

Taxonomic Reasoning for Rare Arthropods: Combining Dense Image Captioning and RAG for Interpretable Classification

Nathaniel Lesperance^{†, ‡}, Sujeewan Ratnasingham[◊], Graham W. Taylor^{†, ‡, ◊, *}

[†] University of Guelph, Guelph, ON

[‡] Vector Institute for AI, Toronto, ON

[◊] Centre for Biodiversity Genomics, Guelph, ON

Abstract

In the context of pressing climate change challenges and the significant biodiversity loss among arthropods, automated taxonomic classification from organismal images is a subject of intense research. However, traditional AI pipelines based on deep neural visual architectures such as CNNs or ViTs face limitations such as degraded performance on the long-tail of classes and the inability to reason about their predictions. We integrate image captioning and retrieval-augmented generation (RAG) with large language models (LLMs) to enhance biodiversity monitoring, showing particular promise for characterizing rare and unknown arthropod species. While a naïve Vision-Language Model (VLM) excels in classifying images of common species, the RAG model enables classification of rarer taxa by matching explicit textual descriptions of taxonomic features to contextual biodiversity text data from external sources. The RAG model shows promise in reducing overconfidence and enhancing accuracy relative to naïve LLMs, suggesting its viability in capturing the nuances of taxonomic hierarchy, particularly at the challenging family and genus levels. Our findings highlight the potential for modern vision-language AI pipelines to support biodiversity conservation initiatives, emphasizing the role of comprehensive data curation and collaboration with citizen science platforms to improve species identification, unknown species characterization and ultimately inform conservation strategies.

Keywords: Taxonomy, Artificial Intelligence, RAG, LLM, Biodiversity

1. Introduction

The ongoing global climate change polycrisis contributes significantly to the loss of biodiversity worldwide, particularly in Arthropoda [1]. Arthropods are found in every habitat on Earth, playing a myriad of ecological roles [2]. They contribute significantly to the health of a given ecosystem [3] and they are the most diverse phylum, making up approximately 84% of known animal species [4]. Despite their diversity, many more species are unknown than known to science [5] [6], making the monitoring of species diversity, distributions, and phenology challenging. Arthropods' overrepresentation among invasive species further threatens biodiversity conservation. The precise taxonomic classification of known and unknown species of arthropods is fundamental for biodiversity conservation and to limiting the negative impacts of climate change for all life on Earth.

Taxonomic classification using DNA-based methods is the primary driver to recent breakthroughs in biodiversity understanding, initiated by the characterization of DNA barcodes and the adoption of species-like clusters called Barcode Index Numbers (BINs) that dramatically expanded our biodiversity catalogue [7, 8]. Recent improvements in deep learning have provided new opportunities for automated DNA-based classification at scale such as the Nucleotide Transformer [9] and HyenaDNA [10]. Unimodal automated taxonomic systems optimized for and trained on DNA barcodes have emerged, including BarcodeBERT [11], MycoAI [12] and BarcodeMAMBA [13]. Jain *et al.* [14] and Bjerger *et al.* [15] trained

* gwtaylor@uoguelph.ca

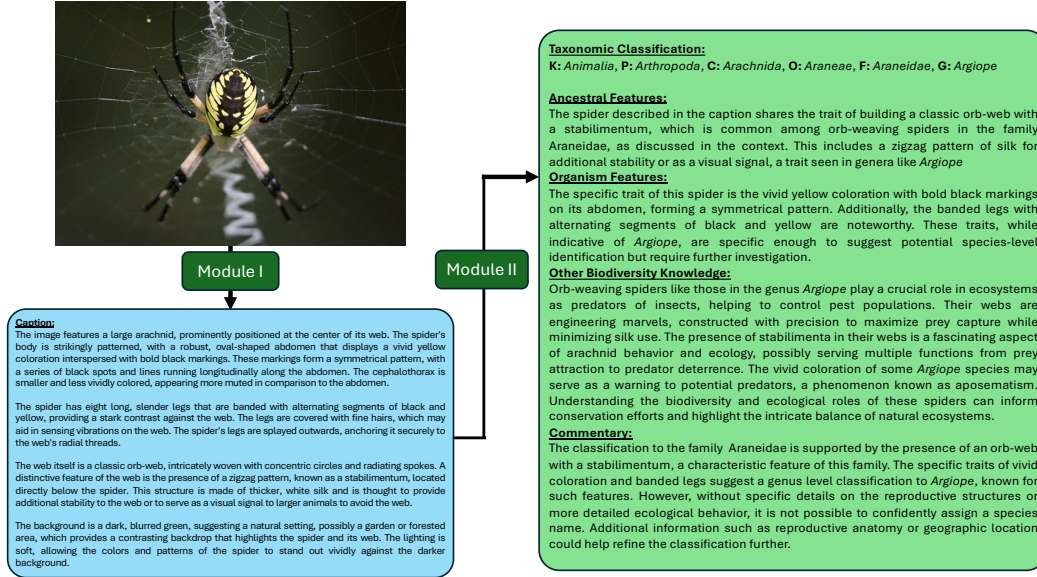


Figure 1. Example output from Module I (light blue) and Module II (light green) for an image of a yellow garden spider (Photo attribution: David Illig). In taxonomic classification: K=Kingdom, P=Phylum, C=Class, O=Order, F=Family, G=Genus. The RAG model generated accurate taxonomic labels to the genus level and refrained from providing a species level classification.

unimodal automated taxonomic systems using images from camera traps. In multimodal approaches, CLIBD [16] extends BioCLIP’s [17] contrastive learning of text and image embeddings to include DNA, addressing the difficulty of obtaining comprehensive expert taxonomic labels at fine scale. These two models and others [18] also mention the characterization of unknown species as a motivation for their architecture and learning strategies. These models have facilitated biodiversity monitoring at scale and reduce the manual labour of domain experts. Yet, to ecologists, they are effectively black boxes.

Prior to DNA’s widespread adoption, taxonomists relied on detailed textual descriptions and references to classify specimens and identify new species. In contrast, today’s automated taxonomic systems typically process raw DNA sequences or images in isolation, without incorporating this rich contextual knowledge. Citizen science resources like iNaturalist [19], Wikipedia and Wikispecies are well maintained and continually updated, allowing them to play a key role in data collection and curation for biodiversity research. While DNA-based methods like Barcode Index Numbers (BINs) have expanded our biodiversity catalogue, newly discovered species often lack the dense trait descriptions that traditionally accompany taxonomic classification. The combination of untapped text-based biodiversity resources and the success of Large Language Models (LLMs) at text-based reasoning enables automated taxonomic classification using LLMs that leverage visual features.

