

# Investigating Robustness of VLMs using LLMs

Kanishk Jain, Nathan Cormerais, Youry Macius

Université de Montréal

{kanishk.jain, nathan.cormerais, youry.macius}@umontreal.ca

## Abstract

We address the challenge of out-of-distribution (OOD) robustness in vision-language models (VLMs) for visual question-answering tasks. Current state-of-the-art VLMs demonstrate remarkable capabilities in answering visual questions, nearly matching human performance on benchmark datasets. However, these models exhibit significant weaknesses when encountering questions beyond their training distribution, often producing incorrect answers by leveraging spurious statistical correlations present in the training data. Our study aims to elucidate the failure modes of VLMs by establishing a two-agent chat framework involving a VLM and a large language model (LLM). Within this framework, we employ the LLM to generate OOD questions that the VLM fails to answer correctly. To facilitate this, we utilize few-shot prompting and proximal policy optimization techniques with the LLM. Furthermore, to systematically identify the failure modes of VLMs, we perform clustering on the incorrectly answered questions. This analysis allows us to categorize and understand the types of errors VLMs make. We present both quantitative and qualitative comparisons to underscore the effectiveness of our approach. The code is available at [github.com/kanji95/red-team-vlm](https://github.com/kanji95/red-team-vlm).

## 1 Introduction

Advancements in hardware capabilities and the availability of large-scale datasets have ushered in a new era of progress in the fields of computer vision and natural language processing. The emergence of Large Language Models (LLMs) has narrowed, and in some cases surpassed, human performance in tasks such as question answering, text summarization, sentiment analysis, and machine translation. Concurrently, Vision Language Models (VLMs) have gained popularity for various vision-language tasks including image recognition, visual question answering, image captioning, visual grounding,

and image-text retrieval. However, recent studies ([Chen et al., 2023](#)) have highlighted vulnerabilities in the robustness of VLMs when faced with changing distributions of images and text.

We study the robustness of VLMs to the changing data distribution for the task of visual question answering. Despite achieving strong performance on benchmark VQA datasets, these models struggle with questions out of their training distribution. A recent work ([Dancette et al., 2021](#)) found that models for VQA leverage shortcuts in the dataset in the form of biases and answer questions without utilizing the visual context in the image. We believe that large VLMs also suffer from the same problem but they come to light in face of out-of-distribution textual prompts.

In this project, we aim to investigate the robustness of Vision Language Models (VLMs) by utilizing Large Language Models (LLMs). Conventionally, assessing the robustness of VLMs involves human participants who formulate questions based on a given image. The VLMs are then evaluated based on their accuracy in responding to these questions. This method, while effective, is time-consuming and necessitates substantial human involvement.

To streamline this process, we propose the use of LLMs to replace human questioners. LLMs, trained on extensive web-scale textual corpora, possess a vast repository of textual knowledge. They are capable of generating diverse and complex questions that effectively challenge the VLMs' responses. However, since LLMs cannot process visual data directly, we bridge this gap by providing image captions that succinctly yet comprehensively describe the visual content. These captions, generated by VLMs, equip the LLMs to pose relevant questions about the images, mimicking the variability and complexity of human-generated queries. This method not only reduces reliance on human labor but also enhances the scalability of testing VLM robustness across diverse visual question-

answering scenarios.

We propose a two-agent chat framework between a VLM and an LLM, wherein the LLM generates questions and the VLM answers these questions. The goal is for the LLM to produce questions that VLM fails to answer correctly consistently. We use the accuracy of the VLM’s answers for providing feedback for subsequent question generation. We leverage techniques such as Few-Shot Prompting and Proximal Policy Optimization (PPO) to guide the question generation by LLM. Subsequently, we automatically identify the VLM’s failure modes by clustering the questions where VLM responded incorrectly. To summarize, our contributions are:

1. We established a two-agent chat framework between a Vision Language Model (VLM) and a Large Language Model (LLM), simulating the interaction between a VLM and a human user. In this setup, the LLM generates questions based on the image, which the VLM then answers.
2. We employ few-shot prompting and proximal policy optimization (PPO) to guide the question generation process by the LLM. Using PPO, we utilize the feedback from the VLM’s responses to inform the generation of subsequent questions.
3. We automate the discovery of VLMs’ failure modes by clustering questions that were answered incorrectly.
4. We provide both qualitative and quantitative results to demonstrate the effectiveness of our approach.

