

# Fine-grained evaluation of VLMs using LLM-as-Judge

Sarang Ravi Chouguley

<sup>1</sup>Institute for Information Systems, Hof University of Applied Sciences,  
Hof, Germany.

## Abstract

Evaluation of visual language models is a challenging task due to their input output format. The aim of this study is to understand if LLM-as-a-Judge approach can be used to perform fine-grained evaluation of VLMs, with small language models. The study evaluates responses generated by a VLM with focus on image identification task. Non-fine-tuned language models with less than 15B parameters were used as LLM-Judge. This study found that small non-fine-tuned language models can achieve high accuracy (94.3 percent) and correlation (0.91) to human evaluation using fine-grained evaluation criteria. The limitations of such evaluation include the limitations of the language model being used along with the need to design and annotate custom ground truth and evaluation criteria.

**Keywords:** Language Model, Visual Language Model, LLM-as-a-judge, fine-grained evaluation

## 1 Introduction

Large language models have become pretty popular now. Seeing the capabilities of these models, researchers combined them with other forms of input like audio, video or image. This combination of input and output gave rise to multi-modal language models like Visual-Language Models (VLMs) or Large Visual Language Models (LVLMs).

Large Vision-Language models have various multi-modal capabilities. Some of the capabilities are visual perception, visual knowledge acquisition, visual reasoning, visual common sense, embodied intelligence. Using these capabilities, a lot of visual language models have achieved remarkable progress in tasks like visual question answering and multi-modal conversation. Basically, LVLMs capitalise on the knowledge from LLMs

and align visual features with textual space. Falmingo [1], a pioneering LVLM, integrates visual features into LLMs through cross-attention layers. Later studies proposed more efficient vision-text interactions [10] more efficient training methods [3, 21] and employing instruction tuning [3, 12, 21, 23].

However, despite the great success, few efforts have been made to provide systematic evaluation of VLMs. Evaluations plays a crucial role in understanding their strengths and weaknesses. Some work has been done to create techniques and tools for evaluating these VLMs. But these tools, techniques face following challenges in evaluation:

- **Coherent Sentence Responses:** Sometimes, the generated responses from the models deviate from the standard image classification benchmarks and provide coherent sentence-style answers.
- **Diverse response structure:** VLMs generate text in diverse and different styles from the ground truth.

Recent work[20] has tried to counter these challenges in following ways:

- **Model based evaluation:** This uses models like sentence transformers to calculate feature similarity between ground-truth and generated answers. It is generally more robust but sometimes suffers due to model limitations.
- **Language model based evaluation:** Some studies have used powerful models like GPT-4 as a judge to evaluate VLMs response. However this method also has limitations like dependence on proprietary models, etc.
- **Human based evaluation:** In this method, the model’s responses are manually judged by humans. This is the most common method and provides best results. The main drawback of this method is dependency on humans, which makes this method slow and cumbersome for large scale evaluation.

This project is undertaken to understand language model based evaluation of VLMs using small language models. This project aims to find answers to the following research questions:

- **Q1.** How effective is LLM-as-a-judge technique for VLM evaluation ?
- **Q2.** What are the challenges and limitations of LLM-as-a-judge in context of VLM evaluation ?
- **Q3.** What techniques and measures can be taken to improve LLM-as-a-judge based evaluation of VLM ?
- **Q4.** Do we really need complex fine-grained evaluation criteria instead of a simple string match scoring ?

This study is restricted to only evaluation of small sized language models. Number of models used as LLM Judge is restricted to 8, to keep it manageable. The model parameter size is restricted to 15B and below.

The rest of the paper is structured in following manner. First it discusses related work. Then, it discusses the dataset used for evaluation, choice of models, evaluation

setup and evaluation task definition. After that, experiment results and analysis are presented. The paper ends with discussion, limitations, conclusion and outlook.

## 2 Related Work

Following discusses some recent related research in visual language models / multi-modal language models evaluations and LLM-as-Judge based evaluations.

### 2.1 Recent trends in VLM evaluation

Xu, Peng et al.[20] released Lvlm-ehub, a comprehensive evaluation benchmark for LVLMS which provides a set of techniques to evaluate different abilities of VLMs including visual perception. It uses different datasets like ImageNet1K[15], CIFAR10[8], Pets37[14] and Flowers102[13] for evaluation. Metrics like Object counting, Top-1 accuracy, per-class accuracy, F1 score, CIDEr score (measures the similarities between generated and ground-truth answers) are used for measuring accuracy in these evaluations[20]. Another benchmark, Mementos[19] evaluates multimodal large language models (MLLMs) like GPT-4v and Gemini on their ability to describe dynamic image sequences. It assesses models using metrics such as recall, precision, and F1 score. Lvlm-ehub also proposed LVLm Arena, an innovative evaluation framework that utilizes a 1v1 LVLm battle with human judgment for open-world evaluation, leading to more accurate and realistic evaluation results. Gunjal et al.[4] released M-HalDetect dataset which can be used to evaluate and mitigate hallucinations in LVLm-generated image descriptions. This dataset includes fine-grained annotations for identifying accurate and inaccurate segments within generated descriptions. In [11], the evaluation metric "POPE" is proposed to evaluate hallucinations in LVLms by polling questions about generated text.

### 2.2 Recent trend in LLM-as-Judge evaluation

Some research has also been done in the domain of using one language model to evaluate another. Studies have shown that using LLMs like GPT-4 as judges can match human preferences well, achieving over 80 percent agreement with human judgements. This approach is scalable and can approximate human evaluations, making it a practical solution for assessing LLM outputs[16]. Other works like [18] showed that even though these LLM can achieve high correlation to human judgements, they are not fail evaluators and contain certain bias like verbosity bias, beauty bias and positional bias in evaluations. Another work by Verga et al.[17] shows that instead of relying on a single large model, using a panel of diverse, small models (PoLL) can reduce intra-model bias and improve evaluation reliability. An empirical study[5] showed that LLM-as-Judge can be used in different ways like pointwise grading, pairwise comparisons and multi-turn evaluations. Studies indicate that fine-tuned judge models perform best in the evaluation schemes they were trained on, but may lack generalisability across different evaluation tasks. This led to the creation of Prometheus v1 and v2 and PrometheusVision[6, 7, 9], which are fine-tuned versions of Llama-2, specifically tuned for acting as fine-grained evaluator, with performance equal to or sometimes

better than GPT-4. Some studies[2, 16, 22] have also used standard benchmarks like TriviaQA and MT-bench to evaluate LLM judges.