Retrieval Augmented Generation (RAG), which can provide additional context to LLMs at inference has been shown to help in overcoming knowledge limitations in training data, obfuscation of reasoning processes and hallucinations seen in LLM responses [20–22]. In the biological domain, RAG has been deployed in exploration of protein-protein interactions [23] but there are no known applications of RAG in taxonomic classification. Additionally, multimodal LLMs show promise in augmenting human biological research [24] and have been used to generate text captions of scientific figures [25]. Images of living arthropods can provide additional context clues that assist in taxonomic classification like habitat,

interactions and behaviour. These supplementary context clues are present in taxonomy literature for retrieval and mostly missing from specimen images common to large arthropod image datasets like TOL-10M [17] and BIOSCAN-5M [26]. The following exploration pairs multimodal LLMs as an image captioning tool with RAG systems to reason over curated biodiversity databases with the aim to generate taxonomic classifications and accelerated biodiversity knowledge for images of arthropods (Figure 1).

Automated taxonomic classification of organismal images using captioning paired with RAG against a vector database built from rich text-based knowledge bases enables minimally invasive and scalable zero-shot classification at fine-grained taxonomic levels. Our approach brings a generalization benefit to moving beyond traditional fixed-class datasets, especially when many classes remain unknown, as is common in biodiversity research [27]. Combining computer vision, natural language processing, and biodiversity informatics can address taxonomic classification challenges, strengthen citizen science platforms, and support global biodiversity efforts like BIOSCAN [28] and Biodiversity Genomics Europe [29].

2. Methods

Figure 2 summarizes our pipeline, detailing data source preprocessing, Module I - Image Captioning, and Module II - RAG System, used for generating taxonomic classifications and supplemental knowledge from organismal images. We now discuss each step in detail.

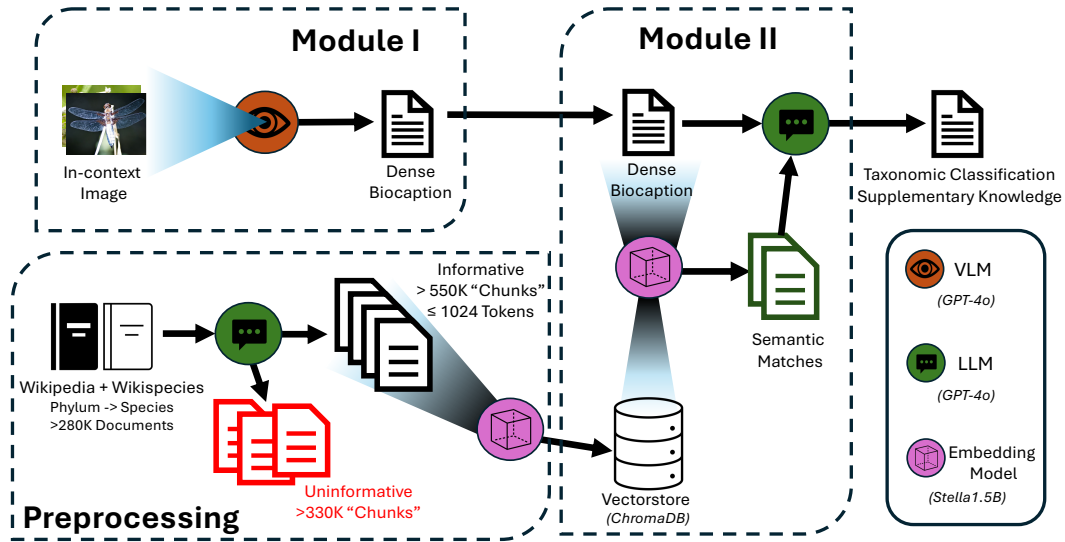


Figure 2. Preprocessing of Wikipedia and Wikispecies text data includes chunking, filtering and contextualization before generating a vector database of embeddings. In-context images of Arthropoda are inputs to the RAG model. They are initially fed through a VLM (Module I) to generate dense image biocaptions which are the subject of RAG queries to the vector database in Module II. In Module II, the Simple RAG and Advanced RAG models differ by search criterion and additional LLM passes for reranking and multi-query; see subsection 2.3 for more detail. Both RAG models generate taxonomic classifications, accelerated biodiversity knowledge and commentary on LLM confidence for each image.

2.1. Data Collection and Preprocessing

Work with VLMs and LLMs is carried out using LangChain [30], instructor [31], and custom prompts (Appendix B, Appendix C, Appendix D) that also contain thinking dots [32].

2.1.1. Image Datasets

Living Arthropods: We curated a development dataset of 240 arthropod images photographed primarily in their natural habitats from Wikimedia and Ontario Nature Magazine. We used this dataset to validate our approach and optimize hyperparameters while managing computational costs. Every image has comprehensive taxonomic labels spanning 5 classes, 18 orders, 65 families, 120 genera, and 153 species within the Phylum Arthropoda with significant Class Insecta representation reflecting typical phylum distributions.

Rare Species subset: We selected all 951 Arthropoda images from the Rare Species dataset [17], including both in-situ habitat photographs and laboratory specimen images. These species are classified by the IUCN Red List [33] as Near Threatened, Vulnerable, Endangered, Critically Endangered, or Extinct in the Wild. Each image has complete taxonomic labels across 4 classes, 9 orders, 17 families, 26 genera, and 32 species within the Phylum Arthropoda, again with Class Insecta being most represented. We provide detailed taxonomic counts for both datasets in Appendix A.

2.1.2. Text-based Knowledge Sources

We tokenized all text from Wikipedia and Wikispecies articles within the Kingdom Animalia, (280,120 documents) using Byte-Pair Encoding and recursively split it into chunks using whitespace as delimiters, generating 879,611 total starting chunks. We then performed semantic filtering of uninformative chunks, reducing the total to 550,711 chunks. For each remaining chunk, we used custom LLM prompts (Appendix B) to generate contextualizing text [34], processing each chunk along with its full corresponding source document where all chunks post-processing were ≤ 1024 tokens.