## 2 Related Work

**Large Language Models (LLMs)** are deep neural architectures with billions of parameters, pre-trained on vast amounts of data using masked language modeling objectives. These models typically consist of multiple layers of transformer-based encoder-decoder blocks and utilize self-attention mechanisms. Pre-trained LLMs (Wei et al., 2022; Jiang et al., 2023) demonstrate exceptional generalization capabilities across a variety of downstream tasks in both generative and discriminative settings. Recent developments (Schick et al., 2023) have equipped LLMs with tools like calculators, search engines, and more. Pertinent to our work is the study by (Perez et al., 2022), which involves using one LLM to “red team” or “jailbreak” another

target LLM. In our research, we plan to employ an LLM to identify the failure modes of a Vision Language Model.

**Vision Language Models (VLMs)** are multi-modal architectures that accept both image and textual prompts as input and generate responses to the query in text. These architectures are usually pre-trained on large-scale datasets (Sharma et al., 2018; Schuhmann et al., 2022), which results in strong generalization on various downstream tasks like image captioning (Xu et al., 2015), visual question answering (VQA) (Antol et al., 2015) and image-text retrieval (Luo et al., 2022). Depending on the pre-training objective, we have different VLMs, like CLIP (Radford et al., 2021) which is trained on contrastive learning objectives between image-text pairs. Following the success of (Radford et al., 2021), BLIP (Li et al., 2022) effectively utilizes the image-caption pairs by bootstrapping the captions. Recently proposed, InstructBLIP (Dai et al., 2023) and Llava (Liu et al., 2023) are VLMs that combine a vision encoder along with a Large Language Model (LLM) and perform instruction tuning on the VLM.

**Robustness in VLMs:** The growing application of VLMs has resulted in many recent works Chen et al. (2023); Dancette et al. (2021) that scrutinize the robustness of VLMs to changes in input image and text. Chen et al. (2023) indicates that real-world test samples often differ from the data used for adaptation during fine-tuning, highlighting the need for improved robustness. They find that full fine-tuning doesn’t consistently provide better robustness and that adapter-based methods can sometimes outperform them. In addition, Dancette et al. (2021) identifies that shortcut learning during the training phase reduces the robustness of VLMs as it enables the model to leverage shortcuts based on dataset biases while minimizing the learning objective. They introduce VQA-CounterExamples (VQA-CE) to detect shortcuts in model performance. Similarly, recent works like Gu et al. (2023); Zhou et al. (2023) propose techniques to improve the robustness of VLMs. Gu et al. (2023) proposes a more robust version of Prompt Learning (ProL) by integrating multiple-scale image features into the prompt. In addition, Zhou et al. (2023) proposes a framework combining knowledge distillation and data augmentation to enhance the robustness of VLMs. They introduce Discrete Adversarial Distillation (DAD) as a method to transfer knowledge from larger to