### 3 Dataset

The dataset used in this project consist of 25 responses generated by LLaVA v1.5 Vicuna 7B. The task for VLM was "given a image of an animal, identify the animal in the given image as specifically as possible".

### 4 Models selected for Study

8 Models were selected for the study to keep it manageable. Only models which are publicly available on hugging-face are selcted. While choosing the models, the following criteria are kept in mind:

- Model having no more than 15B parameters.
- Models are recently released.
- Models which are non-fine-tuned on the task.
- Give preference to the most popular models on hugging-face.

Refer appendix A1 for model details.

### 5 Experiment

#### 5.1 Setup

The prompt template provided by hugging face was used for each model to avoid invalid responses generated due to incorrect prompt template. NVIDIA A100 40GB GPU was used for generating the response. The responses generated by each of the models were collect for evaluation.

#### 5.2 Task Definition

For a given input string and ground truth, LM should provide a detailed(fine-grained) feedback and score according to the evaluation criteria.

#### 5.3 Evaluation 1

Using a simple scoring criteria (inspired by Prometheus score rubric[6]) and ground truth consisting of class, family and species of the animal, evaluate the VLM responses. Only a subset of dataset was used to quickly validate the accuracy of this evaluation strategy.

##### 5.3.1 Scoring criteria

The scoring criteria used is as follows:

- Score 0: The model fails to identify the Class, Family, or Species of the animal.
- Score 1: The model correctly identifies the Class of the animal.
- Score 2: The model correctly identifies the Family of the animal.
- Score 3: The model correctly identifies the Species of the animal.

### 5.3.2 Models used

- Prometheus-13B-v1.0
- Prometheus-7B-v2.0
- Mistral-7B-Instruct-v0.3
- Llama-3-8B-Instruct

### 5.3.3 Prompt structure

The prompt used in this evaluation is based on Prometheus prompt structure[6]. The prompt consists of a system instruction (asking LM to ask as a fair LLM evaluator), a task description (explaining the task and how to evaluate and give a response), the question given to VLM(instruction to evaluate), VLM response (response to evaluate), ground truth (reference answer) and the scoring criteria (score rubric). For detailed prompt structure please refer to A.1.

### 5.3.4 Evaluation Trials

One evaluation trial was conducted, where each of the 4 models were given the prompt, and responses were recorded for 4 data points. Refer to A7 for dataset details.

### 5.3.5 Evaluation Issues

Analysing the responses in detail reveals the following issues with this evaluation strategy. The models are hallucinating and unable to correctly classify the class, family and species. The scoring criteria fails to cover all the cases present in dataset.

## 5.4 Evaluation 2

Learning from the issues identified in evaluation 1, evaluation 2 is created with better evaluation process and scoring criteria. Taking inspiration from techniques like chain-of-thoughts and task-specific-agents, the evaluation is divided into 3 steps. Each step performs one specific action.

Step 1: Extract keywords related to the class, family and species of the animal from the VLM response.

Step 2: Classify the extracted keywords into Class, Family and Species.

Step 3: Evaluate the response using the classification from step 2, ground truth and score matrix and provide a score and detailed feedback for each score.

### 5.4.1 Scoring criteria

The scoring criteria is redesigned version of previous criteria. The key idea here is that a score should be assigned for each identifier and the score for each identifier should be independent from other scores. A score should be assigned if identifier is present and correct, present and incorrect and not present. Fig 1 shows the final score matrix.

Score id	Scoring criteria	Scores
<b>S1</b>	Does the model identify class of animal ?	1: for correct 0: for none -1: for incorrect
<b>S2</b>	Does the model identify family of animal ?	1: for correct 0: for none -1: for incorrect
<b>S3</b>	Does the model identify species of the animal ?	1: for correct 0: for none -1: for incorrect

**Fig. 1** Score Matrix evaluation 2

### 5.4.2 Models used

Prometheus v1 and v2 were excluded from this evaluation, since they cannot be used in step-wise evaluation with different scoring criteria, other than the one they were fine-tuned on. Following were the models included in this evaluation.

- Mixtral-8x7B-Instruct-v0.1
- Llama-2-13B-chat
- Phi-3-medium-128k-instruct

### 5.4.3 Prompt structure

Improving on the issues with eval-1 prompt. This prompt consist of a task description (consisting of the instruction to perform the task), and 3 examples (to reduce the hallucinations and inconsistency in performing the task). 3 prompts are designed specific to each step. For exact prompt refer to appendix [A.2](#) [A.3](#) [A.4](#)

### 5.4.4 Evaluation Trials

Two trials of this evaluation were conducted. Trial 1 consisted of 4 data points and trial 2 consisted of 25 data points. In each trial, the 3 models were given the prompt and responses for each evaluation step was recorded. Refer to [A8](#) for dataset details.

### 5.4.5 Evaluation Issues

Analysing the responses in detail reveals following issues with this strategy. The scores do not match well with human scores when scaled to 25 data points. The models sometimes perform well in one step but fail in other steps. The scoring criteria fails to capture other features of the response like overall accuracy, relevance and other descriptive details of the animal.

## 5.5 Evaluation 3

Learning from the issues identified in evaluation 1 and evaluation 2, evaluation 3 is created with better evaluation process and scoring criteria. This evaluation strategy combines the step-wise evaluation into a single comprehensive prompt. Also the scoring criteria is updated to include all the cases and features of the response.

### 5.5.1 Scoring criteria

The scoring criteria is redesigned version of previous criteria. The key idea here is to have a dedicated metric for each feature of the response. Each metric should have a score for all the possible scenarios (cases). Fig 2 shows the final score matrix.

