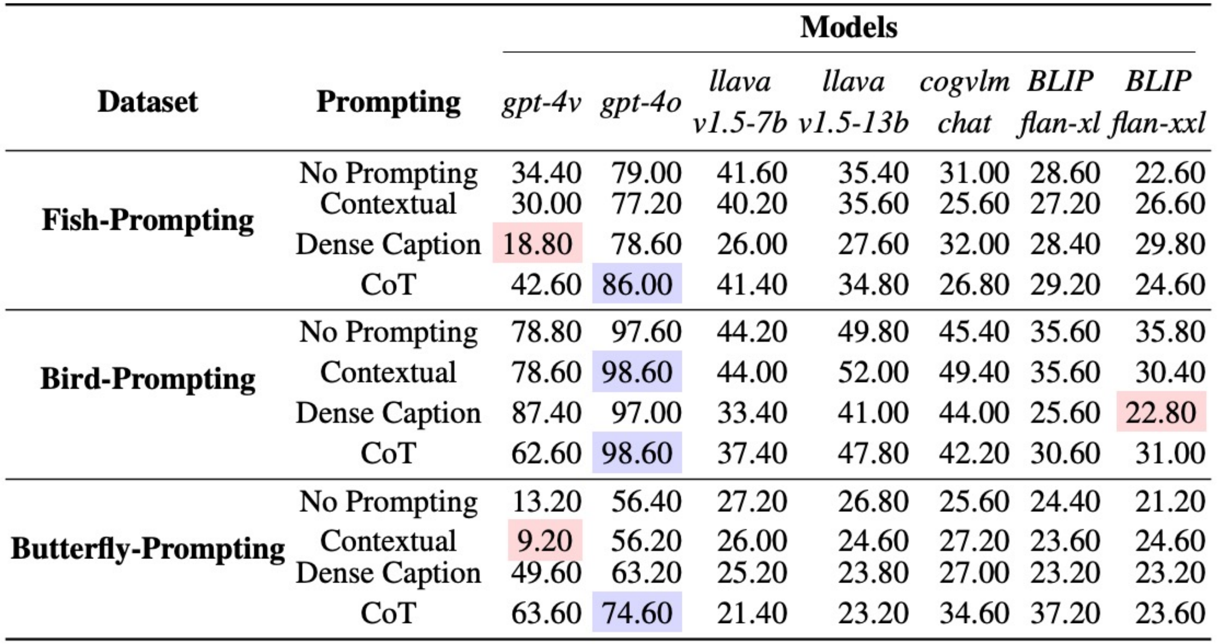
Three prompting techniques:

1. **Contextual Prompting:**
   - We provided a single-line description of the tasks with the question.
   - For example, for species classification task:
     - *Each biological species has a unique scientific name composed of two parts: the first for the genus and the second for the species within that genus.*

2. **Dense Caption Prompting:**
   - We prompt the VLM to generate a dense caption for the specimen image.
   - We add the dense caption before the question and prompt, "*Use the above dense caption and the image to answer the following question.*" to generate responses.

3. **Chain-of-Thought (CoT) Prompting:**
   - We prompt "*Let's think step by step*" to the VLM to generate the reasoning for a given VQA and multiple choices.
   - We then add the reasoning after the VQA and prompt, *"Please consider the following reasoning to formulate your answer."* to generate the VLM response.

