

Ranking de modelos Multimodais para geração de descrições de espécies animais em armadilhas fotográficas da Amazônia

Ronald Vieira Cardoso Silva¹, Eulanda M. dos Santos¹, Juan G. Colonna^{1,2}

¹ Instituto de Computação - Universidade Federal do Amazonas (UFAM)

² School of Mathematics and Statistics - Victoria University of Wellington (VUW)

{ronald.silva, emsantos, juancolonna}@icomp.ufam.edu.br

Abstract. *Automatic analysis of camera-trap images is often limited by classifiers that fail to leverage ecological context. This work in progress investigates the use of Large Multimodal Models (LMMs) to address this gap, evaluating them through taxonomic metrics and an “Arena” methodology based on blind pairwise comparisons. Fourteen OpenAI models were tested using Elo Rating and the Bradley-Terry model. As a result, the EcoLMM Arena web platform was developed to systematize expert voting, generating human-preference-driven rankings for both performance and descriptive quality.*

Resumo. *A análise automática de imagens de armadilhas fotográficas é frequentemente limitada por classificadores que ignoram o contexto ecológico. Este trabalho em andamento investiga o uso de Grandes Modelos Multimodais (LMMs) para mitigar essa lacuna, avaliando-os por meio de métricas taxonômicas e uma metodologia de “Arena” baseada em comparações pareadas às cegas. Foram testados quatorze modelos da OpenAI utilizando os sistemas de pontuação Elo Rating e Bradley-Terry. Como resultado, desenvolveu-se a plataforma web EcoLMM Arena, que sistematiza a coleta de votos de especialistas para gerar rankings de desempenho e qualidade descritiva orientados pela preferência humana.*

1. Introdução

O monitoramento da biodiversidade na Amazônia via armadilhas fotográficas (*camera traps*) é padrão para estudos de fauna contínuos e não invasivos [O’Connell et al., 2011]. Contudo, o volume massivo de imagens gera um gargalo no processamento manual, elevando custos e demandando automação [Norouzzadeh et al., 2018]. Embora técnicas de Visão Computacional sejam aplicadas, a informação contida nas imagens vai além da taxonomia, englobando comportamentos e atributos fenotípicos [Gabeff et al., 2024].

Nesse contexto, os LMMs (*Large Multimodal Models*), como o ChatGPT e o Gemini, possuem a capacidade de processar dados visuais e gerar descrições textuais correspondentes [Wu et al., 2023]. Apesar da aplicação desses modelos em diversas áreas, ainda há uma carência de estudos que avaliem seu desempenho em imagens de armadilhas fotográficas. Outro desafio técnico reside na definição de métricas automáticas para avaliar a qualidade subjetiva das descrições geradas por tais modelos. Para contornar essa limitação, este projeto adapta a metodologia de “Arena” [Chiang et al., 2024], na qual os modelos são avaliados por meio de comparações pareadas (duelos) julgadas por seres



Figura 1. Diversidade de registros no dataset Mamirauá ilustrando variações de iluminação e horário: (a, b) *Panthera onca*; e (c) *Sapajus macrocephalus*.

humanos em testes cegos. Essa abordagem viabiliza a construção de um ranking qualitativo baseado na preferência humana, oferecendo um método objetivo para validar o uso de modelos generativos na análise de dados de monitoramento ambiental.

2. Trabalhos relacionados

Embora Redes Neurais Convolucionais (CNNs) dominem o processamento de armadilhas fotográficas, elas esbarram no desbalanceamento das classes e na exigência de gerar anotações em larga escala. Modelos de Visão-Linguagem (VLMs) e LMMs surgem como alternativas viáveis para cenários *zero-shot* e *few-shot*, reduzindo a dependência de dados rotulados [Alencar et al., 2024]. Ferramentas como *Mega Detector* e *Wildlife Insights* facilitam essa aplicação, enquanto estudos recentes já investigam LMMs para filtragem de imagens vazias.