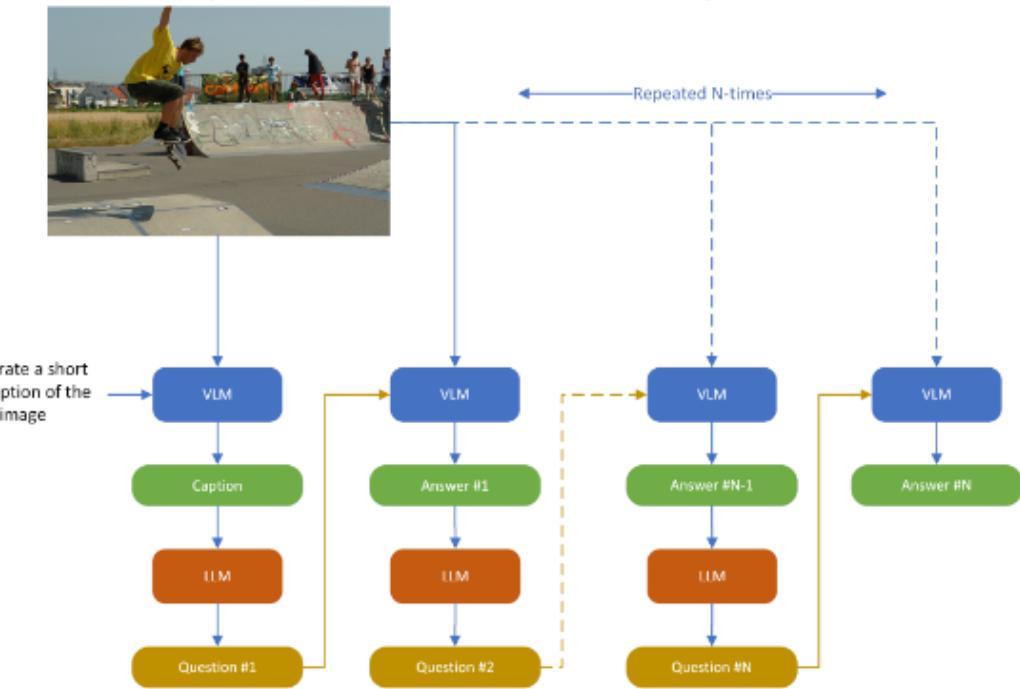


Figure 1: Our pipeline for two-Agent chat framework between a VLM and a LLM

smaller models, improving their performance on new data. In this work, we focus on the VQA task for evaluating the robustness of VLMs.

**OOD (Out-Of-Distribution) question generation** refers to the task of generating questions that are outside the distribution of the training data, but still relevant and meaningful. It can help improve the robustness and generalization of language models and question-answering systems. Recent works (Chen et al., 2023) have shed light on the vulnerability in the robustness of VLMs in changing the distribution of images and text. Extensive research(Jiang et al., 2024) has been dedicated to exploring OOD detection in the vision modality. Generating paraphrased versions of existing questions is one main approach to OOD question generation. Paraphrases(Zhou and Bhat, 2021) are texts that convey the same meaning while using different words or sentence structures. The generation of paraphrases is a longstanding problem for natural language learning. For example, the question How do I improve my English could be equivalently phrased as What is the best way to learn English. Paraphrasing can play an important role in language understanding tasks, such as question answering(Dong et al., 2017; Shuguang et al., 2017), machine translation (Ramtin Mehdizadeh et al., 2015; Brian and Matt, 2020a), and semantic parsing (Ruisheng et al., 2020). And it is also a

good way for data augmentation (Ashutosh et al., 2019; Silin et al., 2020). Given a sentence, paraphrase generation aims to create its paraphrases that can have a different wording or different structure from the original sentence, while preserving the original meaning.

### 3 Methods

Our goal is to automatically identify the failure modes of Vision Language Models (VLMs) using a Large Language Model (LLM). Given an image  $I$ , the LLM’s task is to generate questions  $Q$  pertaining to the image, while the VLM’s task is to provide answers  $A$  to these questions. The answers from the VLM are used as feedback to guide the LLM in generating subsequent questions. We employ two approaches to facilitate this process: (1) Few-Shot Prompting, and (2) Proximal Policy Optimization, both of which are detailed in Section 3.1. To utilize the VLM’s responses effectively, we must first verify the correctness of these answers to ensure they are guiding the LLM towards generating questions that the VLM consistently struggles with. To this end, we have developed a heuristic-based approach to validate the accuracy of the VLM’s answers, as described in Section 3.2. Once we have identified questions that the VLM cannot consistently answer, we perform clustering on these questions to automatically discern the failure modes of the VLMs,

and discuss the approach in Section 3.3.

### 3.1 Question Generation

We utilize Large Language Models (LLMs) to generate questions relevant to a given image. However, for the questions to be effective, they must be grounded in the image itself. Since LLMs only accept textual inputs, they face challenges in generating image-grounded questions directly. To address this, we provide visual context to the LLMs using captions generated by Vision Language Models (VLMs). Based on these captions, the LLM indirectly generates questions that are grounded in the image, which are then answered by the VLM. The responses to these questions are used to guide the generation of subsequent questions. We employ two specific approaches to direct this next-stage question generation, which we will describe in detail subsequently. The full pipeline for the framework is illustrated in Figure 1.

**Few Shot Prompting:** Here, we provide in-context examples to the LLM, sourced from an interactive chat framework explicitly set up between a human and a VLM. In this framework, the human asks a question to the VLM, and the subsequent question is generated based on the VLM’s answer. In this setting, no explicit feedback is provided to the LLM when generating the next question. As a result, this approach generates questions that do not take the VLM’s answer into account beyond the in-context input provided.