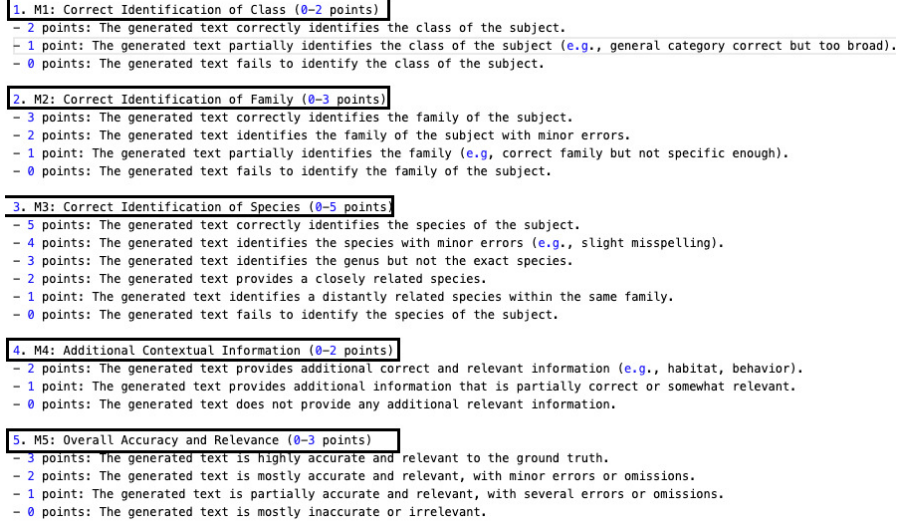
- 
- 1. M1: Correct Identification of Class (0-2 points)**
    - 2 points: The generated text correctly identifies the class of the subject.
    - 1 point: The generated text partially identifies the class of the subject (e.g., general category correct but too broad).
    - 0 points: The generated text fails to identify the class of the subject.
  - 2. M2: Correct Identification of Family (0-3 points)**
    - 3 points: The generated text correctly identifies the family of the subject.
    - 2 points: The generated text identifies the family of the subject with minor errors.
    - 1 point: The generated text partially identifies the family (e.g., correct family but not specific enough).
    - 0 points: The generated text fails to identify the family of the subject.
  - 3. M3: Correct Identification of Species (0-5 points)**
    - 5 points: The generated text correctly identifies the species of the subject.
    - 4 points: The generated text identifies the species with minor errors (e.g., slight misspelling).
    - 3 points: The generated text identifies the genus but not the exact species.
    - 2 points: The generated text provides a closely related species.
    - 1 point: The generated text identifies a distantly related species within the same family.
    - 0 points: The generated text fails to identify the species of the subject.
  - 4. M4: Additional Contextual Information (0-2 points)**
    - 2 points: The generated text provides additional correct and relevant information (e.g., habitat, behavior).
    - 1 point: The generated text provides additional information that is partially correct or somewhat relevant.
    - 0 points: The generated text does not provide any additional relevant information.
  - 5. M5: Overall Accuracy and Relevance (0-3 points)**
    - 3 points: The generated text is highly accurate and relevant to the ground truth.
    - 2 points: The generated text is mostly accurate and relevant, with minor errors or omissions.
    - 1 point: The generated text is partially accurate and relevant, with several errors or omissions.
    - 0 points: The generated text is mostly inaccurate or irrelevant.

Fig. 2 Score Matrix evaluation 3

### 5.5.2 Models used

In addition to the models used in previous evaluation, one more model (Hermes-2-Theta-Llama-3-8B) was included in this evaluation.

### 5.5.3 Prompt structure

Improving on the issues with eval-1 and eval-2 prompts. This prompt consist of a system instruction (instructing the llm to act as expert judge), Scoring criteria (Evaluation Criteria), 3 examples (to reduce the hallucinations and inconsistency in performing the task), Evaluation Steps (instructing the model how to think and evaluate), output format instructions and ground truth. For exact prompt please refer to appendix A.5

### 5.5.4 Evaluation Trials

One trial was conducted, consisting of all the 25 data points in dataset. The response from each model was recorded.

### 5.5.5 Evaluation Issues

This evaluation produced the least issues and errors compared to the previous evaluations. Still the issues consisted of hallucinations and incorrect evaluations.

## 6 Results and Analysis

### 6.1 Evaluation 1

Fig 3 shows the scores given by the models and the human evaluator. Fig 4 shows the Root mean square error values for each model, compared to human evaluator.

Model	DSI1	DSI2	DSI3	DSI4
M1 (Prometheus-13b-v1.0)	2	3	3	1
M2 (Prometheus-7b-v2.0)	2	1	0	1
M3 (Mistral-7B-Instruct-v0.3)	1	1	1	1
M4 (Meta-Llama-3-8B-Instruct-AQLM)	NA (-1)	3	0	NA (-1)
Human Standard	1	1	2	1

Fig. 3 Evaluation 1 Results

Model	M1 (Prometheus-1)	M2 (Prometheus-2)	M3 (Mistral)	M4 (Llama-3)
Root mean square error (RMSE)	1.22	1.12	0.5	2.0

Fig. 4 Evaluation 1 results RMSE analysis

### 6.2 Evaluation 2

Fig 5 shows the scores given by the models and the human evaluator. Fig 6 shows the Root mean square error values for each model, compared to human evaluator scores. From the rmse analysis it is clear that llama-2 with better prompt (slight change in step 3 prompt) performs well for 4 data points, but the error increases to 1.16 from 0.50 when the evaluation is scaled to 25 data points for the same model. Refer to A6 for detailed scores.



Model	DSI1	DSI2	DSI3	DSI4
M1 (Mixtral)	S1: 0, S2: -1, S3: 0	S1: 1,S2: -1,S3: -1	S1: 1,S2: -1,S3: -1	S1: 1,S2: -1,S3: 0
M2 (Llama-2)	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 1	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 0
M3 (Llama-2 with better prompt)	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 1	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 0
M4 (Phi-medium)	S1: 1, S2: -1, S3: 0	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: -1
Human Standard	S1: 1, S2: -1, S3: 0	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: -1	S1: 1, S2: -1, S3: 0