2.1.3. RAG Vectorstore

We generated embeddings with ChromaDB [35] using the `stella_en_1.5B_v5` model [36]. We normalized all embedding vectors to length unity and tagged each chunk with metadata containing taxonomic rank, data source (Wikipedia or Wikispecies) and taxonomic name.

2.2. Module I - Image Captioning

Traditional captioners are trained to generate short descriptions for generic scenes and lack the detail required to produce detailed biologically-informed captions (biocaptions). Image-caption datasets are also generic and there are no known datasets geared towards this producing biocaptions. To pair well with the RAG task, biocaptions should exclusively describe visible features without inappropriate inference that could cloud taxonomic classification. Instead, each image in the dataset is passed to OpenAI’s multimodal GPT-4o model [37] with custom prompting for a dense descriptive caption that could be informative for the taxonomic classification task (Appendix C), including all visible features of the primary organism and its surroundings. Negative prompting helps prevent unnecessary inference about taxonomic classification, unseen behaviour or non-visible body parts.

2.3. Module II - RAG System

Query construction: We used the dense descriptive captions from Module I to query the vector database. Our querying process begins by generating a normalized semantic embedding query vector using the same `stella_en_1.5B_v5` embedding model.

Retrieval: We built two retrieval pipelines to test the impact of advanced retrieval methods on model performance. In the ‘Simple RAG’ model, we use dense search, maximizing cosine similarity to retrieve chunks that semantically match the query. We use default search parameters except for *top k* (30 chunks per query). In the ‘Advanced RAG’ model, we use multiquery retrieval with an instance of GPT-4o-mini to generate multiple artificial semantically similar queries. We use dense search selecting by Maximal Marginal Relevance (MMR) to obtain diverse chunks matching these queries. The set of unique chunks retrieved is passed to Cohere’s reranking model `rerank-english-v3.0` [38] and the top 10 chunks are provided as context for response generation.

Response generation: We passed the dense descriptive caption of the image, along with the text and metadata from the retrieved chunks, to another instance of GPT-4o with a custom prompt (Appendix D). This prompt instructed the model to generate taxonomic classifications at each rank, but only when confident in its classifications. We instructed the model to refrain from making classifications at or below any rank where the evidence in the caption and context was insufficient. In the same pass, we also prompted the model to generate biodiversity knowledge for the organism based on the caption, context and classification, along with commentary on its confidence and choices.

2.3.1. Naïve VLMs and Naïve LLM

To evaluate the contribution of each component in our pipeline, we implemented two baseline variants. Another instance of the same GPT-4o model used in Module I (known here as ‘Naïve VLM’) generates responses from images alone without Module I’s captioning or Module II’s retrieval of additional context from the database. Another instance of the same GPT-4o model used in Module II (known here as ‘Naïve LLM’) generates responses directly from dense captions alone without retrieval of additional context from the database. In evaluations on the Rare Species subset, we test the ‘Naïve VLM’ and an instance of Google DeepMind’s Gemini 2.0 Flash [39] (known here as ‘Naïve Gemini Flash 2.0’). As a result, ‘Naïve VLM’ is replaced with ‘Naïve GPT-4o’ to disambiguate.

2.4. Evaluation

2.4.1. Classification

To evaluate our hypothesis that RAG-based classification would perform better on rare taxa by leveraging external knowledge, we organized the Living Arthropod samples into rare taxa and common taxa based on the cumulative distribution of number of species observations (n_{obs}) in the iNaturalist database [40]. **Rare taxa** ($n = 109$) with n_{obs} above the 95th percentile of the cumulative distribution ($n_{\text{obs}} < 6699$) and **common taxa** ($n = 24$) with n_{obs} below the median of the cumulative distribution, ($n_{\text{obs}} > 60,562$). We then evaluated classification macro accuracy of predicted labels at each taxonomic level for each of the Naïve VLM, Naïve LLM and Simple RAG Model. Since models were instructed to refrain from classification when uncertain, we also tracked the number of classification attempts made by each model — this helps distinguish models that achieve high accuracy by being selective from those that maintain accuracy while attempting more classifications.

2.4.2. RAG

Retrieval augmented generation of accelerated biodiversity knowledge was evaluated using two custom measures: answer relevancy and faithfulness, provided by RAGAS [41] that use instances of GPT-4o-mini. We compared the Simple RAG model with the Advanced RAG model (both described in subsection 2.3).

To assess **answer relevancy**, an LLM generates artificial questions based on the model’s final output, then takes the mean of cosine similarities between embeddings of each artificial question and the model input query. To assess **faithfulness**, an LLM determines the total number of factual claims in the response and the number of context-supported claims. The ratio of context-supported claims to total claims is the faithfulness score. Both of these measures generate a score between 0 and 1 from each response.

3. Results

Example output from Module I and the Module II (Simple RAG) is shown in Figure 1.

3.1. Validation: Image Captioning

Module I was able to sufficiently describe visible features of arthropods including informed morphology, behaviour and contextualization. However, despite negative prompting, not all speculation or inferences about function not immediately evident in the image were eliminated during captioning. For example, in an example caption (Figure 1) the description “...and is thought to provide additional stability to the web or to serve as a visual signal to larger animals to avoid the web” is inappropriate speculation about the function that should be left to Module II during context comparison.

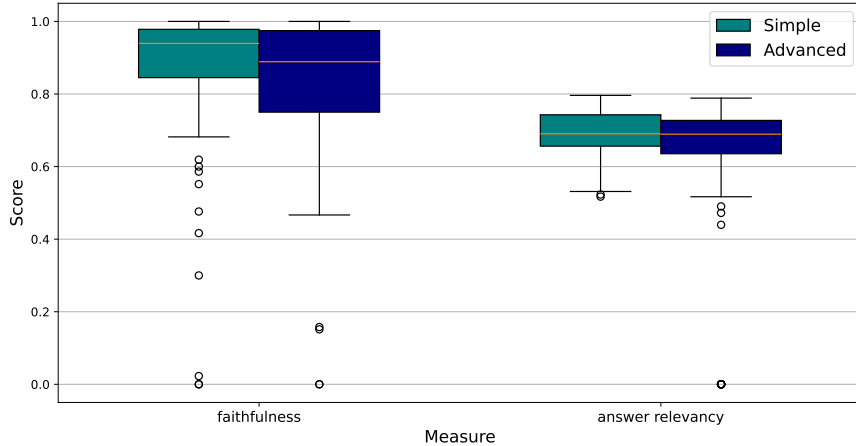


Figure 3. Faithfulness and answer relevancy score distributions for accelerated biodiversity knowledge for Simple and Advanced RAG models. Medians (red lines), Inter quartile ranges (colour-filled boxes), maxima, minima as well as outliers more than 1.5 IQR outside the 1st and 3rd quartiles are shown. The Advanced RAG model uses MMR search criterion, multi-query and reranking while the Simple RAG model uses cosine similarity search criterion, no multiquery and no reranking.