**Proximal Policy Optimization (PPO):** Here, we fine-tune the LLM using a reinforcement learning algorithm, specifically Proximal Policy Optimization (PPO). We calculate the reward based on the correctness of the VLM’s answers, details of which are provided in Section 3.2. In our PPO setup, one episode is equivalent to one round of question and answer, with the reward derived from the VLM’s response used to update the LLM via PPO. Given the large number of parameters in LLMs, we utilize parameter-efficient LoRA fine-tuning to optimize the LLM. This feedback loop consisting of a question, an answer, a reward, and a subsequent PPO update is performed over a fixed number of steps, specifically eight, in our study. Because explicit feedback is provided to the LLM during PPO, the questions it generates are more deeply rooted in the ongoing dialogue between the VLM and LLM, mimicking the way humans interact with VLMs.

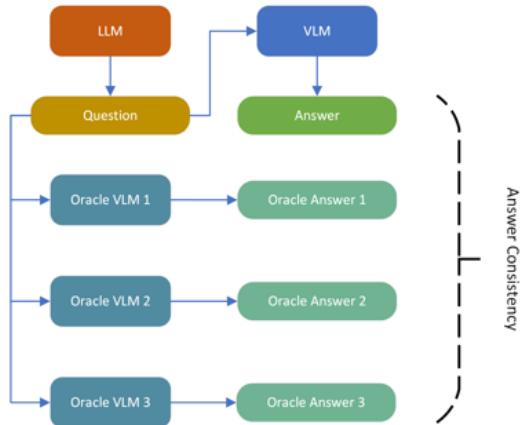


Figure 2: Our heuristic approach for verifying the correctness of VLM’s answer

### 3.2 Answer Verification

In this section, we describe our approach for assessing the correctness of answers provided by Vision Language Models (VLMs) within our proposed two-agent chat framework. The full approach is illustrated in Figure 2 In this setup, since the Large Language Model (LLM) generates questions on the fly, pre-established ground truth answers are unavailable. To address this, we employ a heuristic-based method to verify answer correctness. Specifically, we query multiple Oracle VLMs with the same image-question pair and collect their responses. To determine the correctness of the VLM under consideration, we calculate a semantic similarity score between its answer and those provided by the Oracle VLMs. The rationale behind using multiple Oracle VLMs is to mitigate the risk of relying on a potentially incorrect response from a single Oracle; it is unlikely that all Oracle VLMs would provide an incorrect answer simultaneously, making this a robust heuristic for verification. The reward for the LLM is computed based on the mean semantic similarity score with the Oracle VLMs' answers. If this mean score falls below a predetermined threshold, it indicates that the answer from the VLM under consideration significantly deviates from those of the Oracle VLMs and is likely incorrect. This result should yield a high reward, motivating the LLM to generate similar types of challenging questions that the VLM struggles to answer correctly.

### 3.3 Question Clustering

To evaluate the performance of VLMs, we want to organize the questions into different clusters.

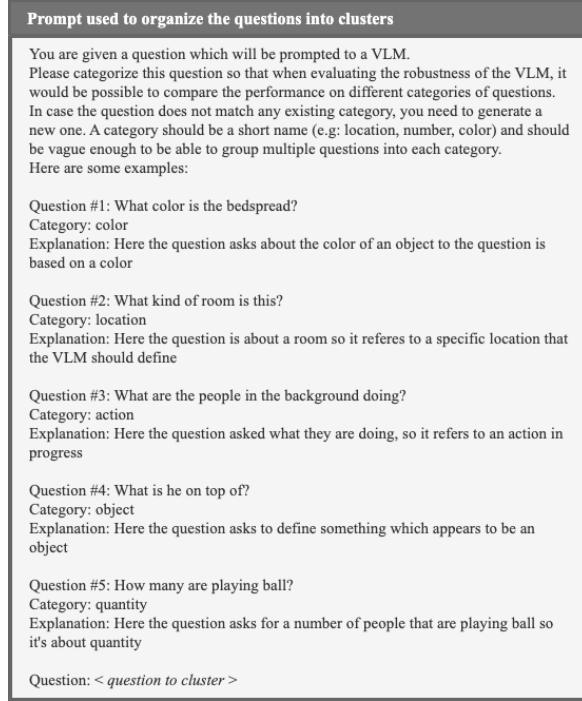


Figure 3: Prompt that was used for questions clustering

Indeed, this technique will help us identify the domains for which the evaluated VLMs perform well and the ones for which the VLMs show some incorrect answers. To filter the OOD (Out-Of-Distribution) questions into categories, we used *OpenHermes-2.5-Mistral-7B* which is a state of the art Mistral fine-tuned LLM. By exercising some prompt engineering on the model, we have built a prompt to organize a given question into a cluster (see Figure 3).