Como os dados de origem das armadilhas fotográficas são naturalmente desbalanceados, ou seja algumas espécies são fotografadas com maior frequência, a literatura adota o Macro F1-Score como métrica fundamental. No entanto, pela natureza subjetiva das descrições textuais geradas por LMMs, a avaliação baseada em preferência humana destaca-se como o estado da arte. O Chatbot Arena [Chiang et al., 2024], por exemplo, converte duelos cegos em um ranqueamento global via ELO ranking ou o modelo estatístico de Bradley-Terry [Hunter, 2004], gerando uma avaliação mais alinhada com as expectativas humanas.

3. Metodologia


3.1. Base de Dados e Métricas

Neste trabalho foi utilizado o dataset do Instituto Mamirauá entre 2012-2016 (Figura 1) [Alencar et al., 2024]. O acervo de imagens compreende registros das classes *background* (5.046 amostras) e de 8 espécies: *Crax globulosa* (3.444), *Didelphis albiventris* (4), *Leopardus wiedii* (2.446), *Panthera onca* (5.161), *Pauxi tuberosa* (5.350), *Sapajus macrocephalus* (2.159), *Sciurus spadiceus* (45) e *Tupinambis teguixin* (4.344). Para lidar com o desbalanceamento natural das classes, adotou-se uma amostragem para os duelos estruturada em duas etapas com reposição: sorteia-se aleatoriamente a classe (espécie) e, em seguida, uma imagem pertencente a ela.

3.2. Sistema de Duelos e Engenharia de Prompts

A avaliação humana ocorre no sistema web EcoLMM Arena <https://ecollmduel-2.streamlit.app/>. Após a autenticação, é realizado um

Imagem da Armadilha



Esquilo-vermelho-da-Amazônia (Sciurus spadiceus) |

Prompt Enviado (Blind Test)

Você é um biólogo especialista em vida selvagem e reconhecimento de imagem.

Tarefa:
Analisar esta imagem de armadilha fotográfica e descreva detalhadamente o que é visível. Caso haja animais na imagem, identifique a espécie presente, informando o nome científico, o nome comum e o número de indivíduos detectados, considere imagens no contexto da selva amazônica brasileira.

Formato de saída:
Retorne a análise estritamente em formato JSON, sem qualquer texto adicional, contendo obrigatoriamente os seguintes campos:

Modelo A

Tempo: 11.06s

```
{
  "deteccao": "Sim"
  "nome_cientifico": "Dasyprocta leporina"
  "nome_comum": "Cutia"
  "numero_individuos": "1"
  "descricao_imagem": "
  \"A imagem de armadilha fotográfica mostra o interior de uma floresta, com solo coberto por folhas secas e vegetação densa ao fundo. No centro da cena há um pequeno mamífero de porte médio, com pelagem marrom-avermelhada, corpo alongado, orelhas pequenas e arredondadas e membros posteriores mais longos. O animal está apoiado sobre as patas traseiras, com o corpo ereto e as patas dianteiras próximas ao peito. Não há outros animais visíveis na imagem.\"
  \"razao\": \"
  \"O animal apresenta características típicas de uma cutia: corpo robusto e alongado, pelagem marrom a castanho-avermelhada, ausência de cauda visível, orelhas pequenas e postura ereta apoiada nas patas traseiras. Essa morfologia é compatível com Dasyprocta leporina, espécie comum na Amazônia brasileira. Apenas um indivíduo é visível na imagem.\"
  \"
}
```

Modelo B

Tempo: 5.49s

```
{
  "deteccao": "Sim"
  "nome_cientifico": "Sciurus spadiceus"
  "nome_comum": "Esquilo-de-barriga-vermelha"
  "numero_individuos": "1"
  "descricao_imagem": "
  \"A imagem mostra o sub-bosque de uma floresta densa, com folhas secas cobrindo o solo e árvores altas ao fundo. Em destaque, próximo ao centro inferior da imagem, há um esquilo de pelagem marrom-avermelhada e cauda espessa erguida, forrageando no solo.\"
  \"razao\": \"
  \"A identificação foi baseada na presença de um pequeno mamífero com corpo ágil, cauda longa e peluda, e pelagem marrom-avermelhada, características típicas do esquilo-de-barriga-vermelha, comum na Amazônia brasileira.\"
  \"
}
```

Figura 2. Interface de duelo no EcoLMM Arena. Acima, a imagem com o rótulo real (*Sciurus spadiceus*) e parte do *prompt*. Abaixo, as saídas em JSON, onde o Modelo A erra e o Modelo B acerta o reconhecimento do esquilo.

