

species classification

Exp. 1

No prompt

Exp. 2

No prompt
+
Easy, Medium,
Hard Question
set

Exp. 3

Three different
prompting
Contextual, Dense
Captioning, CoT

Exp. 4

Affects of
hallucination
(i) FCT
(ii) NOTA

(optional)
Exp. 5

Comparing pre-trained
VLMs with
Biologically fine
tuned Model

Dataset Creation

1. No prompt

we consider asking a VLM to provide the scientific name of the organism shown in a given image. There are two types of questions that we consider for this task. First, we consider **open-ended questions**, where we do not provide any answer choices (or options) to the VLM in the input prompt. The second type is **multiple-choice (MC) questions**, where we provide four choices of candidate species names for the VLM to choose from (out of which only one is correct while the remaining three are randomly selected from the set of all species classes).


The scientific names for the images of Fish-10K and Butterfly-10K were obtained directly from their

Task	Prompt Format
Species Classification	<image> What is the scientific name of the <organism> shown in the image? <options> Write the answer after writing the answer is: .

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

Options: A) Cyprinella spiloptera, B) Phenacobius mirabilis, C) Notropis hudsonius, D) Notropis buccatus.

Write the answer after writing 'The answer is: '



Answer:

The answer is: B) Phenacobius mirabilis

Ground Truth: Phenacobius mirabilis

Figure 10: No Prompting. GPT-4o Correct prediction. Section M.1.

respective sources. For Bird-10K, we obtained the scientific names from the iNatLoc500 [62] dataset.

We curated around **31K question-answer pairs in both open and multiple choice (MC) question formats for evaluating species classification tasks**

2. Easy, Medium & Hard Questions

A potential hypothesis for this observation is that VLMs are able to avoid incorrect answer choices (or options) that are too different from the correct option and thus are easy to eliminate. To test this hypothesis, we create three variants of the MC questions for species classification—easy, medium, and hard—where **species choices in each variant have varying degrees of similarity determined by their taxonomic groupings**. In particular, note that the scientific name of an organism contains taxonomic information at three levels: <genus name> <species name> <subspecies name>

Since organisms that share taxonomic information have similar appearances, it is hard to differentiate species choices if they are from the same taxonomic group. On the other hand, it is easier to work with species choices from different taxonomic groups.

Hence, **for the easy set, we selected 50 species from different genera, ensuring that all species choices appear quite different from each other.**

For the medium set, we increased the complexity by constructing species choices from the same genus but from 10 different species.

The hard set presented the highest difficulty level for the butterfly dataset, with the answer choices being from the same genus and species but from 10 subspecies.

Each difficulty level consists of 200 images from each set of organisms

Note – hard questions are only available for butterfly , reason -

4We only have subspecies level information for the Butterfly-10K dataset


3. Different prompting

Contextual prompting, we provided a single-line description (context) of the tasks (e.g., we add “Each biological species has a unique scientific name composed of two parts: the first for the genus and the

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

Options: A) *Lepomis cyanellus*, B) *Notropis biennius*, C) *Esox lucius*, D) *Gambusia affinis*.

Write the answer after writing 'The answer is: '



Answer:

The answer is: *Esox lucius*

Ground Truth: *Esox lucius*

Figure 14: Contextual Prompting. GPT-4o Correct prediction. Section M.2.

Dense Caption: The image showcases a small fish with a slender, elongated body shape. The body is covered in yellowish-brown hues, with a gradient that lightens towards the ventral side and deepens dorsally. The scales are distinctly visible, giving a textured appearance to the body. Starting from the head, the fish has a rounded snout and a relatively large, prominent eye with a dark pupil and a lighter surrounding area. The mouth is small and positioned slightly downturned. The dorsal fin is placed towards the rear of the body, exhibiting a rather elongated and low profile with delicate, translucent rays. The anal fin mirrors the dorsal fin but is placed closer to the rear ventral side, displaying similar morphological traits. The pectoral fin are modest in size, almost triangular, and located near the gill covers. The pelvic fins are also small and positioned ventrally, slightly behind the pectoral ones. The caudal fin is fan-shaped and extends symmetrically, marked by an array of thin rays. This fin appears semi-transparent with slight yellowish tinges that match the overall body coloration. Other distinctive features include a lateral line that is faintly visible and runs along the midline of the body. The skin reveals an intricate network of fine lines, which might suggest sensory functions in identifying environmental changes. Given these detailed anatomical descriptions, the fish could potentially be identified as a member of the Percellidae family, possibly a species of *Gambusia* or *Heterandria*.

Question: Use the above dense-caption and the image to answer the following question. What is the scientific name of the fish in the image?

Options: A) *Notropis volucellus*, B) *Notropis percbromus*, C) *Gambusia affinis*, D) *Lepomis gibbosus*.

second for the species within that genus.” before the species classification question to give some additional context on the task).

Dense Caption prompting involves two stages: (1) first, we prompt the VLM to generate a dense caption of the specimen image such that the caption contains all the necessary trait information of the specimen. (2) 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 from the VLM. Similarly, the **Zero-Shot Chain-of-Thought (CoT)** happens in two stages: (1) first, we prompt the VLM to generate the reasoning for a given VQA and multiple choices (options). Zero-shot CoT appends “Let’s think step by step.” after the question and options to generate the reasoning. (2) We then add the reasoning after the VQA and prompt “Please consider the following reasoning to formulate your answer” to generate the VLM response.

