Understanding Visual Shortcomings of Multimodal Large Language Model Through Training Data Distribution

Stanford CS224N {Custom} Project

Alan Li
Department of Radiation Oncology
Stanford University
`lby@stanford.edu`

Binxu Li
Department of Electrical Engineering
Stanford University
`andy0207@stanford.edu`

Abstract

AI-assisted clinical workflows are significantly transforming the healthcare paradigm. One of the specific clinical tasks is the automation of radiation therapy treatment planning, where multimodal-language models (MLLMs) have the potential to play a pivotal role. However, the current automation pipeline faces limitations due to visual shortcomings in MLLMs, where obviously inaccurate text descriptions are generated given images. To resolve this, we sought to explore the root causes of these visual shortcomings in MLLMs. In this project, we hypothesize that the visual shortcomings are fundamentally due to the biases in training data distribution. We aim to systematically explore the effects of training data distribution on visual shortcomings in MLLMs. Our results indicate evidences of a potential positive correlation between the frequencies of semantic units in training data and the accuracies of MLLMs on questions about semantic units.

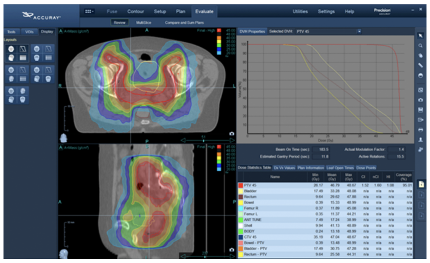
External Collaborators: Sheng Liu, Department of Radiation Oncology, Stanford University, `shengliu66@stanford.edu`
Mentor: Yuhui Zhang, `yuhuiz@stanford.edu`

1 Introduction

AI-assisted care has seen rapid progression in the field of radiation oncology over the past few years, with methods like accurate tumor and organ segmentation using CNNs being actively integrated into clinical workflows for treatment plans with improved dose optimization Huynh (2020). Recently, the emergence of large language models has prompted researchers to explore their integration into the radiation oncology domain Holmes (2023); Fabio Dennstädt (2024); Holmes (2024). During one particular stage of the radiation therapy treatment planning pipeline that we are currently developing, we obtain color-coded treatment plans similar to Fig. 1, and we aim to query the multimodal large language models (MLLMs) to identify rough anatomical regions of the images that are labeled with specific colors.

MLLMs, which integrate data from other modalities into Large Language Models (LLMs), leveraging the powerful capabilities of LLMs to demonstrate exceptional performance in tasks such as image understanding and visual question answering (VQA) Dai et al. (2023); Liu et al. (2023). Identifying rough anatomical regions based on color in radiation therapy treatment planning, while straightforward for clinicians, is a task we aim to fully automate with AI. However, for this seemingly simple task, MLLMs often fail and disrupt the pipeline. Concurrently, it has been observed that many MLLMs such as LLaVA, GPT-4V, do not possess very accurate color recognition and understanding capabilities of natural images, further illustrating the visual shortcomings in MLLMs.

We hypothesize that the observed visual shortcomings in MLLMs may stem from the influence of the distribution of training data on their core component, CLIP, inspired by MetaCLIP Xu et al. (2023).



Beam On Time (sec): 183.5
Actual Modulation Factor: 1.4
Estimated Gantry Period (sec): 11.8
Active Rotations: 15.5

	Name	Min (Gy)	Mean (Gy)	Max (Gy)	CI	nCI	HI	Coverage (%)
🟥	PTV 45	26.17	46.79	48.67	1.52	1.60	1.08	95.01
🟨	Bladder	17.49	33.28	48.08	n/a	n/a	n/a	n/a
🟫	Rectum	9.64	29.62	47.88	n/a	n/a	n/a	n/a
🟨	Bowel	0.39	15.33	48.99	n/a	n/a	n/a	n/a
🟦	Femur R	0.37	11.89	45.08	n/a	n/a	n/a	n/a
🟩	Femur L	0.35	11.37	44.21	n/a	n/a	n/a	n/a
🟩	ANT TUNE	7.49	17.24	38.99	n/a	n/a	n/a	n/a
⬜	Shell	9.94	41.13	48.89	n/a	n/a	n/a	n/a
🟩	BODY	0.24	13.18	48.99	n/a	n/a	n/a	n/a
🟦	CTV 45	35.19	47.04	48.67	n/a	n/a	n/a	n/a
🟥	Bowel - PTV	0.39	13.48	48.99	n/a	n/a	n/a	n/a
🟧	Bladder - PTV	17.49	30.75	47.28	n/a	n/a	n/a	n/a
🟨	Rectum - PTV	9.64	25.58	44.31	n/a	n/a	n/a	n/a

Figure 1: Example treatment plan. Credit: Google

MetaCLIP make a efficient data curation pipeline and retrain the metaclip by it, which achieves similar result compared with clip in many downstream tasks. During the data curation process, MetaCLIP specifically points out the issues that arises when the frequencies of an entry in the training data is below a certain threshold and mitigates this problem by forcing the frequencies to match the threshold.