sorteio de uma imagem de armadilha fotográfica e são apresentados ao avaliador o *prompt* do LMM, a imagem, o rótulo real (oculto aos modelos) e as respostas em JSON dos dois modelos em um duelo lado a lado (Figura 2). A identidade dos modelos permanece anonimizada para evitar vieses dos avaliadores humanos. A tarefa do avaliador é escolher a melhor resposta baseada na justificativa e na descrição da imagem, bem como no acerto ou erro da espécie presente na fotografia. Os votos podem ser a favor de um modelo, empate satisfatório ou empate insatisfatório, este último quando ambos os modelos erram ou produzem descrições imprecisas em relação ao contexto.

Para mitigar ambiguidades, a saída textual é estruturada em JSON. O sistema seleciona, de forma equiprovável, entre dois *zero-shot prompts* distintos: um totalmente cego e outro condicionado com informações *a priori* sobre as espécies da região monitorada. Estas abordagens permitem avaliar o desempenho dos modelos tanto em contextos genéricos quanto em estudos específicos direcionados a espécies-alvo.

4. Resultados e Próximos Passos

No primeiro conjunto de experimentos, foram avaliados 14 modelos da OpenAI acessados via API (incluindo as famílias GPT-4, GPT-4o e GPT-5), o que permitiu estabelecer um ranqueamento inicial entre eles. Este ranking integra o julgamento humano ao desem-

penho taxonômico: após cada nova avaliação, a plataforma atualiza o *Elo Rating* e aplica o modelo de Bradley-Terry, focando na qualidade das descrições.

Objetivamente, a predição das espécies é aferida por métricas taxonômicas clássicas: *One-vs-Rest* (F1-score binário), matrizes de confusão e **Macro F1-Score**. Tais métricas são calculadas exclusivamente sobre o subconjunto de espécies com as quais cada modelo interagiu durante o duelo. Tais resultados parciais podem ser acessados pela plataforma de avaliação, a qual contém todos os resultados parciais expostos.

Como a coleta de dados na EcoLMM Arena é contínua, os resultados preliminares validam a funcionalidade, estabilidade e usabilidade da plataforma. Embora a arquitetura do sistema suporte IAs de múltiplos provedores, o escopo inicial restringiu-se aos modelos da OpenAI devido ao tempo hábil de testes. Como trabalhos futuros, planeja-se expandir o catálogo com modelos de código aberto e de outros provedores, visando estabelecer um novo *benchmark* multimodal para o monitoramento ecológico na Amazônia.

Agradecimentos

Este trabalho foi apoiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Código de Financiamento 001 (AUXPE-PROEX), e pela Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM), por meio dos projetos PDPG/-CAPES e POSGRAD 2026/2027. Agradecemos à OpenAI pelo apoio por meio da parceria estabelecida com seu representante, Nicolas Robinson Andrade.

Referências

- L. Alencar, F. Cunha, and E. M. Dos Santos. Zero and few-shot learning with modern mllms to filter empty images in camera trap data. In *2024 37th SIBGRAP Conference on Graphics, Patterns and Images (SIBGRAP)*, pages 1–6, 2024. doi: 10.1109/SIBGRAP62404.2024.10716305.
- W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, and J. E. Gonzalez. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024. doi: 10.48550/arXiv.2403.04132.
- V. Gabeff, M. Rußwurm, D. Tuia, and A. Mathis. WildCLIP: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. *International Journal of Computer Vision*, 132(9):3770–3786, 2024. doi: 10.1007/s11263-024-02026-6.
- D. R. Hunter. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406, 2004. doi: 10.1214/aos/1079120141.
- M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725, 2018. doi: 10.1073/pnas.1719367115.
- A. F. O’Connell, J. D. Nichols, and K. U. Karanth. *Camera traps in animal ecology: methods and analyses*. Springer Tokyo, 2011. ISBN 978-4-431-99495-4. doi: 10.1007/978-4-431-99495-4.
- J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu. Multimodal Large Language Models: A Survey. In *IEEE International Conference on Big Data (BigData)*, pages 2247–2256, 2023. doi: 10.1109/BigData59044.2023.10386743.