3.2. Validation: Retrieval Augmented Generation

Figure 3 shows distributions of the two key RAG metrics: answer relevancy and faithfulness, for the Simple RAG and Advanced RAG models. Both methods maintain high median faithfulness and answer relevancy. Multiple ablations of advanced RAG methods like reranking, alternative search criterion, and multi-query as well as variation in hyperparameters showed minimal impact on RAG performance (Figure 3) indicating a simpler RAG system was sufficient to maximize context retrieval of data available in the vector database.

High median faithfulness and median answer relevancy (Figure 3) with minimal outliers indicates the generated biodiversity knowledge aligns with the assigned task and factually aligns with the retrieved context. This supports effective retrieval of context as well as effective text generation based on context.

Table 1. Model classification attempts and classification macro accuracy (Acc.) for the Simple RAG model, Naïve LLM and Naïve VLM. Best performing models at each rank are in **bold** unless all models performed equally. Second best model is underlined.

(a) All 240 arthropod images.

Model	Phylum		Class		Order	
	Attempts (%)	Acc.	Attempts (%)	Acc.	Attempts (%)	Acc.
Simple RAG	240 (100%)	1.00	240 (100%)	0.971	217 (90.4%)	<u>0.977</u>
Naïve LLM	240 (100%)	1.00	240 (100%)	0.971	225 (93.8%)	0.942
Naïve VLM	240 (100%)	1.00	240 (100%)	0.971	240 (100%)	0.983

Model	Family		Genus		Species	
	Attempts (%)	Acc.	Attempts (%)	Acc.	Attempts (%)	Acc.
Simple RAG	35 (14.6%)	<u>0.657</u>	3 (1.3%)	1.00	0 (0%)	–
Naïve LLM	61 (25.4%)	<u>0.541</u>	11 (4.6%)	0.364	0 (0%)	–
Naïve VLM	240 (100%)	0.746	240 (100%)	<u>0.442</u>	230 (95.8%)	0.248

(b) Common taxa ($n = 24$, $n_{\text{obs}} > 60562$).

Model	Phylum		Class		Order	
	Attempts (%)	Acc.	Attempts (%)	Acc.	Attempts (%)	Acc.
Simple RAG	24 (100%)	1.00	24 (100%)	1.00	24 (100%)	1.00
Naïve LLM	24 (100%)	1.00	24 (100%)	1.00	24 (100%)	1.00
Naïve VLM	24 (100%)	1.00	24 (100%)	1.00	24 (100%)	1.00

Model	Family		Genus		Species	
	Attempts (%)	Acc.	Attempts (%)	Acc.	Attempts (%)	Acc.
Simple RAG	5 (20.8%)	0.600	1 (4.2%)	1.00	0 (0%)	–
Naïve LLM	12 (50.0%)	<u>0.750</u>	3 (12.5%)	0.667	0 (0%)	–
Naïve VLM	24 (100%)	0.958	24 (100%)	<u>0.729</u>	24 (100%)	0.792

(c) Rare taxa ($n = 109$, $n_{\text{obs}} > 6699$).

Model	Phylum		Class		Order	
	Attempts (%)	Acc.	Attempts (%)	Acc.	Attempts (%)	Acc.
Simple RAG	109 (100%)	1.00	109 (100%)	0.982	99 (90.8%)	<u>0.970</u>
Naïve LLM	109 (100%)	1.00	109 (100%)	0.982	100 (91.7%)	0.930
Naïve VLM	109 (100%)	1.00	109 (100%)	0.982	109 (100%)	0.982

Model	Family		Genus		Species	
	Attempts (%)	Acc.	Attempts (%)	Acc.	Attempts (%)	Acc.
Simple RAG	18 (16.5%)	<u>0.556</u>	2 (1.8%)	1.00	0 (0%)	–
Naïve LLM	24 (22.0%)	0.458	3 (2.8%)	0	0 (0%)	–
Naïve VLM	109 (100%)	0.661	109 (100%)	<u>0.275</u>	109 (100%)	0.028

3.3. Validation: Taxonomic Classification

Table 1(a) shows classification accuracy as well as number of taxonomic classification attempts made by the Naïve VLM, Naïve LLM, and RAG models at each taxonomic rank in the Living Arthropod dataset. The RAG model and Naïve VLM perform similarly at the phylum, class, and order levels, both achieving higher accuracy and lower overconfidence than the Naïve LLM. At the family level, Naïve VLM outperforms the RAG model with both exceeding the Naïve LLM in accuracy. The RAG model obtains perfect performance on its three genus-level attempts but is likely too conservative. At the species level, only the Naïve VLM makes classification attempts and shows a drop in accuracy despite little change to its confidence. For rare taxa (Table 1(c)), however, the RAG model maintained its performance better than both Naïve LLM and Naïve VLM, which showed significant accuracy drops compared to their performance on common species (Table 1(b)).

Table 2. Model classification attempts, classification macro accuracy (Acc.) and F1 scores for the Simple RAG model, Advanced RAG model, Naïve GPT-4o and Naïve Gemini 2.0 Flash on 951 Rare Species arthropod images. Best performing models at each rank are in **bold** unless all models performed equally. Second best model is underlined. We count accuracy and F1 scores within ± 0.001 as ties.