To assess the model's understanding of visual and textual pairs composed of different frequencies, we developed a novel evaluation pipeline. We first selected an object and one of its attributes, such as {building, color} or the {animal, action}, to combine into a semantic unit. Then we attempted to compute the frequencies of semantic units from the captions in metadata. These semantic units then serve as the basis for constructing questions designed to test the CLIP models' visual interpretative capabilities. This pipeline is generalizable to all semantic units consisted of an object and its attributes, which includes the problem in treatment planning and many other VQA tasks.

Through our proposed evaluation pipeline and detailed analysis, our results indicate evidences of a potential positive correlation between the frequencies of semantic units in training data and the accuracies of MLLMs. Utilizing this correlation, MLLMs may be further refined during the training stage or the fine tuning stage, enhancing their understanding of complex visual-textual interactions. Additionally, we may develop an accuracy prediction model based on the frequencies of different semantic units for existing MLLMs.

2 Related Work

2.1 Importance of Understanding CLIP's Data Distribution

The distribution of training data directly impacts the performance, robustness of any machine learning model and also, it is a critical factor in the success of data curation which is the key for understanding how OpenAI trained CLIP Słowik and Bottou (2021). Ensuring a balanced and representative distribution of training examples allows researchers to devise more effective and efficient algorithms.

2.2 Handling Noisy Internet Data

Handling Noisy Internet Data. Addressing noisy data from the Internet is a significant challenge, and existing approaches often heavily rely on human-designed filter systems. Classical methods involve dataset cleaning and outlier removal to discard samples that may introduce undesirable biases to models Jiang et al. (2001).

2.3 Evaluating MLLM's visual Understanding Capabilities.

Existing works evaluates the model's visual understanding capabilities on visual multichoice or captioning tasks with metrics such as accuracy, recall, and CIDer Peng et al. (2023); Chen et al. (2023); You et al. (2023). However, these metrics fall short when it comes to evaluating visual dialogue for large multimodal models in an open-world setting. To evaluate MLLM's capability in engaging in visual conversations for image-level understanding, two families of evaluation are proposed: multiple-choice or using GPT-4 as a judge for free-form answers. However, these methods do not adequately explain the issues between understanding capabilities and training data. Therefore, we propose our evaluation method to more fairly assess the model's visual understanding abilities.

3 Approach

Our goal is to show a positive correlation between the frequency of semantic units in the training data and the accuracy of multimodal language models. This section describes how the frequency of semantic units are calculated and how the accuracies of multimodal language models are computed.

3.1 Extracting frequency of semantic units from training data

The semantic units are composed of an object and one of its attributes that describes the object's state or action. Although the semantic unit in the case of treatment planning composed of "anatomical region" and "color", the amount of testing data was limited to show a general relationship between training data distribution and MLLMs accuracy. Additionally, the training data does not contain medical images with treatment planning labels. Therefore, we looked for another object to pair with the "color" in the semantic unit. This turned out to be very noisy as "color" is a concept that is very widely used for many objects, making the relationshp between frequencies and accuracies extremely complex. Since to accurately compute frequencies of semantic units is a work in progress, we have decided to use a surrogate semantic unit that is easier to demonstrate the principle relationship. To do that, we have chosen "animal" and "action" as our units, which has a wide frequency range for each element in the semantic unit. Since only certain animals will perform certain actions, this strong association makes it simpler to compute frequencies.

We sought to count the number of times a semantic units appeared in the training data caption. The training data that we were parsing consisted of too many captions that were unrealistic to perform complicated dependency parsing for each caption. Thus, as a first pass, we applied a very coarse parsing scheme using python multiprocessing to find all captions that contained the object and its actions in the same sentence. From this step, we can calculate an unfiltered noisy frequency of semantic units from the training data.

3.2 Training data denoising for frequency calculation

The noise in the training data is defined as deviation from ideal meanings of the semantic units. From our experience dealing with the training data, this noise can be loosely categorized into the three categories, the intra-caption noise, the intra-image noise and the inter-caption-image-noise. We will describe how we attempted to filter out the first two kind of noises, where the filtering are applied in the order of computation complexity to optimize for computation time. Then, we will describe how it can be challenging to filter out the third kind of noise.

3.2.1 Intra-caption noise

The intra-caption noise exists when the description word in the caption is not directly describing the object of interest. For example, some possible intra-caption noises are listed in Table 1, considering the semantic units consist of an animal and its action, and the animal is cat and the action is sit.

Table 1: Example of intra-caption noises

Example numbers	Example captions
1	While the cat roamed, we were told to sit.
2	I saw a cat dash past as I was about to sit.
3	The story mentioned a cat right when the character decided to sit.