## 4 Experiments

### 4.1 Datasets

We utilized the VQA v2 dataset introduced by [Goyal et al. \(2017\)](#), which contains over 214,000 (question, image) pairs in the validation split. However, to investigate the robustness of VLMs, we use only a subset of images from the VQA v2 dataset within our multi-agent chat framework. It is important to note that our methodology is not restricted to a specific dataset and can be applied to any set of images.

### 4.2 Baselines

We choose Large Language and Vision Assistant (LLaVA-1.5) ([Liu et al., 2023](#)) as the VLM for which we investigate the robustness. Llava combines the capabilities of both language and vision

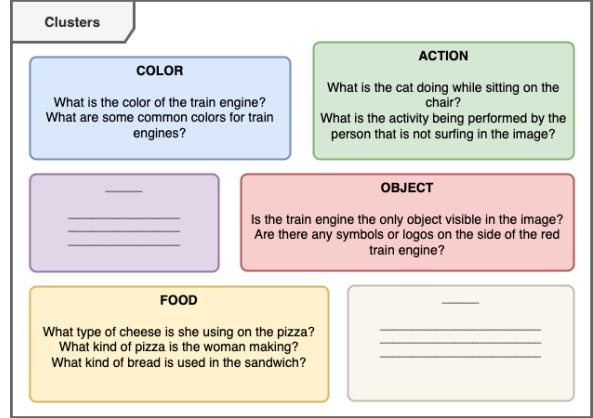


Figure 4: Samples of clusters and questions obtained from the clustering technique

models and can process both images and text, allowing for a wide range of applications and interactions. We use *OpenHermes 2.5 Mistral 7B* as the LLM in our study. For oracle VLMs we use *Internlm-xcomposer2-vl-7b*, *Deepseek-vl-7b-chat* and *Llava-v1.6-mistral-7b-hf*. All the models including VLMs and LLM have 7 billion parameters and are open-sourced and downloaded from huggingface. Furthermore, the Oracle VLMs have better performance than Llava-1.5 (the VLM for which we investigate robustness) on the [OpenVLM leaderboard](#) on huggingface. We also experiment with older VLMs like ViLT ([Wonjae et al., 2021](#)) which is a transformer-based architecture that processes visual and textual inputs in a unified manner, enabling the model to understand and generate both images and text. ViLT processes visual inputs without relying on convolutional neural networks (CNNs), making it much faster than previous vision-and-language models. We investigate the robustness of both ViLT and Llava on the most trivial type of OOD question distribution, ie: paraphrased questions. ViLT was fine-tuned on the VQA dataset and is evaluated as a supervised baseline and Llava is evaluated in a zero-shot manner.

### 4.3 Evaluation Methods

To evaluate the robustness of the models, we use the answer accuracy, ie: *String-Matching* as the evaluation metric. However, relying solely on string matching to determine if a predicted answer aligns with the ground truth can lead to inaccuracies. This is because string matching might not recognize correct answers that are semantically similar but not identical in form. For instance, if the model responds with "one" to a counting question, but

**LLM:** Are there any other animals visible in the field?  
**Llava 1.5:** No, there are no other animals visible in the field.  
**Internlm-xcomposer2:** No, there are no other animals visible in the field besides the zebra.  
**DeepSeek Answer:** Yes, there is a bird visible in the field near the zebra.  
**Llava 1.6:** Yes, there is a bird visible in the field behind the zebra.



**LLM:** What is the color of the zebra's stripes?  
**Llava 1.5:** The zebra's stripes are black and white.  
**Internlm-xcomposer2:** The zebra has black and white stripes  
**DeepSeek Answer:** The zebra's stripes are brown and white.  
**Llava 1.6:** The zebra's stripes are black and white.

A zebra standing in a grassy field.