Fig. 5 Evaluation 2 results for 4 data points

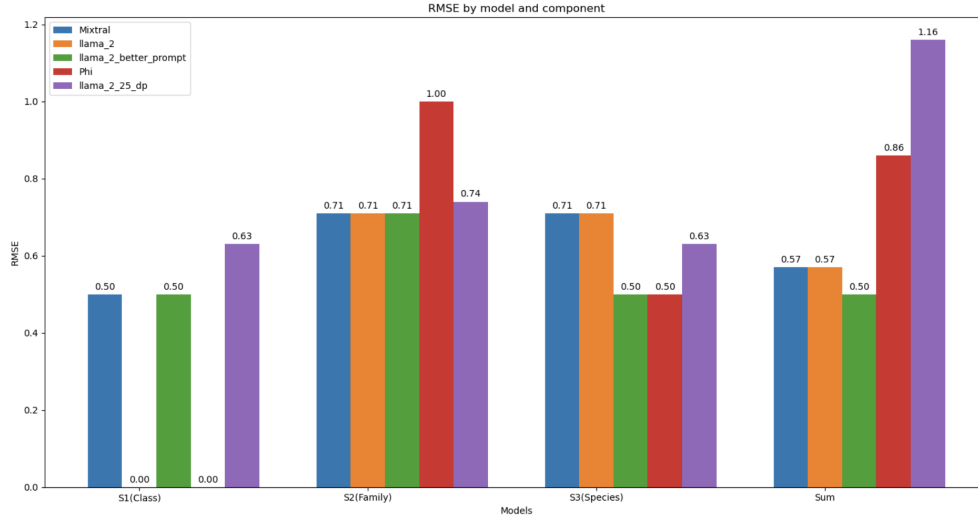
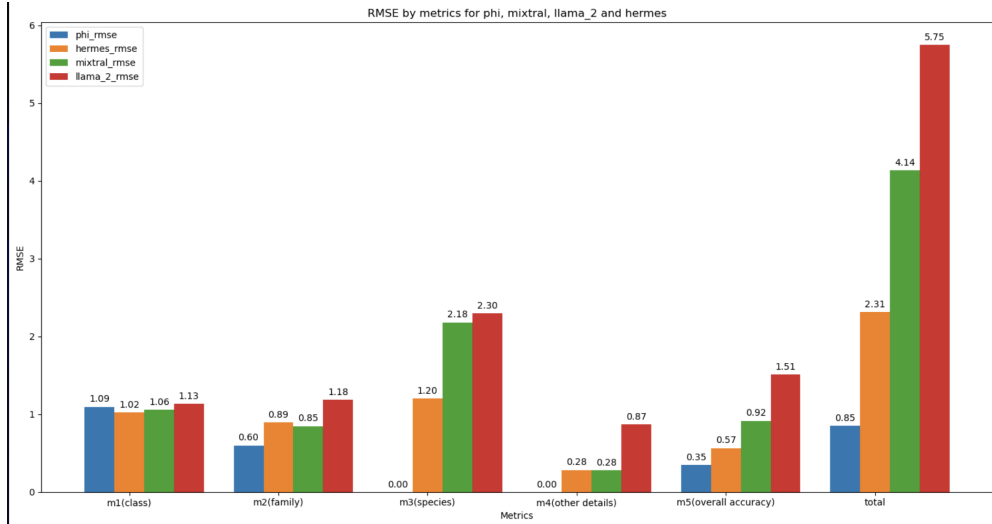


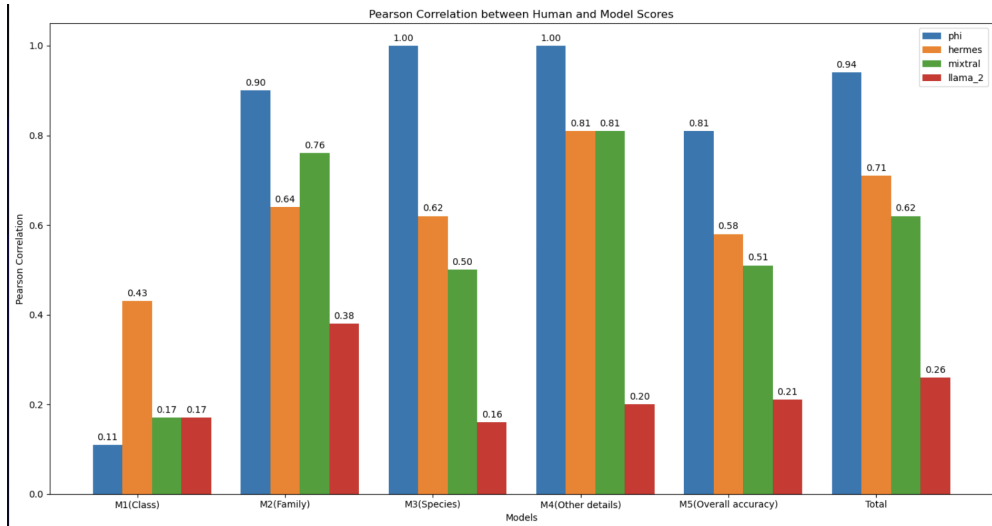
Fig. 6 Evaluation 2 results RMSE analysis

### 6.3 Evaluation 3

Fig 7 shows the Root mean square error values for each model, compared to human evaluator scores. Fig 8 shows the correlation analysis to human evaluator scores. From the RMSE analysis it is clear that phi performs best with error of 0.85 in total score and llama-2 performs worst with error of 5.75 in total score. Fig 9 shows the accuracy achieved by each of the models across each metric (which was calculated by normalising the rmse and then converting it to percentage by multiplying the normalised rmse with 100, and then subtracting this normalised percentage rmse from 100). From the table, it is clear that Phi achieves the highest accuracy across the metrics (except M1) ranging from 100 to 94.3%. And llama-2 achieves lowest accuracy ranging from 92.5 to 61.6%. Refer to Appendix A1 A2 A3 A4 A5 for detailed model scores.



**Fig. 7** Evaluation 3 RMSE scores for 25 data points



**Fig. 8** Evaluation 3 Correlation analysis

## 7 Discussion

### 7.1 How effective is LLM-as-Judge based evaluation ?

The evaluations and result analysis show that language models can perform evaluation close to human standards, achieving accuracy between 61.6% to 100.0%. Also the

Model Name / Metric	M1	M2	M3	M4	M5	Total
Phi	92.7	96.0	100.0	100.0	97.7	94.3
Hermes	93.2	94.0	92.0	98.1	96.2	84.6
Mixtral	92.9	94.3	85.5	98.1	93.9	72.4
Llama-2	92.5	92.1	84.6	98.1	89.9	61.6

**Fig. 9** Evaluation 3 Accuracy analysis

scores have a positive high correlation to human scores, ranging from 0.11 to 1.0. The models used as VLM-Judge are small having parameters not more than 15B and not fine-tuned on the evaluation task. This shows that small non-fine-tuned models can perform well if the prompt and evaluation criteria are good enough.