First, we used the Natural Language Toolbox (NLTK) to Bird et al. (2009) filter out the verbs in each caption. Then, we applied the spaCy to compute the dependency of every verbs to the object and removed the captions that don't have a direct relationship between the object and the action wordHonnibal and Montani (2017).

3.2.2 Intra-image noise

The intra-image noise exists when the caption by itself is reasonable where the action word is describing the object, but the image is entirely unrelated to the caption for some reason. As seen in

Fig. 2, for the animal hamster, many captions consisted phrases similar to "hamster chew toys" but the images are actually about the "chew toys" without the presence of any hamsters.

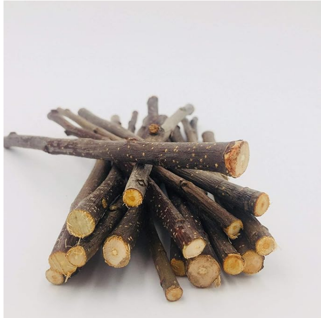


Figure 2: Example of intra-image noises: Hamster chew toys

We tried to use the CLIP model to reduce intra-image noise. Specifically, we computed the cosine similarity between the CLIP embedding for the image and the one word that represent the object. In the previous hamster case, the cosine similarity is computed between the embedding of the word "hamster" and the images of chew toys, which results in a low cosine similarity and is filtered out.

3.2.3 Inter-caption-image noise

The inter-caption-image noise exists usually when multiple meanings can be a reasonable match to a caption corresponding to one semantic unit. For example, for an image of a hamster eating something, the hamster can also simultaneously be standing or sitting. In this case, it is challenging to pin point the caption-image pair to a specific semantic unit. We think it will be difficult to accurately and robustly capture all possible scenarios with rule-based methods for calculation of both the frequencies of semantic units and later on the accuracy of multimodal language models.

3.2.4 Further denoising using action word categorization

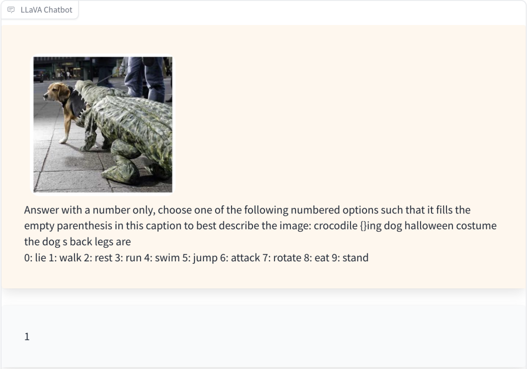
Since many of the action words have similar visual appearance on images and are reasonable replacements for each other in the caption, we attempted to group them together. For example, the words "eat" and "feed" mean the same action and show up similarly on images usually. The word "feed" may have less frequency, but the multimodal language models have learned the action of feeding from many other examples of eating. This phenomenon may introduce bias for certain low frequency words that have shared meanings with high frequency words. Additionally, the action words that have similar visual appearance may be different between animals. To solve this problem, we queried chatGPT to provide a list of lists of action words that a specific animal can possibly perform. Then, we computed the frequency and accuracy by category instead of by each action word.

3.3 Testing CLIP model and LLaVA accuracy on caption-image pairs

For each of the caption-image pair, we generated the a list of possible captions where the ground truth action word is replaced with other action words on the list for a specific animal. Then, for testing CLIP, we computed the probablity of CLIP selecting one of the captions for an image, and we selected the caption with the largest probablity as the final answer of CLIP. For testing LLaVA, as Figure3 we directly set the multiple-choice question as prompt and obatin a LLaVA reponse to see if the answer is right.

4 Experiments

Based on the semantic units we designed {animal, action}, we tested similarity of image and caption through CLIP and conduct multiple-choice questions on LLaVA, using the captions generated by us to pose questions.



Answer with a number only, choose one of the following numbered options such that it fills the empty parenthesis in this caption to best describe the image: crocodile {}ing dog halloween costume the dog s back legs are
0: lie 1: walk 2: rest 3: run 4: swim 5: jump 6: attack 7: rotate 8: eat 9: stand

1

Figure 3: LLaVA test demo

4.1 Data

This project involves the LAION-400M dataset, which is a open source and large-scale collection of image and corresponding caption pairs. And it's filtered with OpenAI‘s CLIP by calculating the cosine similarity between the text and image embeddings and dropping those with a similarity below 0.3Schuhmann et al. (2021).