Model	Phylum			Class			Order		
	Attempts (%)	Acc.	F1	Attempts (%)	Acc.	F1	Attempts (%)	Acc.	F1
Simple RAG	936 (98.4%)	0.994	0.997	928 (97.6%)	0.976	<u>0.979</u>	860 (90.4%)	0.900	0.919
Advanced RAG	932 (98.0%)	0.995	0.997	926 (97.4%)	0.976	<u>0.978</u>	835 (87.8%)	<u>0.923</u>	<u>0.935</u>
Naïve GPT-4o	922 (96.9%)	<u>0.985</u>	<u>0.992</u>	922 (96.9%)	0.976	0.983	921 (96.9%)	0.921	0.933
Naïve Gemini 2.0 Flash	949 (99.8%)	<u>0.985</u>	<u>0.993</u>	949 (99.8%)	0.976	0.984	948 (99.7%)	0.929	0.942

Model	Family			Genus			Species		
	Attempts (%)	Acc.	F1	Attempts (%)	Acc.	F1	Attempts (%)	Acc.	F1
Simple RAG	419 (44.1%)	0.928	0.934	89 (9.4%)	<u>0.843</u>	<u>0.904</u>	43 (4.5%)	0.512	0.547
Advanced RAG	420 (44.1%)	<u>0.924</u>	<u>0.927</u>	82 (8.6%)	0.854	0.919	47 (4.9%)	<u>0.468</u>	<u>0.544</u>
Naïve GPT-4o	849 (89.3%)	0.873	0.909	459 (48.3%)	0.534	0.598	159 (16.7%)	<u>0.157</u>	0.223
Naïve Gemini 2.0 Flash	944 (99.3%)	0.798	0.836	763 (80.2%)	0.329	0.406	369 (38.8%)	0.195	0.265

3.4. Evaluation on IUCN Rare Species Dataset

Having validated our approach on the smaller development dataset, we conducted our primary evaluation on the more challenging IUCN Rare Species Dataset (Arthropod subset). Table 2 shows classification attempts, accuracy and F1 scores for the Simple RAG, Advanced RAG, Naïve GPT-4o and Naïve Gemini 2.0 Flash models on all 951 rare species images. Both RAG models consistently outperform both Naïve VLMs at the family, genus and species level. Both RAG pipelines show higher F1 scores compared to accuracy at the family, genus and species level, indicating elevated performance on minority classes. The Advanced RAG model’s more sophisticated retrieval techniques (multiquery, MMR, and reranking) did not yield significant improvements over the Simple RAG model, suggesting that simpler retrieval methods are sufficient for the available knowledge base. Processing 951 rare-species images, the Simple and Advanced RAG models accrued API costs totalling \$30.45 and \$19.98 (amounts in USD), respectively. These modest costs—together with the fact that key steps such as image captioning and retrieval are trivially parallelizable—make the system practical to scale for low-latency, large-volume biodiversity monitoring.

4. Discussion

In Table 1(a) we see that the classification performance of the RAG model relative to the Naïve LLM shows reasoning over the additional retrieved context provides a small boost to LLM confidence and a modest gain in classification accuracy, particularly at the family and genus level where it is most needed. However, the improved classification performance of the Naïve VLM indicates many taxa present in the Living Arthropod image dataset were likely seen during pre-training. Table 1(b) and Table 1(c) support this idea with high performance in more common taxa classification and poorer performance classifying more rare taxa. The RAG models did not see the same sharp decrease in accuracy and confidence as the Naïve LLM and Naïve VLMs in the rare taxa subset nor in the Rare Species subset Table 2, signalling the RAG model could be a viable method for rare and unknown species identification, particularly as more curated text-based knowledge sources become available.

Our RAG-based approach to taxonomic classification represents a promising new paradigm that leverages explicit textual descriptions rather than relying solely on visual patterns. While our current implementation using Wikimedia and Wikispecies data showed modest improvements over the Naïve LLM, the approach has significant untapped potential.

As citizen science initiatives develop and species trait characterization become more comprehensive, there is potential for our RAG pipeline to harness those improvements to better assist in identification and characterization of unknown species. Additionally, LLMs have recently been recruited to mine and curate data sources for biodiversity research [42] which would automate data curation and could soon boost the relevancy of RAG systems in Arthropoda taxonomy.

Complex retrieval methods like multi-query and reranking provided only modest improvements to classification and our efforts to improve the embedding space using semantic filtering of uninformative source chunks and contextualizing text did not yield marked improvements. This suggests that the current limiting factor is the breadth and depth of curated taxonomic text data, particularly at the genus and species levels. Alternative embedding structures such as GraphRAG [43] could improve retrieval, as demonstrated in other biological domains [23]. Additionally, many models have demonstrated the classification benefits of multimodal over unimodal embedding spaces [16, 17, 44], suggesting that RAG systems could similarly benefit from multimodal embeddings of images paired with text.

Our approach first converts visual features into text through image captioning, creating an abstraction layer for taxonomic classification. While our prompting strategy reduced inappropriate inferences during captioning (Appendix C), developing an organismal image-caption dataset with domain experts could enable fine-tuning of VLMs specifically for biological image captioning. Fine-tuning would help ensure captions contain only visually-evident features and could provide a tool to generate image-biocaption datasets. A fine-tuned captioner would, on its own, be valuable for automated trait descriptions of potentially novel taxa (i.e. BINs), aiding in both indexing of operational taxonomic units and functional inference. These detailed descriptions would in turn contribute to expanding the contextual knowledge bases available for taxonomic classification.

Recent work has shown VLMs can exhibit shape and texture biases [45], potentially limiting their inference to species seen during pre-training. By reasoning over explicit visible traits in captions combined with retrieved contextual knowledge, our RAG approach can better identify novel species compared to direct image classification. This advantage is particularly evident through evaluation on the Rare Species subset of uncommon species and could be extended by evaluation on the full Rare Species dataset [17]. Additionally, incorporating human feedback during fine-tuning, which has proven successful for calibrating LLM confidence scores [46], could further improve the quality of our generated biodiversity knowledge.

5. Conclusion

We have demonstrated that pairing RAG models with VLMs offers a promising approach to taxonomic classification of unknown Arthropoda species, particularly as biodiversity knowledge bases continue to improve. While Naïve VLMs excelled in identifying familiar species, our RAG approach can leverage contextual information to classify rare and unknown taxa. The growing integration of citizen science data will further enhance species identification and inform conservation strategies. We envision future work expanding these models to more diverse datasets, especially those including underrepresented species. Our work establishes a novel bridge between modern AI tools and traditional text-based knowledge bases in biodiversity conservation, combining advanced RAG techniques, complex LLM-tooling, and LLM-based evaluation. By adapting and incorporating emerging AI techniques, we can significantly advance our understanding and conservation of biodiversity.

Acknowledgements

Scott C. Lowe provided valuable feedback on results visualizations and manuscript readability. Elizabeth Campolongo provided technical assistance for Rare Species experiments. This project was initiated within the Vector Institute RAG Bootcamp. We thank the organizers and the staff who supported its implementation.