We curated a prompting dataset of **500 multiple-choice (MC) VQAs** for each set of organisms, which is a subset of the VLM4Bio dataset for species classification.

4. Hallucination Test

For FCT we randomly select an option from the list of given choices and prompt it to the VLM as a “suggested correct answer” along with the question and options.

To evaluate VLMs on FCT, we use Accuracy as well as the Agreement score, which is the percentage of times the VLM agrees with the suggested answer, irrespective of whether that is right or wrong. A high agreement score with a low overall accuracy indicates poor performance as it suggests that the model is simply following the suggestion either because of a lack of knowledge or low confidence in its own response.

On the other hand, in the **NOTA Test**, we replace the correct option with “None of the Above”, requiring the model to produce “None of the above” for all the questions

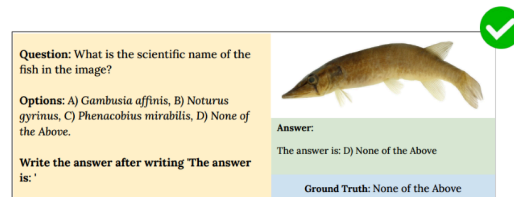


Figure 32: GPT-4o Correct prediction. Actual species name is Esox Americanus. Section N.2.

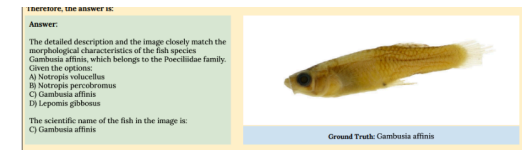


Figure 18: Dense Captions in Prompts. GPT-4o Correct prediction. Section M.3.

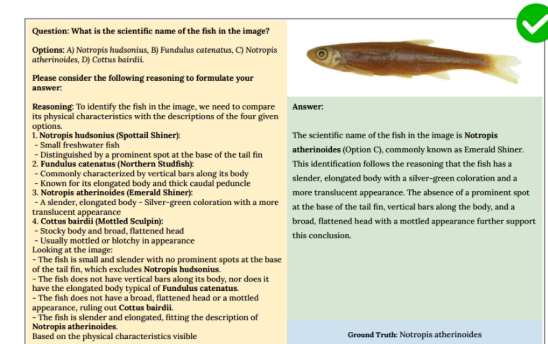


Figure 22: Chain-Of-Thought Prompting. GPT-4o Correct prediction. Section M.4.

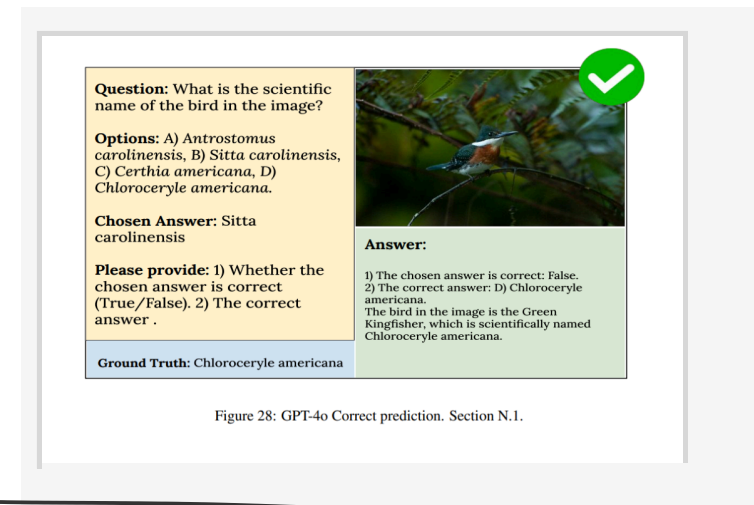


Figure 28: GPT-4o Correct prediction. Section N.1.

Additional :

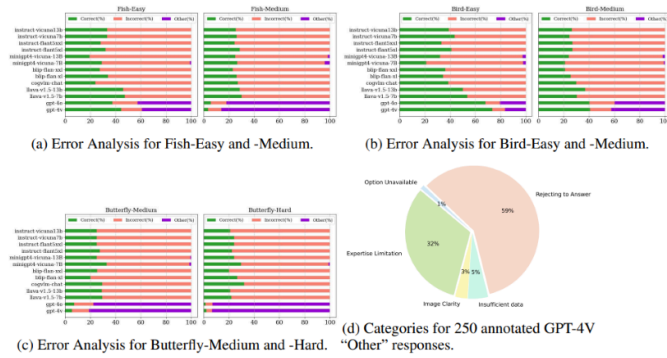


Figure 8: Analysis of errors for the pretrained VLM responses.

include *Image Clarity* issues and *Option Unavailable* (i.e., GPT-4V could not find a suitable option from the list of options provided in the prompt).

Dataset	Question type	Models				
		<i>gpt-4v</i>	<i>llava v1.5-7b</i>	<i>cogvlm chat</i>	<i>CLIP</i>	<i>BioCLIP</i>
Species Classification						
Fish-10K	Open	1.01	2.32	0.11	0.57	1.24
	MC	35.91	40.20	31.72	42.45	50.65
Bird-10K	Open	17.40	1.45	0.86	7.74	67.12
	MC	82.58	50.32	44.73	45.78	93.93
Butterfly-10K	Open	0.04	0.05	0.01	5.33	15.95
	MC	28.91	50.24	36.45	45.60	62.32

Table 7: Zero-shot accuracy comparison of VLM baselines (in % ranging from 0 to 100) with BioCLIP for the species classification task. Results are color-coded as **Best**, and **Worst**.