We have selected ten animal objects from the image caption, which are *cobra, hippo, hamster, elephant, spider, snake, crocodile, alpaca, mosquito, frog* and use spacy to extract the action vocabulary related to these animals, then we make prompt:

> Title: Organizing {animal} Behavior Terms into Categories
>
> Prompt: I have a list of verbs and terms, some relevant to {animal} behaviors and others not. My goal is to first filter out the irrelevant terms and then organize the {animal}-specific actions into distinct groups based on visual or thematic similarities. This effort aims to thoroughly represent {animal} life across a spectrum of activities, care routines, interactions, and emotional expressions.
>
> Given the list below, could you summarize the approach you would take to:
>
> 1.Filter out non-{animal}-related terms from the list.
>
> 2.Categorize the relevant {animal} actions into groups, ensuring these categories are visually or thematically distinct. Furthermore, please ensure that the final categories are mutually exclusive, contain only the appropriate terms from the provided list, and present the categorized actions in a Python list of lists format for clarity.

to extract animal actions categories by GPT-3.5, Table 3 list all the categories and the corresponding meanings.

For testing each image-caption pair, we extract the object word from the caption, and compute the cosine similarity between the object word and the image with CLIP to make sure the object is present in the image. We set a threshold of 0.2 where image-caption pairs that have cosine similarity higher than the threshold are then subjected to further testing with CLIP and LLaVA.

4.2 Evaluation method

Since the output of the model is limited to two types ("correct" or "incorrect"), it is convenient to measure the metrics of accuracy for both CLIP and LLaVA. For similar words in a category, we calculated the categorical frequency by summing the frequencies and the categorical accuracy by combining the correct counts. Then, we computed a linear interpolation and R^2 for accuracy vs frequency plots for both using keyword and categories.

4.3 Results

4.3.1 CLIP test result

Results from keyword frequency don't show a definitive relationship while results from categorical frequency show a positive relationship for 8 out of 10 testing objects. The results are recorded in Table 2. Figure 4 through Figure 23 contains the scatter plots and linear fittings.

animals	R^2-keyword	keyword relationship(positive:+/negative:-)	R^2-category	category relationship(positive:+/negative:-)
alpaca	0.075	-	0.976	+
cobra	0.001	-	0.211	+
crocodile	0.049	+	0.233	+
elephant	0.000	-	0.379	+
frog	0.047	+	0.137	+
hamster	0.210	-	0.807	-
hippo	0.126	+	0.835	+
mosquito	0.547	-	0.444	+
snake	0.307	+	0.963	+
spider	0.108	+	0.148	-

Table 2: CLIP test result

4.3.2 LLaVA test result

Due to computational resource and time constraints, we only present the accuracy results (Figure 24, 25) of answers for 2 objects and their associated actions after categorization in the testing of the LLaVA model because the testing for these two objects is more comprehensive.

5 Analysis

We noticed that certain words that share very similar meanings may have drastically different computed frequencies. For example, the word feed and the word eat have similar visual appearances for many animals, but the frequency of feed is usually much lower compared to eat. Mostly due to this effect, the accuracy vs frequency plots show a relatively random relationship when using only keyword frequency and accuracy, where we see half and half negative and positive correlation.

The categorical frequency and accuracy was designed to solve this issue. From the results, we can indeed see that 8 out of 10 testing objects showed a positive correlation. While this can't prove that there is a positive correlation, this is suggesting that there seem to be evidence of such positive correlation. This signal is especially interesting since our denoising methods are far from optimized.

The results of the two objects in LLaVA also indicate that frequency has a significant impact on the model's visual understanding ability. Of course, the testing of LLaVA requires a more comprehensive testing process to be more statistically significant and to better prove our hypothesis.

6 Conclusion

In conclusion, we have built a pipeline to extract the frequencies of specified semantic units from a large image url + caption dataset. We tested the accuracy of CLIP models by computing the probability between images from URLs and multiple caption options. The result shows 8 out of 10 test objects demonstrated a positive correlation between the frequencies of semantic units and the accuracy. We also tested the accuracy of LLaVA for 2 objects by prompting multiple choice question related to the semantic unit based on caption, the result also indicates the positive correlation between frequencies and accuracy. In the future, we plan to further refine the caption denoising methods and increase the sample size for further denoising, and test more MLLMs with semantic units to makes our conclusions more robust. With proper denoising, we aim to apply the method to the treatment planning problem to analyze the more subtle relationship between training data distribution and

MLLMs accuracy. Then, strategies to improve the accuracy may be adopted or accuracy prediction models may be employed based on the discovered relationship.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Paul Martin Putora Erwin Vu Galina F. Fischer Krisztian Süveg Markus Glatzer Elena Riggenbach Hông-Linh Hà Nikola Cihoric Fabio Dennstädt, Janna Hastings. 2024. Exploring capabilities of large language models such as chatgpt in radiation oncology. *Advances in Radiation Oncology*, 9.

Liu Z. Zhang L. Ding Y. Sio T. T. McGee L. A. Ashman J. B. Li X.-Liu T. Shen J. Liu W. Holmes, J. 2023. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Frontiers in oncology*, 13.