## 7.2 What are the challenges and Limitations in LLM based evaluation ?

The main challenge is to design best possible prompt and evaluation criteria, fitting the evaluation task as closely as possible. From the evaluations it can be concluded that atleast for small language models prompt engineering, ground truth and fine-grained evaluation criteria play an important role in the accuracy of evaluation. Limitations observed during the evaluations were hallucinations and inconsistencies. Some times the model would hallucinate about the text to evaluate or the ground truth given. Also if everything was kept same, and the evaluations were done again, the models seem to generate different scores for few data points.

## 7.3 What techniques and measures can be used to improve the evaluation ?

During the evaluations it was observed that some techniques yield better results than others. Using few-shot prompts yields better results than zero-shot prompts. Using task specific evaluation metrics covering all the scoring scenarios rather than general evaluation metrics yield better evaluations. Providing detailed and step-wise instructions on how the model should think and evaluate also yields better results than using simple prompts with ambiguous instructions.

## 7.4 Do we really need complex fine-grained evaluation criteria instead of a simple string match scoring ?

To answer this question, I designed a simple scoring scheme, which assigns 0 if no matching words are present in ground truth (only species name) and VLM response, 0.5 if some matching words are present in the response and 1 if all the words match with the ground truth. I used the best performing model from evaluation 3, Phi to run this evaluation on 25 data points. The total sum of scores for all the data points

was 12.5, which converted to percentage equals 50%. To compare this score to the evaluation 3 criteria scores generated by phi, I normalised the total score given by phi for each data point and calculated the sum, which equals to 6.7, which converted into percentage equals 26%. This shows that using a simple scoring criteria, the LLM scores the VLM as 50% accurate, whereas using a complex fine-grained evaluation criteria, the LLM scores the VLM as 20% accurate. My hypothesis is that using a fine-grained evaluation criteria evaluates VLM more strictly resulting in a low accuracy score compared to simple word match evaluation. Hence, we need fine-grained evaluation criteria to evaluate the models more strictly.

## 8 Limitations

### 8.1 Limitation due to size of dataset

The dataset used in this study only includes set of 25 VLM responses and only for one category i.e. Birds. This small size and category can undermine the conclusions made in this project.

### 8.2 Limitation due to Number of human evaluators

This study used only one human evaluator scores as baseline to calculate the accuracy of models. This can lead to bias in the baseline scores.

## 9 Conclusion and Outlook

This study shows that doing custom fine-grained evaluation of VLMs using small language models is a difficult task. But if successful, such evaluation can help in understanding capabilities of VLMs at a fine-grained level. But more work needs to be done, to find if such evaluation can be generalised to other tasks. Following studies can be done in this direction:

- Comparative analysis of custom fine-grained evaluation criteria and simple string match evaluation.
- Comparative analysis of LLM based evaluations for different VLMs.

## Declarations

I certify that the work was solely undertaken by myself without any third person or party helping and it is not AI generated. All sections of the paper that use quotes or describe an argument or concept developed by another author have been referenced, including all secondary literature used, to show that this material has been adopted to support my paper. This is especially true for contents generated by artificial intelligence like GPT-4 or Bard.

Sarang Ravi Chouguley  
19th July 2024

## References

- [1] Alayrac JB, Donahue J, Luc P, et al (2022) Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35:23716–23736
- [2] Chen GH, Chen S, Liu Z, et al (2024) Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:240210669*
- [3] Gao P, Han J, Zhang R, et al (2023) Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:230415010*
- [4] Gunjal A, Yin J, Bas E (2024) Detecting and preventing hallucinations in large vision language models. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp 18135–18143
- [5] Huang H, Qu Y, Liu J, et al (2024) An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:240302839*
- [6] Kim S, Shin J, Cho Y, et al (2023) Prometheus: Inducing fine-grained evaluation capability in language models. In: *The Twelfth International Conference on Learning Representations*
- [7] Kim S, Suk J, Longpre S, et al (2024) Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:240501535*
- [8] Krizhevsky A, Hinton G, et al (2009) Learning multiple layers of features from tiny images
- [9] Lee S, Kim S, Park SH, et al (2024) Prometheusvision: Vision-language model as a judge for fine-grained evaluation. *arXiv preprint arXiv:240106591*
- [10] Li J, Li D, Savarese S, et al (2023) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *International conference on machine learning*, PMLR, pp 19730–19742
- [11] Li Y, Du Y, Zhou K, et al (2023) Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:230510355*
- [12] Liu H, Li C, Wu Q, et al (2024) Visual instruction tuning. *Advances in neural information processing systems* 36
- [13] Nilsback ME, Zisserman A (2008) Automated flower classification over a large number of classes. In: *2008 Sixth Indian conference on computer vision, graphics & image processing*, IEEE, pp 722–729

- [14] Parkhi OM, Vedaldi A, Zisserman A, et al (2012) Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE, pp 3498–3505
- [15] Russakovsky O, Deng J, Su H, et al (2015) Imagenet large scale visual recognition challenge. *International journal of computer vision* 115:211–252
- [16] Thakur AS, Choudhary K, Ramayapally VS, et al (2024) Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:240612624*
- [17] Verga P, Hofstatter S, Althammer S, et al (2024) Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:240418796*
- [18] Wang P, Li L, Chen L, et al (2023) Large language models are not fair evaluators. *arXiv preprint arXiv:230517926*
- [19] Wang X, Zhou Y, Liu X, et al (2024) Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:240110529*
- [20] Xu P, Shao W, Zhang K, et al (2023) Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:230609265*
- [21] Zhang A, Fei H, Yao Y, et al (2024) Vpgtrans: Transfer visual prompt generator across llms. *Advances in Neural Information Processing Systems* 36
- [22] Zheng L, Chiang WL, Sheng Y, et al (2024) Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36
- [23] Zhu D, Chen J, Shen X, et al (2023) Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:230410592*

## Appendix A Prompts, Results and Dataset Details

### A.1 Eval-1 Prompt

''' **Task Description:** An instruction (might include an Input inside it), a response to evaluate, a reference answer that gets a score of 3, and a score rubric representing a evaluation criteria are given. 1. Write a detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general. 2. After writing a feedback, write a score that is an integer between 1 and 3. You should refer to the score rubric. 3. The output format should look as follows: Feedback: (write a feedback for criteria) [RESULT] (an integer number between 1 and 3)4. Please do not generate any other opening, closing, and explanations.