This work was supported by the [AI and Biodiversity Change \(ABC\) Global Center](#), which is funded by the US National Science Foundation under [Award No. 2330423](#) and Natural Sciences and Engineering Research Council of Canada under [Award No. 585136](#). We also acknowledge the support of the Government of Canada’s New Frontiers in Research Fund under [Award No. NFRFT-2020-00073](#). This research was supported, in part, by the Province of Ontario and the Government of Canada through the Canadian Institute for Advanced Research (CIFAR), and [companies sponsoring](#) the Vector Institute.

GWT is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada Research Chairs program, and the Canada CIFAR AI Chairs program.

References

- [1] S. Seibold et al. “Arthropod decline in grasslands and forests is associated with landscape-level drivers”. In: *Nature* 574.7780 (2019), pp. 671–674. ISSN: 1476-4687. DOI: [10.1038/s41586-019-1684-3](#).
- [2] L. H. Yang and C. Gratton. “Insects as drivers of ecosystem processes”. In: *Current Opinion in Insect Science* 2 (2014), pp. 26–32. ISSN: 2214-5745. DOI: [10.1016/j.cois.2014.06.004](#).
- [3] T. W. Culliney. “Role of Arthropods in Maintaining Soil Fertility”. In: *Agriculture* 3.4 (2013), pp. 629–659. ISSN: 2077-0472. DOI: [10.3390/agriculture3040629](#).
- [4] Z.-Q. Zhang. “Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness (Addenda 2013)”. In: *Zootaxa* 3703 (2013). Place: New Zealand, pp. 1–82. ISSN: 1175-5326. DOI: [10.11646/zootaxa.3703.1.1](#).
- [5] E. Hartop, L. Lee, A. Srivathsan, M. Jones, P. Peña-Aguilera, O. Ovaskainen, T. Roslin, and R. Meier. “Resolving biology’s dark matter: species richness, spatiotemporal distribution, and community composition of a dark taxon”. In: *BMC Biology* 22.1 (2024), p. 215. ISSN: 1741-7007. DOI: [10.1186/s12915-024-02010-z](#).
- [6] F. Ødegaard. “How many species of arthropods? Erwin’s estimate revised”. In: *Biological Journal of the Linnean Society* 71.4 (2000), pp. 583–597. ISSN: 0024-4066. DOI: [10.1111/j.1095-8312.2000.tb01279.x](#).
- [7] P. D. N. Hebert, A. Cywinska, S. L. Ball, and J. R. deWaard. “Biological identifications through DNA barcodes.” In: *Proceedings. Biological sciences* 270.1512 (2003). Place: England, pp. 313–321. ISSN: 0962-8452 1471-2954. DOI: [10.1098/rspb.2002.2218](#).
- [8] S. Ratnasingham and P. D. N. Hebert. “A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System”. In: *PLOS ONE* 8.7 (July 2013), pp. 1–16. DOI: [10.1371/journal.pone.0066213](#).
- [9] H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. de Almeida, H. Sirelkhatim, G. Richard, M. Skwark, K. Beguir, M. Lopez, and T. Pierrot. “The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics”. In: *Nature Methods* 22.2 (2025), pp. 287–297. ISSN: 1548-7105. DOI: [10.1038/s41592-024-02523-z](#).
- [10] E. Nguyen, M. Poli, M. Faizi, A. Thomas, C. Birch-Sykes, M. Wornow, A. Patel, C. Rabideau, S. Massaroli, Y. Bengio, S. Ermon, S. A. Baccus, and C. Ré. “HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 43177–43201. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/86ab6927ee4ae9bde4247793c46797c7-Paper-Conference.pdf.

- [11] P. Millan Arias, N. Sadjadi, M. Safari, Z. Gong, A. T. Wang, J. Bruslund Haurum, I. Zarubiieva, D. Steinke, L. Kari, A. X. Chang, S. C. Lowe, and G. W. Taylor. “Barcode-BERT: Transformers for Biodiversity Analysis”. In: *arXiv pre-print arXiv.2311.02401* (2023), arXiv:2311.02401. DOI: [10.48550/arXiv.2311.02401](https://doi.org/10.48550/arXiv.2311.02401).
- [12] L. Romeijn, A. Bernatavicius, and D. Vu. “MycoAI: Fast and accurate taxonomic classification for fungal ITS sequences”. In: *Molecular Ecology Resources* 24.8 (2024), e14006. DOI: [10.1111/1755-0998.14006](https://doi.org/10.1111/1755-0998.14006).
- [13] T. Gao and G. W. Taylor. “BarcodeMamba: State Space Models for Biodiversity Analysis”. In: *arXiv pre-print arXiv.2412.11084* (2024). DOI: [10.48550/arXiv.2412.11084](https://doi.org/10.48550/arXiv.2412.11084).
- [14] A. Jain, F. Cunha, M. Bunsen, L. Pasi, A. Viklund, M. Larrivée, and D. Rolnick. “A machine learning pipeline for automated insect monitoring”. In: *arXiv pre-print arXiv.2406.13031* (2024). DOI: [10.48550/arXiv.2406.13031](https://doi.org/10.48550/arXiv.2406.13031).
- [15] K. Bjerger, H. M. R. Mann, and T. T. Høy. “Real-time insect tracking and monitoring with computer vision and deep learning”. In: *Remote Sensing in Ecology and Conservation* 8.3 (2022), pp. 315–327. DOI: [10.1002/rse2.245](https://doi.org/10.1002/rse2.245).
- [16] Z. Gong, A. T. Wang, X. Huo, J. B. Haurum, S. C. Lowe, G. W. Taylor, and A. X. Chang. “CLIBD: Bridging Vision and Genomics for Biodiversity Monitoring at Scale”. In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: <https://openreview.net/forum?id=d5HUnyByAI>.
- [17] S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong, W. M. Dahdul, C. Stewart, T. Berger-Wolf, W.-L. Chao, and Y. Su. “BioCLIP: A Vision Foundation Model for the Tree of Life”. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2024, pp. 19412–19424. DOI: [10.1109/CVPR52733.2024.01836](https://doi.org/10.1109/CVPR52733.2024.01836).
- [18] S. Badirli, C. J. Picard, G. Mohler, F. Richert, Z. Akata, and M. Dundar. “Classifying the unknown: Insect identification with deep hierarchical Bayesian learning”. In: *Methods in Ecology and Evolution* 14.6 (2023), pp. 1515–1530. DOI: [10.1111/2041-210X.14104](https://doi.org/10.1111/2041-210X.14104).
- [19] iNaturalist. Available from <https://www.inaturalist.org>. Accessed Jan 31st, 2025. 2025.
- [20] S. Gupta, R. Ranjan, and S. N. Singh. “A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions”. In: *arXiv pre-print arXiv.2410.12837* (2024). DOI: [10.48550/arXiv.2410.12837](https://doi.org/10.48550/arXiv.2410.12837).
- [21] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.
- [22] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou. “Hallucination of Multimodal Large Language Models: A Survey”. In: *arXiv pre-print arXiv.2404.18930* (2024). DOI: [10.48550/arXiv.2404.18930](https://doi.org/10.48550/arXiv.2404.18930).
- [23] Z. Li, X. A. Chen, and Y. Jeon. “GraPPI: A Retrieve-Divide-Solve GraphRAG Framework for Large-scale Protein-protein Interaction Exploration”. In: *arXiv pre-print arXiv.2501.16382* (2025). DOI: [10.48550/arXiv.2501.16382](https://doi.org/10.48550/arXiv.2501.16382).
- [24] S. Zhang, G. Dai, T. Huang, and J. Chen. “Multimodal Large Language Models for Bioimage Analysis”. In: *Nature Methods* 21.8 (2024), pp. 1390–1393. ISSN: 1548-7105. DOI: [10.1038/s41592-024-02334-2](https://doi.org/10.1038/s41592-024-02334-2).
- [25] J. Kim, J. Lee, H.-J. Choi, T.-Y. Hsu, C.-Y. Huang, S. Kim, R. Rossi, T. Yu, C. L. Giles, T.-H. K. Huang, and S. Choi. “Multi-LLM Collaborative Caption Generation in Scientific Documents”. In: *arXiv pre-print arXiv.2501.02552* (2025). DOI: [10.48550/arXiv.2501.02552](https://doi.org/10.48550/arXiv.2501.02552).
- [26] Z. Gharaee, S. C. Lowe, Z. Gong, P. M. Arias, N. Pellegrino, A. T. Wang, J. B. Haurum, I. Zarubiieva, L. Kari, D. Steinke, G. W. Taylor, P. Fieguth, and A. X. Chang. “BIOSCAN-5M: A Multimodal Dataset for Insect Biodiversity”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37. Curran Associates, Inc., 2024, pp. 36285–36313. URL: <https://proceedings>