Liu Z. Zhang L. Ding Y. Sio T. T. McGee L. A. Ashman J. B. Li X.-Liu T. Shen J. Liu W. Holmes, J. 2024. Exploring the capabilities and limitations of large language models for radiation oncology decision support. *International Journal of Radiation Oncology*Biology*Physics*, 118.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Hosny A. Guthier C. et al Huynh, E. 2020. Artificial intelligence in radiation oncology. *Nat Rev Clin Oncol*, 17:771–781.

M. F. Jiang, S. S. Tseng, and Chih-Ming Su. 2001. Two-phase clustering process for outliers detection. *Pattern Recognit. Lett.*, 22:691–700.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.

Agnieszka Słowik and Léon Bottou. 2021. Algorithmic bias and data bias: Understanding the relation between distributionally robust optimization and data curation.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023. Demystifying clip data.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity.

7 Appendix

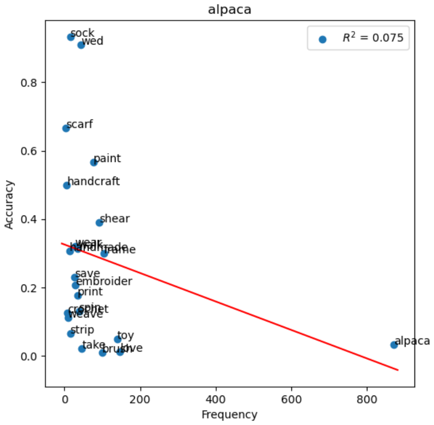


Figure 4: alpaca accuracy vs frequency

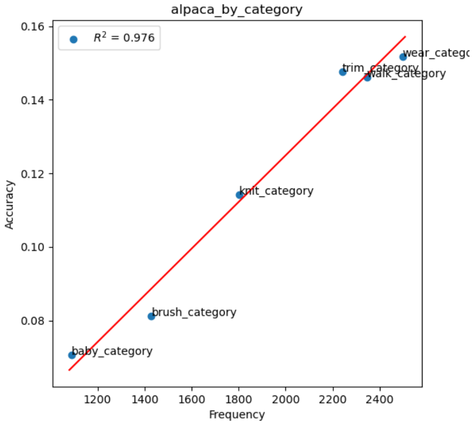


Figure 5: alpaca accuracy vs frequency by verb category

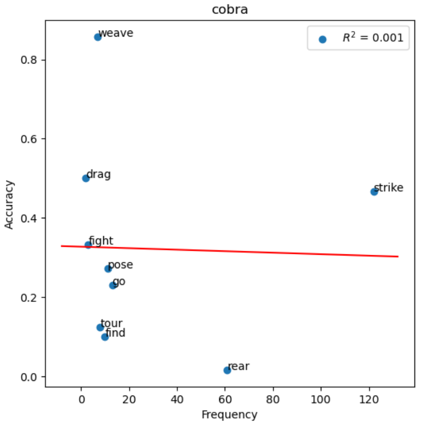


Figure 6: cobra accuracy vs frequency

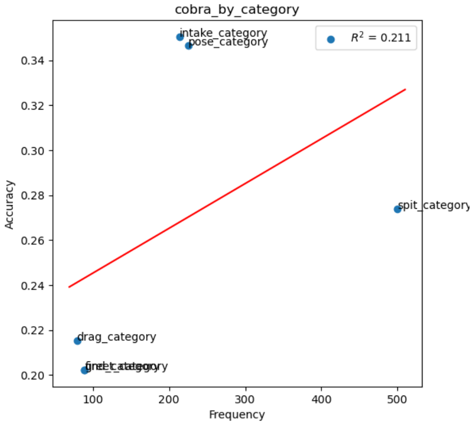


Figure 7: cobra accuracy vs frequency by verb category