**The instruction to evaluate:** "What kind of animal is that? Be as specific as possible! Fish and bird for example are too coarse."

**Response to evaluate:** {orig\_response}

**Reference Answer (Score 5):** {orig\_reference\_answer}

**Score Rubrics:** *Correctness*

Score 1: The response is completely incorrect and does not give exact name of the animal.

Score 2: The response is partially correct because it does not contain full name of the animal.

Score 3: The response completely correct and contains the exact name of the animal.

**Feedback:** '''

## A.2 Eval-2 Step-1 Prompt

''' Perform the task according to the task description. DO NOT return any extra text or explanation

**Task Description:**

Take the input and extract words which refer to class, family, or species of the animal. Enclose the result between following tags: <RESULT> </RESULT>

**Example 1:**

**Input:**

A peacock is a large, colorful bird known for its distinctive plumage and elaborate tail feathers.

**Output:**

<RESULT>peacock, bird</RESULT>

**Example 2:**

**Input:**

The bald eagle is a bird of prey found in North America. It has a white head and tail with a dark brown body and wings.

**Output:**

<RESULT>Bald eagle, bird</RESULT>

**Example 3:**

**Input:**

a large, long-legged bird, possibly a type of ostrich or emu. These birds are known for their distinctive appearance and long legs, which enable them to run at high speeds.

**Output:**

<RESULT>bird, ostrich, emu</RESULT>

**Your Turn:**

**Input:**

{input}

**Output:** '''

## A.3 Eval-2 Step-2 Prompt

''' Perform the task according to the task description. DO NOT return any extra text or explanation.

**Task Description:**

For the given keywords, classify them into the following groups: class of animal, family of animal, species of animal. Enclose the result between the following tags: <RESULT> </RESULT>

**Example 1:**

**Input:**

Keywords: peacock, bird

**Output:**

<RESULT>

Class: bird

Family: peacock

Species: None

</RESULT>

**Example 2:**

**Input:**

Keywords: Bald eagle, bird

**Output:**

<RESULT>

Class: bird

Family: None

Species: Bald eagle

</RESULT>

**Example 3:**

**Input:**

Keywords: bird, ostrich, emu

**Output:**

<RESULT>

Class: bird

Family: ostrich, emu

Species: None

</RESULT>

**Your Turn:**

**Input:**

{input}

**Output:** '''

## A.4 Eval-2 Step-3 Prompt

''' Perform the task according to the task description. DO NOT return any extra text or explanation.

**Task Description:**

Evaluate the input based on the following score rubric and ground truth. For each criterion, assign the appropriate score and provide a brief justification. Enclose the result between the following tags: <RESULT> </RESULT>

**Score Rubric:**

- **S1: Does the model identify the class of the animal?**
  - 1: for correct



- 0: for none
- -1: for incorrect
- **S2: Does the model identify the family of the animal?**
  - 1: for correct
  - 0: for none
  - -1: for incorrect
- **S3: Does the model identify the species of the animal?**
  - 1: for correct
  - 0: for none
  - -1: for incorrect

**Example 1:**

**Input:**

Class: bird, Family: None, Species: Great Bustard

**Ground Truth:**

Class: Bird, Family: Bustard, Species: Great Bustard

**Output:**

Scores and Justifications:

S1: 1 (Correctly identifies the class as "bird")

S2: 0 (Does not identify the family)

S3: 1 (Correctly identifies the species as "Great Bustard")

<RESULT>S1: 1, S2: 0, S3: 1</RESULT>

**Example 2:**

...

**Example 3:**

...

**Your Turn:**

**Input:**

{input}

**Ground Truth:**

{ground\_truth}

Scores and Justifications:

S1:

S2:

S3: '''

## A.5 Eval-3 Prompt

''' **Evaluation Task**

You are an expert judge evaluating the Retrieval Augmented Generation applications. Your task is to evaluate a given answer based on a ground truth and question using the criteria provided below.

**Evaluation Criteria (0-1):**

{evaluation\_criteria}

**Evaluation Steps:**  
{evaluation\_steps}

**Output format:**  
{json\_schema}

**Examples:**  
{examples}

Now, please evaluate the following:

**Question:**  
{question}  
**Ground Truth:**  
{ground\_truth}  
**Answer:**  
{answer} '''

Sr. No.	Model Name	Model provided	Parameter size	Quantisation
1	Prometheus-v1	Prometheus-eval	13B	None
2	Prometheus-v2	Prometheus-eval	7B	None
3	Mixtral-8x7B-v0.1	MistralAI	7B	None
4	Mistral-7B-Instruct-v0.3	MistralAI	7B	None
5	Meta-Llama-3-8B	Meta	8B	None
6	Phi-3-medium-128k-Instruct	Microsoft	14B	None
7	Llama-2-13B-chat	TheBloke	13B	GPTQ
8	Hermes-2-Theta-Llama-3-8B	NousResearch	8B	None

**Fig. A1** Model Details

id	M1	M2	M3	M4	M5	Total
DSI1	1	0	0	0	0	1
DSI2	2	0	0	0	1	3
DSI3	2	0	2	1	1	6
DSI4	2	0	0	1	1	4
DSI5	2	1	0	0	1	4
DSI6	2	2	0	0	2	6
DSI7	2	1	0	0	1	4
DSI8	2	0	0	0	1	3
DSI9	2	2	0	0	2	6
DSI10	1	0	0	0	0	1
DSI11	2	0	0	0	1	3
DSI12	2	0	0	0	1	3
DSI13	2	1	0	0	1	4
DSI14	2	0	0	0	1	3
DSI15	2	0	0	0	1	3
DSI16	2	1	0	0	1	4
DSI17	2	0	0	0	1	3
DSI18	2	1	0	0	1	4
DSI19	2	0	0	0	1	3
DSI20	2	2	5	0	3	12
DSI21	2	0	0	0	1	3
DSI22	2	1	1	2	2	8
DSI23	2	0	0	0	1	3
DSI24	2	0	0	0	1	3
DSI25	2	1	0	0	1	4