- ngs.neurips.cc/paper_files/paper/2024/file/3fdbb472813041c9ecef04c20c2b1e5a-Paper-Datasets_and_Benchmarks_Track.pdf.
- [27] R. A. Collins and R. H. Cruickshank. “The seven deadly sins of DNA barcoding”. In: *Molecular Ecology Resources* 13.6 (2013), pp. 969–975. DOI: [10.1111/1755-0998.12046](https://doi.org/10.1111/1755-0998.12046).
 - [28] International Barcode of Life Consortium. *IBOL’s BIOSCAN Initiative*. 2025. URL: <https://ibol.org/bioscan/> (visited on 02/16/2025).
 - [29] Biodiversity Genomics Europe. *BGE’s The Project*. 2024. URL: <https://biodiversitygenomics.eu/the-project/> (visited on 02/16/2025).
 - [30] H. Chase. *LangChain*. 2022. URL: <https://github.com/langchain-ai/langchain> (visited on 02/05/2025).
 - [31] J. Liu. *Instructor*. 2023. URL: <https://github.com/instructor-ai/instructor> (visited on 02/10/2025).
 - [32] J. Pfau, W. Merrill, and S. R. Bowman. “Let’s Think Dot by Dot: Hidden Computation in Transformer Language Models”. In: *First Conference on Language Modeling*. 2024. URL: <https://openreview.net/forum?id=NikbrdtVvG>.
 - [33] IUCN. *The IUCN Red List of Threatened Species. Version 2024-2*. 2024. URL: <https://www.iucnredlist.org> (visited on 02/16/2025).
 - [34] Anthropic. *Introducing Contextual Retrieval*. Anthropic. Sept. 2024. URL: <https://www.anthropic.com/news/contextual-retrieval> (visited on 02/05/2025).
 - [35] Chroma. *ChromaDB*. 2022. URL: <https://github.com/chroma-core/chroma> (visited on 02/10/2025).
 - [36] D. Zhang, J. Li, Z. Zeng, and F. Wang. “Jasper and Stella: distillation of SOTA embedding models”. In: *arXiv pre-print arXiv.2412.19048* (2025). DOI: [10.48550/arXiv.2412.19048](https://doi.org/10.48550/arXiv.2412.19048).
 - [37] OpenAI et al. “GPT-4o System Card”. In: *arXiv pre-print arXiv.2410.21276* (2024). DOI: [10.48550/arXiv.2410.21276](https://doi.org/10.48550/arXiv.2410.21276).
 - [38] Cohere. *Cohere’s Rerank Models*. 2024. URL: <https://docs.cohere.com/v2/docs/rerank-2> (visited on 02/16/2025).
 - [39] Google DeepMind. *Google DeepMind’s Gemini AI Models*. 2024. URL: <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#gemini-2-0-flash> (visited on 02/16/2025).
 - [40] iNaturalist community. *Observations of all species in Living Arthropod dataset from all geographic areas observed before Jan 30th, 2025. Exported from https://www.inaturalist.org on Jan 30th, 2025*. 2025.
 - [41] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert. “RAGAS: Automated Evaluation of Retrieval Augmented Generation”. In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Ed. by N. Aletras and O. De Clercq. Association for Computational Linguistics, 2024, pp. 150–158. URL: <https://aclanthology.org/2024.eacl-demo.16/>.
 - [42] J. D’Souza, Z. Laubach, T. A. Mustafa, S. Zarriek, R. Frühstückl, and P. Illari. “Mining for Species, Locations, Habitats, and Ecosystems from Scientific Papers in Invasion Biology: A Large-Scale Exploratory Study with Large Language Models”. In: *arXiv pre-print arXiv.2501.18287* (2025). DOI: [10.48550/arXiv.2501.18287](https://doi.org/10.48550/arXiv.2501.18287).
 - [43] H. Han et al. “Retrieval-Augmented Generation with Graphs (GraphRAG)”. In: *arXiv pre-print arXiv.2501.00309* (2025). DOI: [10.48550/arXiv.2501.00309](https://doi.org/10.48550/arXiv.2501.00309).
 - [44] S. Zhang et al. “BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs”. In: *arXiv pre-print arXiv.2303.00915* (2023). DOI: [10.48550/arXiv.2303.00915](https://doi.org/10.48550/arXiv.2303.00915).
 - [45] P. Gavrikov, J. Lukasik, S. Jung, R. Geirhos, B. Lamm, M. J. Mirza, M. Keuper, and J. Keuper. “Are Vision Language Models Texture or Shape Biased and Can We Steer Them?” In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: <https://openreview.net/forum?id=iVMcYxTiVM>.
 - [46] K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. D. Manning. “Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback”. In: *The 2023 Conference on Empirical Methods in Natural Language Processing*. 2023. URL: <https://openreview.net/forum?id=g3faCfrwm7>.