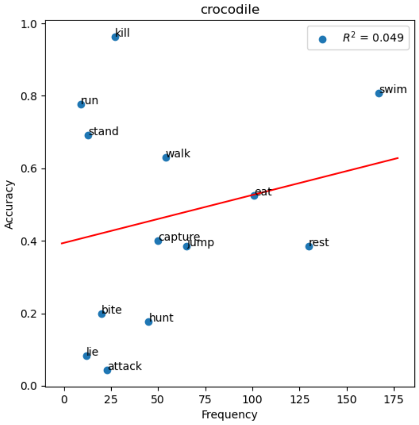


Figure 8: crocodile accuracy vs frequency

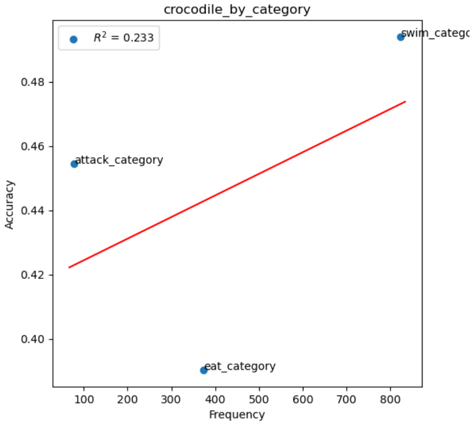


Figure 9: crocodile accuracy vs frequency by verb category

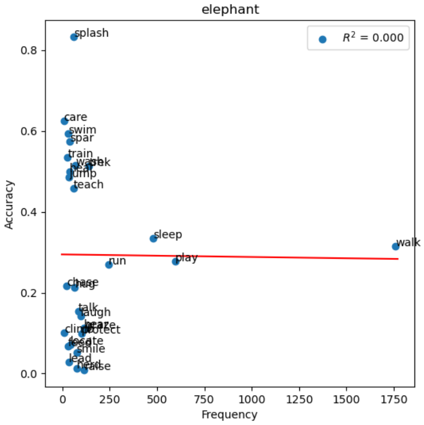


Figure 10: elephant accuracy vs frequency

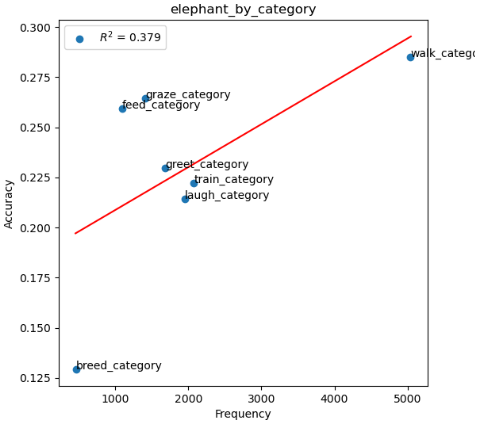


Figure 11: elephant accuracy vs frequency by verb category

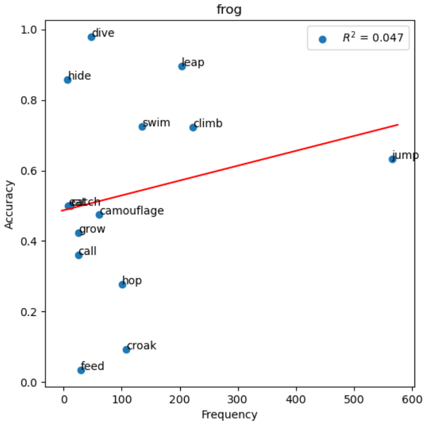


Figure 12: frog accuracy vs frequency

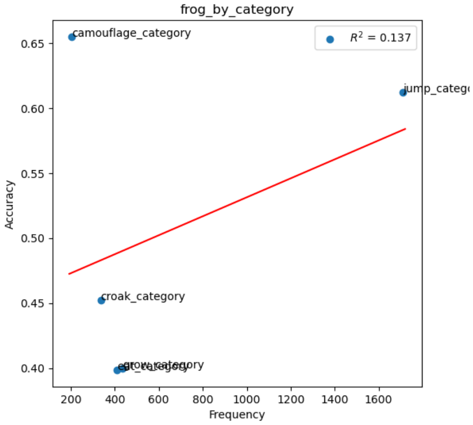


Figure 13: frog accuracy vs frequency by verb category

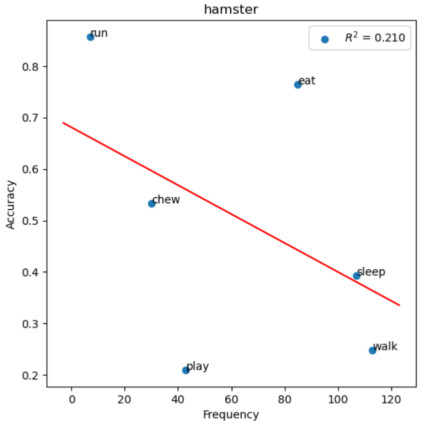


Figure 14: hamster accuracy vs frequency

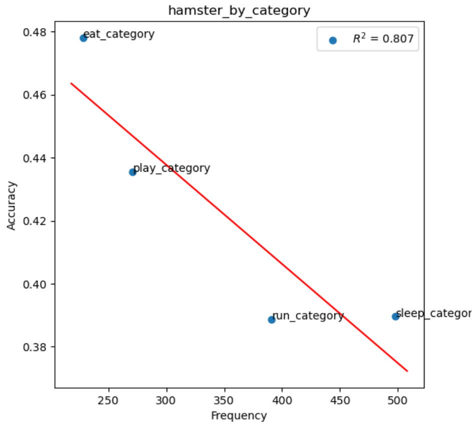


Figure 15: hamster accuracy vs frequency by verb category

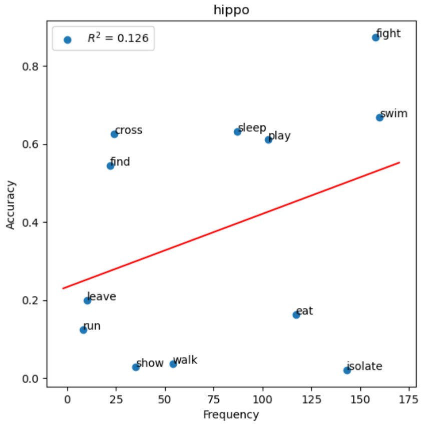


Figure 16: hippo accuracy vs frequency

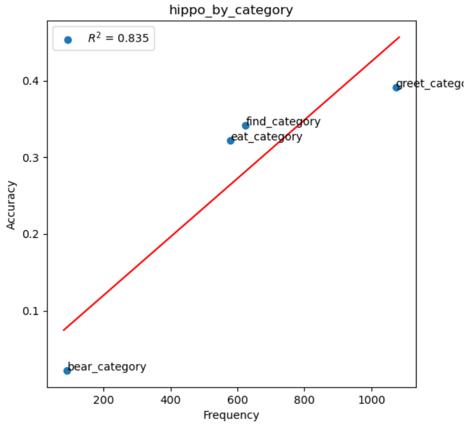


Figure 17: hippo accuracy vs frequency by verb category

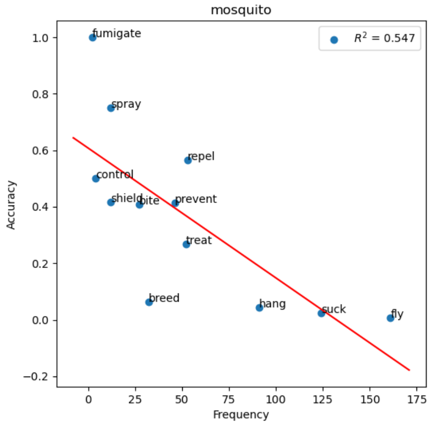


Figure 18: mosquito accuracy vs frequency

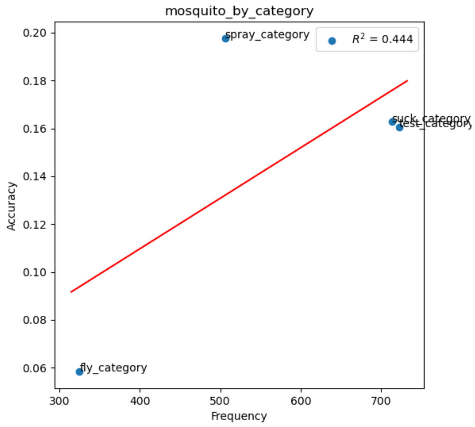


Figure 19: mosquito accuracy vs frequency by verb category

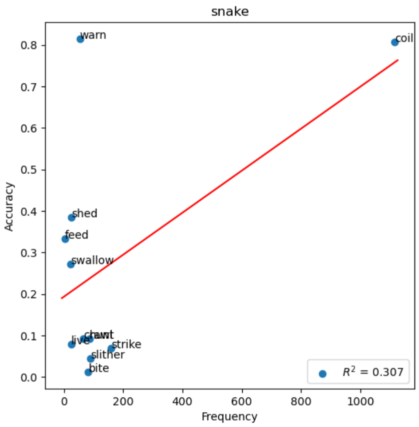


Figure 20: snake accuracy vs frequency

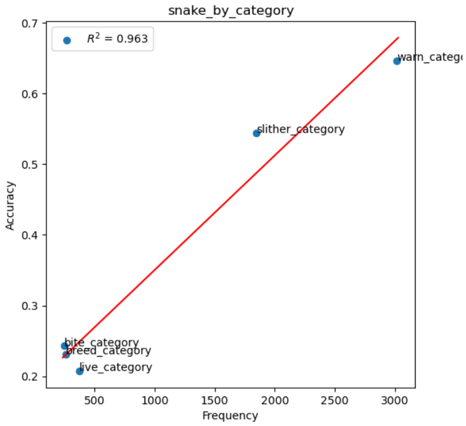


Figure 21: snake accuracy vs frequency by verb category

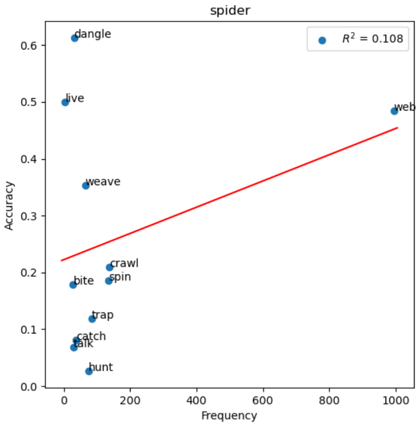