**Table A1** Eval-3 Phi Evaluation Scores

id	M1	M2	M3	M4	M5	Total
DSI1	2	0	0	0	1	3
DSI2	2	0	0	0	1	3
DSI3	2	1	1	0	1	5
DSI4	2	0	0	0	1	3
DSI5	2	2	0	0	2	6
DSI6	1	2	3	0	2	8
DSI7	2	3	4	0	2	11
DSI8	2	0	0	0	1	3
DSI9	2	0	0	0	1	3
DSI10	1	0	0	0	1	2
DSI11	2	0	0	0	1	3
DSI12	2	0	0	0	1	3
DSI13	2	0	0	0	1	3
DSI14	2	0	0	0	1	3
DSI15	2	0	0	0	1	3
DSI16	2	0	0	0	1	3
DSI17	2	0	0	0	1	3
DSI18	2	1	0	0	2	5
DSI19	2	0	0	0	1	3
DSI20	2	3	5	0	3	13
DSI21	2	0	0	0	0	2
DSI22	2	2	2	2	2	10
DSI23	2	0	0	0	1	3
DSI24	2	0	0	0	0	2
DSI25	2	2	3	0	2	9

**Table A2** Eval-3 Hermes Evaluation Scores

id	M1	M2	M3	M4	M5	Total
DSI1	1	0	0	0	0	1
DSI2	2	0	0	0	1	3
DSI3	2	3	5	0	2	12
DSI4	2	1	0	0	1	4
DSI5	2	2	3	0	2	9
DSI6	2	3	5	0	3	13
DSI7	2	3	5	0	3	13
DSI8	2	0	0	0	0	2
DSI9	2	2	0	0	1	5
DSI10	1	0	0	0	0	1
DSI11	2	0	0	0	0	2
DSI12	2	0	0	0	1	3
DSI13	2	2	0	0	2	6
DSI14	2	0	0	0	1	3
DSI15	2	0	0	0	1	3
DSI16	2	2	3	0	3	10
DSI17	2	0	0	0	0	2
DSI18	2	2	5	0	2	11
DSI19	2	0	0	0	1	3
DSI20	2	3	5	0	3	13
DSI21	1	0	0	0	1	2
DSI22	2	3	5	2	3	15
DSI23	1	0	1	0	1	3
DSI24	2	0	0	0	1	3
DSI25	2	0	0	0	1	3

**Table A3** Eval-3 Mixtral Evaluation Scores

id	M1	M2	M3	M4	M5	Total
DSI1	0	0	0	0	0	0
DSI2	2	0	0	0	1	3
DSI3	2	2	3	2	3	12
DSI4	2	0	0	0	1	3
DSI5	2	2	3	1	3	11
DSI6	2	2	3	1	3	11
DSI7	2	2	3	0	3	10
DSI8	2	2	3	0	3	10
DSI9	2	0	0	0	1	3
DSI10	0	0	0	0	0	0
DSI11	2	0	0	0	1	3
DSI12	2	2	4	2	4	14
DSI13	2	2	3	2	3	12
DSI14	2	2	5	1	3	13
DSI15	2	2	3	0	3	10
DSI16	2	2	2	1	2	9
DSI17	2	2	3	1	3	11
DSI18	2	2	3	1	3	11
DSI19	2	2	3	0	2	9
DSI20	2	2	3	1	3	11
DSI21	0	0	0	0	0	0
DSI22	2	1	0	1	1	5
DSI23	0	0	0	0	0	0
DSI24	2	1	0	1	2	6
DSI25	2	1	0	0	2	5

**Table A4** Eval-3 Llama-2 Scores

id	M1	M2	M3	M4	M5	Total
DSI1	2	0	0	0	1	3
DSI2	2	0	0	0	1	3
DSI3	2	0	2	1	1	6
DSI4	2	0	0	1	1	4
DSI5	0	3	0	0	1	4
DSI6	0	3	0	0	1	4
DSI7	0	2	0	0	1	3
DSI8	2	0	0	0	1	3
DSI9	2	2	0	0	1	5
DSI10	0	0	0	0	0	0
DSI11	2	0	0	0	1	3
DSI12	2	0	0	0	1	3
DSI13	0	2	0	0	1	3
DSI14	2	0	0	0	1	3
DSI15	2	0	0	0	1	3
DSI16	0	2	0	0	1	3
DSI17	2	0	0	0	1	3
DSI18	0	2	0	0	1	3
DSI19	2	0	0	0	1	3
DSI20	2	2	5	0	3	12
DSI21	2	0	0	0	1	3
DSI22	2	1	1	2	2	8
DSI23	0	0	0	0	1	1
DSI24	2	0	0	0	1	3
DSI25	2	1	0	0	1	4

**Table A5** Eval-3 Human Scores

Datapoint	Llama-2	Human Baseline
DSI1	S1: 1, S2: 0, S3: 0	S1: 1, S2: -1, S3: 0
DSI2	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 0
DSI3	S1: 1, S2: 0, S3: 1	S1: 1, S2: -1, S3: 0
DSI4	S1: 1, S2: 0, S3: 0	S1: 1, S2: -1, S3: 0
DSI5	S1: 1, S2: 0, S3: 1	S1: 0, S2: 1, S3: 0
DSI6	S1: 1, S2: 0, S3: 0	S1: 0, S2: 1, S3: 0
DSI7	S1: 1, S2: 1, S3: 1	S1: 0, S2: 1, S3: 0
DSI8	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 0
DSI9	S1: 1, S2: 0, S3: 0	S1: 0, S2: 1, S3: 0
DSI10	S1: 1, S2: 0, S3: 0	S1: 0, S2: -1, S3: 0
DSI11	S1: 1, S2: 0, S3: 1	S1: 1, S2: 0, S3: 0
DSI12	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 0
DSI13	S1: 1, S2: 0, S3: 0	S1: 0, S2: 1, S3: 0
DSI14	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 0
DSI15	S1: 1, S2: 0, S3: 1	S1: 0, S2: 1, S3: 0
DSI16	S1: 1, S2: 0, S3: 1	S1: 0, S2: 1, S3: 0
DSI17	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 0
DSI18	S1: 1, S2: 0, S3: 1	S1: 0, S2: 1, S3: 0
DSI19	S1: 1, S2: 0, S3: 1	S1: 1, S2: 0, S3: 0
DSI20	S1: 1, S2: 0, S3: 1	S1: 1, S2: 0, S3: 1
DSI21	S1: 1, S2: 0, S3: 0	S1: 1, S2: -1, S3: 0
DSI22	S1: 1, S2: 0, S3: 0	S1: 1, S2: 1, S3: 1
DSI23	S1: 1, S2: 0, S3: 1	S1: 0, S2: -1, S3: 0
DSI24	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 0
DSI25	S1: 1, S2: 0, S3: 0	S1: 1, S2: 0, S3: 0