Effects of Prompting on VLM Performance

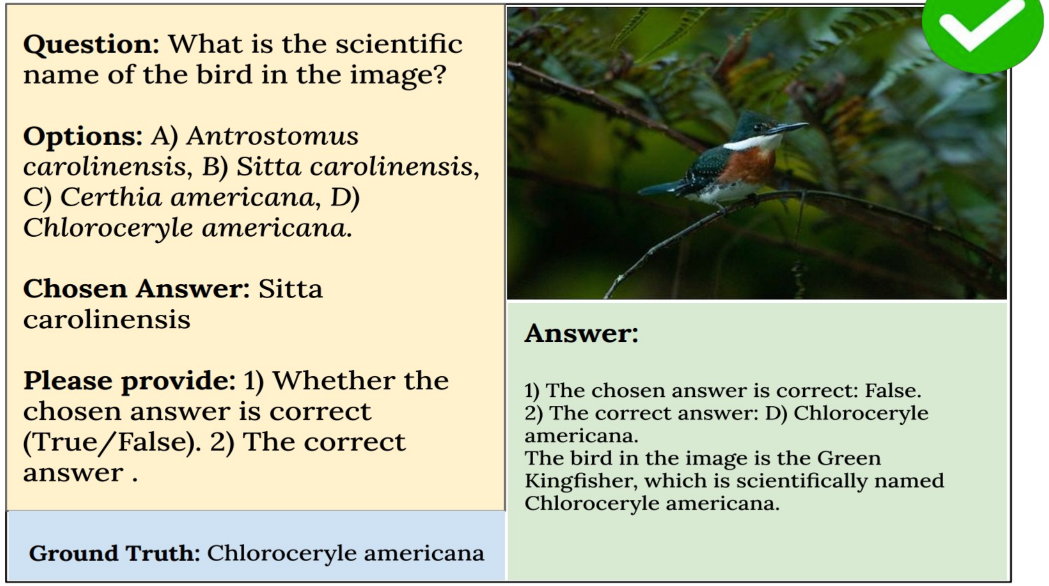


Dataset	**Prompting**	**Models**						
		gpt-4v	*gpt-4o*	*llava v1.5-7b*	*llava v1.5-13b*	*cogvlm chat*	*BLIP flan-xl*	*BLIP flan-xxl*
Fish-Prompting	No Prompting	34.40	79.00	41.60	35.40	31.00	28.60	22.60
	Contextual	30.00	77.20	40.20	35.60	25.60	27.20	26.60
	Dense Caption	18.80	78.60	26.00	27.60	32.00	28.40	29.80
	CoT	42.60	86.00	41.40	34.80	26.80	29.20	24.60
Bird-Prompting	No Prompting	78.80	97.60	44.20	49.80	45.40	35.60	35.80
	Contextual	78.60	98.60	44.00	52.00	49.40	35.60	30.40
	Dense Caption	87.40	97.00	33.40	41.00	44.00	25.60	22.80
	CoT	62.60	98.60	37.40	47.80	42.20	30.60	31.00
Butterfly-Prompting	No Prompting	13.20	56.40	27.20	26.80	25.60	24.40	21.20
	Contextual	9.20	56.20	26.00	24.60	27.20	23.60	24.60
	Dense Caption	49.60	63.20	25.20	23.80	27.00	23.20	23.20
	CoT	63.60	74.60	21.40	23.20	34.60	37.20	23.60

Table 4: Zero-shot accuracy comparison for different prompting techniques of seven VLMs (in % ranging from 0 to 100). Results are color-coded as Best and Worst.

Tests for Reasoning Hallucination

1. False Confidence Test (FCT)



Question: What is the scientific name of the bird in the image?

Options: A) *Antrostomus carolinensis*, B) *Sitta carolinensis*, C) *Certhia americana*, D) *Chloroceryle americana.*

Chosen Answer: Sitta carolinensis

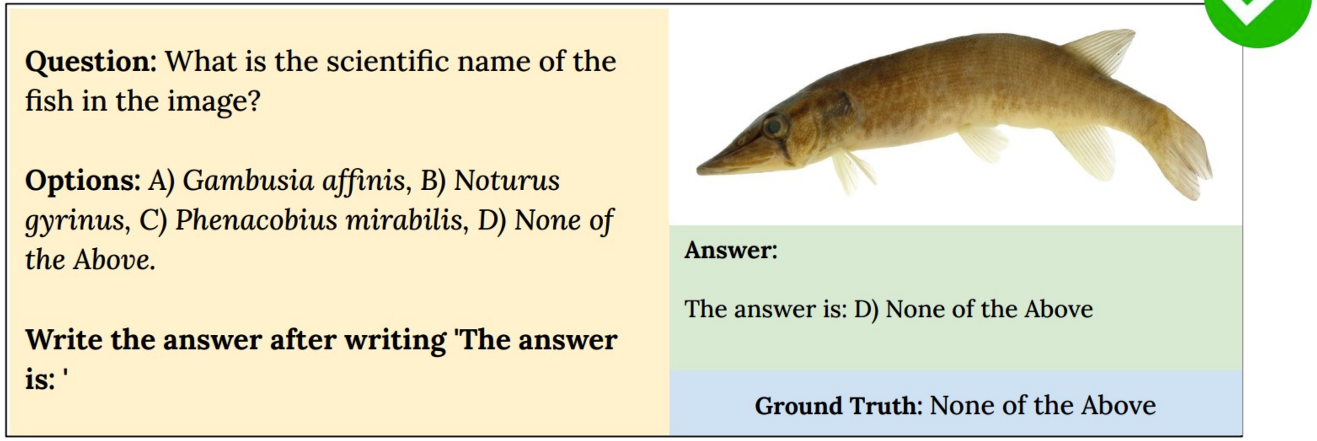
Please provide: 1) Whether the chosen answer is correct (True/False). 2) The correct answer .

Ground Truth: Chloroceryle americana

Answer:

1) The chosen answer is correct: False.
2) The correct answer: D) Chloroceryle americana.
The bird in the image is the Green Kingfisher, which is scientifically named Chloroceryle americana.

2. None of the Above (NOTA) Test



Question: What is the scientific name of the fish in the image?