Figure 22: spider accuracy vs frequency

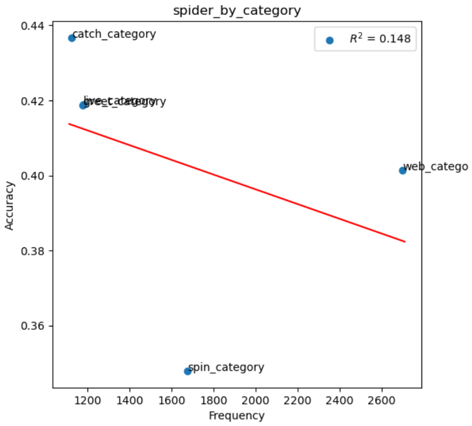


Figure 23: spider accuracy vs frequency by verb category

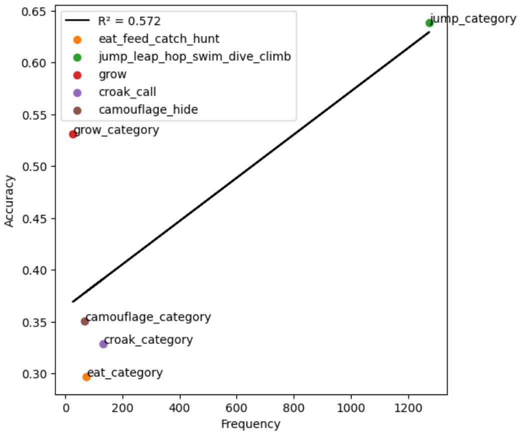


Figure 24: hippo accuracy vs category frequency

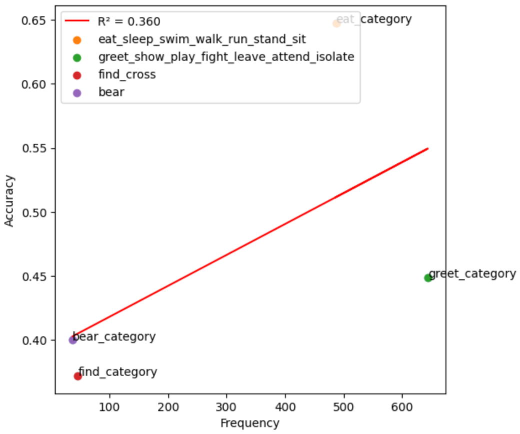


Figure 25: frog accuracy vs category frequency

Animal	Category	Meaning
alpaca	brush, greet, love, shear	Care and Interaction
	walk, show, take	Activity and Presentation
	knit, crochet, weave, spin, dye, paint, handcraft, craft, embroider, frame	Handiwork and Crafts
	wear, sock, scarf, glove, hood, handmade, fringe, pattern, strip, rib, print	Apparel and Textiles
	baby, alpaca, toy	Offspring and Play
	trim, shear, dye, paint, style, wed, look, save	Grooming and Aesthetics
cobra	spit, strike, rear, hold, fight	Defense Mechanisms
	drag, weave, propel, go, tour	Movement
	intake, strike	Feeding and Hunting
	greet	Social Interaction
	pose	Body Language and Communication
	find	Environmental Interaction
crocodile	swim, walk, run, jump, rest, lie, stand, rotate	Movement and Activity
	eat, feed, bite, hunt, capture, kill	Feeding Behavior
	attack	Defensive Behaviors
elephant	greet, herd, meet, roam, share, surround, talk, spend	Social Interactions
	walk, run, climb, swim, play, chase, jump, splash, trek, gather, spar	Physical Activities