Code can be found at <https://github.com/uoguelph-mlrg/taxonomic-rag-system>.

Appendix A. Dataset Summary Tables

Table 3. Summary of taxonomic counts by rank in Living Arthropod dataset

Class	Order Count	Family Count	Genus Count	Species Count
Arachnida	3	16	27	32
Chilopoda	2	2	2	2
Diplopoda	3	3	3	4
Insecta	8	37	81	108
Malacostraca	2	7	7	7

Table 4. Summary of taxonomic counts by rank in Rare Species subset

Class	Order Count	Family Count	Genus Count	Species Count
Arachnida	1	2	2	3
Branchiopoda	1	1	1	1
Insecta	6	13	22	28
Malacostraca	1	1	1	1

Appendix B. Contextualizing Chunks Prompt

You are an expert AI assistant to a taxonomist. You can determine if a user-provided chunk of textual context is useful for a taxonomist and if it is, contextualize it for retrieval in the context of a larger document.

You are analyzing text from documents about organisms. For each chunk: 1. First determine if the chunk (alone) contains actual descriptive content about organisms. A chunk with just a citation, header, references, or other non-descriptive text should be marked as not useful. 2. If the chunk contains descriptive content about organisms, please give a short succinct context to situate this chunk within the overall document for the purposes of improving search retrieval of the chunk. Also include taxonomic classification from the document. 3. If the chunk does NOT contain descriptive content, briefly state what type of content it contains instead (e.g. "citation", "references header", etc).

You do not need to give an overview starting statement such as "This document provides a detailed description of the phylum Chordata" but rather, get right into what is being said about Chordata in the chunk.

Format your response as a parsable JSON with: - useful: boolean indicating if descriptive organism content is present - contextual_text: string containing either the descriptive contextualization or content type explanation

```
<document> {doc_content} </document>
```

Here is the chunk we want to situate within the whole document

```
<chunk> {chunk_content} </chunk>
```

Note: 256 thinking dot tokens are appended to the prompt above.

Appendix C. Module I Prompt

You are an expert AI vision assistant to a taxonomist that describes animals in images.

Your task is to describe in extensive detail all the physical features (body and head shape, appendages, colour pattern, shape, texture, etc) of any organism(s) observable in the image.

Ensure each feature is elaborated upon wherever possible. Elaborate on the visual traits and morphology of subsections of the organism such as any appendages (such as limbs, wings etc) that are visible in the image — aim to describe them thoroughly.

Additionally, include a comprehensive description of the current state and/or life stage of the organism, with nuances in coloration, wear, or any distinctive features.

Please describe the environmental context surrounding the animal in detail. For example, if a butterfly is resting on flowers, you should also delve into the unique characteristics of the flowers, such as their shape, color, and arrangement.

Avoid using species common names (such as 'Monarch' for a butterfly).

Aim for a detailed analysis of at least 7 sentences with no upper limit if the detail demands it.

Do not include emotional descriptors (such as 'peaceful setting') or any other non-visible descriptors. Everything you mention should be evident in the image to an observer. Your response must rely solely on visual cues from the image, avoiding any inferences that are not evident.

Write a extremely detailed caption for the organism in the image and without commenting on the contrast or 'feeling' of the image.

An example of the type of caption you should produce is: Insecta with 4 visible jointed legs, partially translucent wings and compound eyes. There is a three-part body with a head, thorax and abdomen. An anterior lateral view of an adult fly with an abdomen that is mostly black and has a black tail-like taper. The wings have streaks of white as does the thorax and are black elsewhere. The prescutum and scutum are brown and in addition to the head, have small shiny hairs. The wings attach at the middle of the thorax, as do the legs. The legs have an initial black segment but are mostly coppery-brown and terminate into a triangular base. The wings are not as long as the length of the body and lay relatively flat at an angle away from the body with 2 segmented translucent halteres. The head is copper, orange and brown with white bordering. The head is visibly segmented from the thorax but the thorax and abdomen appear continuous and not visibly segmented. One brownish-orange eye with a white border is fully visible and the other eye is partially visible. There are two coppery kidney-shaped mouth parts protruding from the lower front of the head. A single shiny antennae is visible. The fly is standing on a green leaf that has pointed edges.

Note: 256 thinking dot tokens are appended to the prompt above.

Appendix D. Module II Prompt

You are an expert AI taxonomist. Your task is to use the organisms discussed in the caption of a new organism to generate a taxonomic classification for the new organism.

You will also be provided with some context that could or could not match the caption, if there is information in the context that matches the caption, you can use that info to inform your decision about the taxonomic classification, otherwise, if information does not match the details provided in the caption, disregard it.

You will be provided with context and a caption, provide in your response: 1. A Taxonomic classification 2. A description pairing physical traits common to both the new organism described in the caption and other organisms described in the context that indicate and support the choice made in the taxonomic classification. 3. A description of physical traits particular to this new organism described in the caption. These traits may set it apart from other organisms, may suggest it has unique features, and/or contain traits that may be candidates to investigate for a more specific taxonomic classification. 4. Commentary on your choice, including discussion of confidence, what new information about the new organism would help support the taxonomic classification and what new information

would dispute the taxonomic classification. 5. A few paragraphs describing the features present (from the caption) and how they could relate to the biodiversity knowledge that is relevant to the taxa chosen.

Do not include a taxonomic classification for a certain rank unless you are confident from the caption (and/or context) about the classification.

<context> {context} </context>

<caption> {caption} </caption>

Note: 256 thinking dot tokens are appended to the prompt above.