**Table A6** Eval-2 Llama-2 and Human Scores



id	Response	Ground Truth
aenb1	a peacock, which is a large, colorful bird known for its distinctive plumage and elaborate tail feathers.	Great Bustard
aenb2	Bird	Bald Eagle
aenb3	a bird, specifically a parrot.	Bird of Paradise
aenb4	a large bird, specifically a crane.	Marabou Stork
aenb5	Parrot	African Grey Parrot

**Table A7** Dataset for eval-1

id	Response	Ground Truth
DSI1	a peacock, which is a large, colorful bird known for its distinctive plumage and elaborate tail feathers.	Class: Bird, Family: Bustard, Species: Great Bustard
DSI2	Bird	Class: Bird, Family: Eagle, Species: Bald Eagle
DSI3	a bird, specifically a parrot.	Class: Bird, Family: Birds of Paradise, Species: Lesser Birds of paradise
DSI4	a large bird, specifically a crane.	Class: Bird, Family: Stork, Species: Marabou Stork
DSI5	Parrot	Class: Bird, Family: Parrot, Species: African Grey Parrot
DSI6	Flamingo	Class: Bird, Family: Flamingo, Species: American Flamingo
DSI7	Ostrich	Class: Bird, Family: Ostrich, Species: African Ostrich
DSI8	Bird	Class: Bird, Family: Crane, Species: Grey Crowned Crane
DSI9	Parrot	Class: Bird, Family: Parrot, Species: Macaw
DSI10	Seagull	Class: Bird, Family: Albatrosses, Species: Great Albatross
DSI11	Bird	Class: Bird, Family: Parrot, Species: Hyacinth Macaw
DSI12	Bird	Class: Bird, Family: Waxbills, Species: Gouldian Finch
DSI13	Penguin	Class: Bird, Family: Penguins, Species: Humboldt Penguin
DSI14	Bird	Class: Bird, Family: Cuckoos, Species: Roadrunner
DSI15	Bird	Class: Bird, Family: Parrot, Species: Ring-necked Parakeet
DSI16	Hummingbird	Class: Bird, Family: Hummingbird, Species: Blue-chinned sapphire
DSI17	Bird	Class: Bird, Family: Cockatoos, Species: Inca Cockatoo
DSI18	Peacock	Class: Bird, Family: Peafowl, Species: Indian Peafowl
DSI19	Bird	Class: Bird, Family: Toucan, Species: Toco Toucan
DSI20	a bird, specifically a blue-footed booby.	Class: Bird, Family: Booby, Species: Blue-footed Booby
DSI21	a bird, specifically a parrot.	Class: Bird, Family: Frigatebird, Species: Magnificent frigatebird
DSI22	a large, long-legged bird, possibly a type of ostrich or emu. These birds are known for their distinctive appearance and long legs, which enable them to run at high speeds.	Class: Bird, Family: Emu, Species: Emu
DSI23	Parrot	Class: Bird, Family: Cockatoo, Species: Cockatiel
DSI24	Bird	Class: Bird, Family: Eagle, Species: Harpy Eagle
DSI25	a large bird, specifically a vulture.	Class: Bird, Family: New World Vultures, Species: Condor

**Table A8** Dataset for Eval-2

id	Response	Ground Truth
DSI1	The bird is a seagull.	Class: Bird, Family: Sulidae, Species: Northern Gannet
DSI2	Bird	Class: Bird, Family: Eagle, Species: Bald Eagle
DSI3	a pelican, which is a large white bird with a long beak.	Class: Bird, Family: Pelecanidae, Species: Pink Pelican
DSI4	a large bird, specifically a crane.	Class: Bird, Family: Stork, Species: Marabou Stork
DSI5	Parrot	Class: Bird, Family: Parrot, Species: African Grey Parrot
DSI6	Flamingo	Class: Bird, Family: Flamingo, Species: American Flamingo
DSI7	Ostrich	Class: Bird, Family: Ostrich, Species: African Ostrich
DSI8	Bird	Class: Bird, Family: Crane, Species: Grey Crowned Crane
DSI9	Parrot	Class: Bird, Family: Parrot, Species: Macaw
DSI10	Seagull	Class: Bird, Family: Albatrosses, Species: Great Albatross
DSI11	Bird	Class: Bird, Family: Parrot, Species: Hyacinth Macaw
DSI12	Bird	Class: Bird, Family: Waxbills, Species: Gouldian Finch
DSI13	Penguin	Class: Bird, Family: Penguins, Species: Humboldt Penguin
DSI14	Bird	Class: Bird, Family: Cuckoos, Species: Roadrunner
DSI15	Bird	Class: Bird, Family: Parrot, Species: Ring-necked Parakeet
DSI16	Hummingbird	Class: Bird, Family: Hummingbird, Species: Blue-chinned sapphire
DSI17	Bird	Class: Bird, Family: Cockatoos, Species: Inca Cockatoo
DSI18	Peacock	Class: Bird, Family: Peafowl, Species: Indian Peafowl
DSI19	Bird	Class: Bird, Family: Toucan, Species: Toco Toucan
DSI20	a bird, specifically a blue-footed booby.	Class: Bird, Family: Booby, Species: Blue-footed Booby
DSI21	a bird, specifically a parrot.	Class: Bird, Family: Frigatebird, Species: Magnificent frigatebird
DSI22	a large, long-legged bird, possibly a type of ostrich or emu. These birds are known for their distinctive appearance and long legs, which enable them to run at high speeds.	Class: Bird, Family: Emu, Species: Emu
DSI23	Parrot	Class: Bird, Family: Cockatoo, Species: Cockatiel
DSI24	Bird	Class: Bird, Family: Eagle, Species: Harpy Eagle
DSI25	a large bird, specifically a vulture.	Class: Bird, Family: New World Vultures, Species: Condor

**Table A9** Dataset for Eval-3