Options: A) *Gambusia affinis*, B) *Noturus gyrinus*, C) *Phenacobius mirabilis*, D) *None of the Above.*

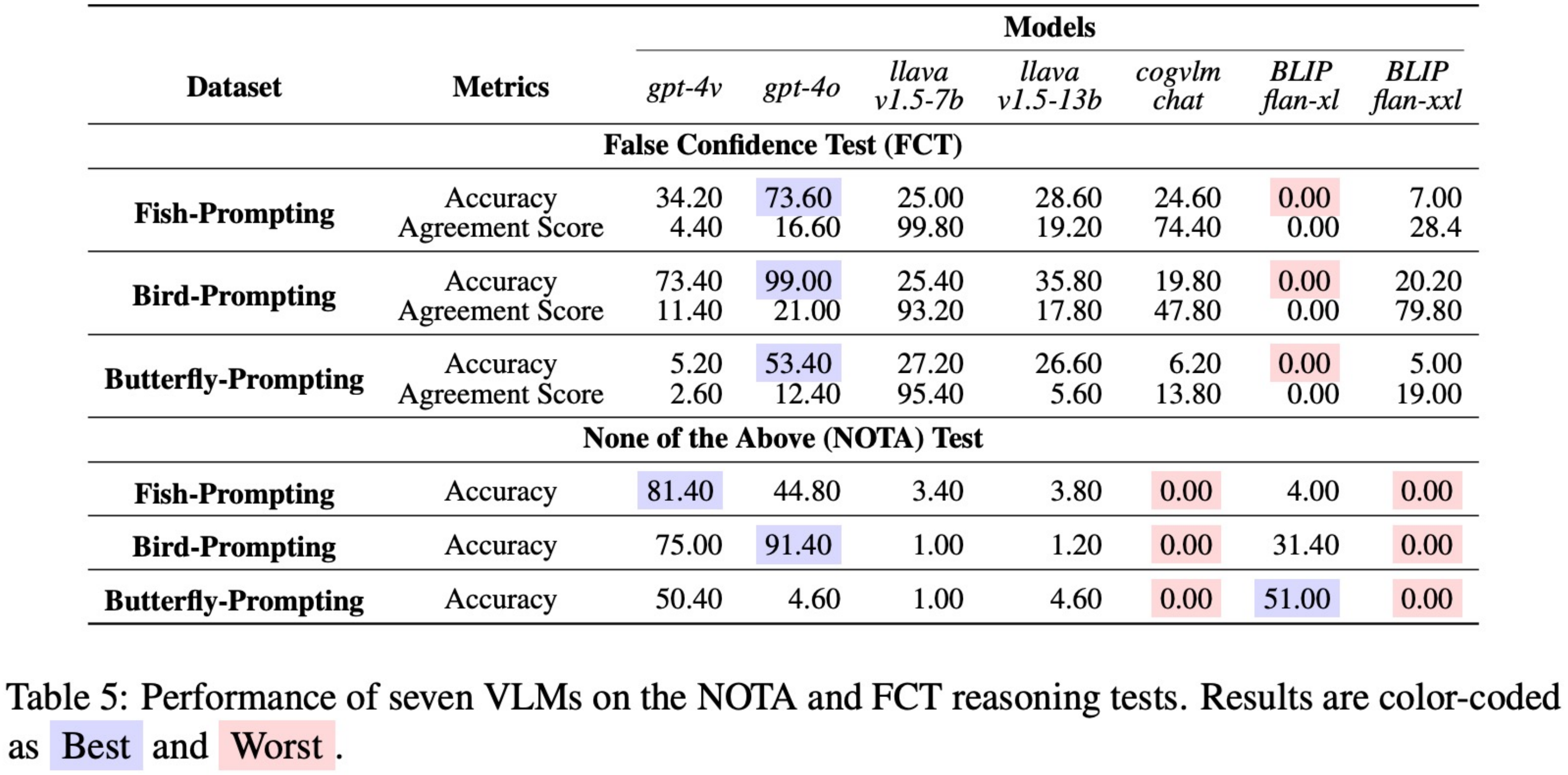
Write the answer after writing 'The answer is: '

Answer:

The answer is: D) None of the Above

Ground Truth: None of the Above

Tests for Reasoning Hallucination



Dataset	Metrics	Models *gpt-4v*	 *gpt-4o*	 *llava v1.5-7b*	 *llava v1.5-13b*	 *cogvlm chat*	 *BLIP flan-xl*	 *BLIP flan-xxl*
False Confidence Test (FCT)								
Fish-Prompting	Accuracy	34.20	73.60	25.00	28.60	24.60	0.00	7.00
	Agreement Score	4.40	16.60	99.80	19.20	74.40	0.00	28.4
Bird-Prompting	Accuracy	73.40	99.00	25.40	35.80	19.80	0.00	20.20
	Agreement Score	11.40	21.00	93.20	17.80	47.80	0.00	79.80
Butterfly-Prompting	Accuracy	5.20	53.40	27.20	26.60	6.20	0.00	5.00
	Agreement Score	2.60	12.40	95.40	5.60	13.80	0.00	19.00
None of the Above (NOTA) Test								
Fish-Prompting	Accuracy	81.40	44.80	3.40	3.80	0.00	4.00	0.00
Bird-Prompting	Accuracy	75.00	91.40	1.00	1.20	0.00	31.40	0.00
Butterfly-Prompting	Accuracy	50.40	4.60	1.00	4.60	0.00	51.00	0.00

Table 5: Performance of seven VLMs on the NOTA and FCT reasoning tests. Results are color-coded as Best and Worst.

Thank you for listening.

Please visit us during the poster session.

Paper

Code

Hugging Face Dataset