	feed, sleep, wash, care, heal	Care and Maintenance
	laugh, hug, smile, wait	Emotional Expressions
	train, lead, teach	Training and Handling
	graze, locate, trek, roam	Environmental Interaction
	breed, bear, raise, protect	Reproduction
frog	eat, feed, catch, hunt	Feeding Behaviors
	jump, leap, hop, swim, dive, climb	Movement
	grow	Reproductive and Life Cycle
	croak, call	Vocalizations and Sound Production
	camouflage, hide	Defensive Mechanisms
hamster	run, walk	Moverment
	eat, chew	Feeding Behaviors
	play	Playing
	sleep	Resting
hippo	eat, sleep, swim, walk, run, stand, sit	Basic Behaviors
	greet, show, play, fight, leave, attend, isolate	Social Interactions
	find, cross	Habitat and Environment
	bear	Reproduction
mosquito	suck, transmit, bite, carry, breed	Feeding and Reproduction
	fly, hang, hunt	Movement
	spray, treat, prevent, repel, fumigate, shield, control	Prevention and Control
	test	Detection and Testing
	breed	Reproduction
snake	slither, strike, coil, swallow	Movement
	bite, feed, swallow, hunt	Feeding and Prey
	live, shed, crawl	Habitat and Environment
	warn, coil	Defense and Interaction
	breed	Reproduction
spider	spin, trap, hunt, crawl, weave	Movement
	web, dangle	Web-related Activities
	catch, feed, bite	Feeding and Prey
	live	Habitat and Environment
	greet, show, talk	Interaction and Communication

Table 3: Animal Actions Category