#### QUALITATIVE RESULTS FOR FEW-SHOT PROMPTING

**LLM:** Is the cat wearing any accessories or collars on its neck?  
**Llava 1.5:** The cat is wearing a collar on its neck  
**Internlm-xcomposer2:** No, the cat is not wearing any collars on its neck.  
**DeepSeek Answer:** The cat is wearing a collar in the image  
**Llava 1.6:** The cat in the image does not appear to wear any collar on its neck

**LLM:** What is the color of the cat's collar?  
**Llava 1.5:** The cat's collar is white.  
**Internlm-xcomposer2:** I cannot see the cat's collar in the image.  
**DeepSeek Answer:** The cat is not wearing a collar.  
**Llava 1.6:** The cat in the image is not wearing a collar.



A cat is sitting on a chair next to a person.

#### QUALITATIVE RESULTS FOR PPO

Figure 5: Correct answers are in green color and incorrect ones in red color

the ground truth is represented by the numeral "1," string matching would unfairly mark a correct prediction as incorrect. In addition, we introduced a *heuristic-metric*, in which we consider partial matches but ignore semantics. Indeed, using the heuristic matching method, we can evaluate if the answer contains the ground truth, therefore, if the answer is "close enough" to the ground truth. To address these limitations, we turn to LLMs known for their advanced language comprehension. We used an LLM-based metric called *LAVE* that was introduced in Mañas et al. (2024). The purpose is to use an LLM to evaluate the accuracy of the answer given to a VLM compared to reference answers. This method outputs a score in the range  $[0, 1]$  where 0 means that the candidate answer is far from the truth and 1 means that the candidate answer is the most likely to be truth. In the paper, the authors show that the *LAVE* score is more aligned with human judgement and also provides an explanation of the given rating.

#### 4.4 Experimental Results

For the question generation, we use 8 rounds of interaction between the VLM and LLM for question answering. For applying PPO update we use AdamW optimizer. The threshold used for computing semantic similarity of Oracle answers with Llava1.5's answers is 0.8. As mentioned in section 3.3, we used the *OpenHermes-2.5-Mistral-7B* model and an approach based on prompt-engineering to group the OOD questions into clusters.

#### 4.5 Results and Analysis

We provide a qualitative comparison between the Few-Shot and PPO-based approaches for question generation in Figure 5. The few-shot approach, does not incorporate feedback from the VLM's answers and generates questions that are not deeply rooted in the conversation between the VLM and LLM. Consequently, the questions do not require an in-depth understanding of the image. In contrast, the questions generated using the PPO approach

	<b>Model</b>	<b>String matching</b>	<b>Heuristic accuracy</b>	<b>LAVE score</b>
<b>VQA v.2.0</b>	ViLT (FINE-TUNED)	70.0%	72.0%	83.25%
	LLAVA (ZERO-SHOT)	61.0%	64.0%	74.50%
<b>OOD questions</b>	ViLT (FINE-TUNED)	54.69%	56.28%	72.65%
	LLAVA (ZERO-SHOT)	54.36%	58.72%	74.06%

Table 1: Quantitative results on a subset of images from VQA v.2 dataset, using the ViLT model fine-tuned on the VQA v.2.0 dataset, and the LLava model

are more closely tied to the dialogue. For instance, the first question pertains to the collar on a cat’s neck, with the follow-up question based on the VLM’s answer, specifically regarding the color of the collar. In this scenario, LLava-1.5 provided an incorrect response to the first question, and interestingly, one of the oracle models also responded incorrectly. However, in the follow-up question, which was based on the incorrect initial answer, all the oracle models provided the correct response, while LLava-1.5 continued to answer incorrectly. This illustrates that the PPO setting enables the generation of questions that specifically target the weaknesses of the VLM under consideration. For questions clustering, as mentioned in section 3.3, samples of clusters containing OOD questions are illustrated in Figure 4. Even though some of the categories were defined in the few-shot examples, others were defined by the LLM.

## 5 Conclusion

In conclusion, our work aims to investigate the robustness of VLMs by leveraging LLMs. We find that older vision-language models (VLMs) like ViLT, struggle with OOD questions in the form of paraphrased questions, leading to significant performance drops. In contrast, newer models such as LLava show improved robustness against such variations, maintaining consistent performance. Therefore, we leverage LLava as the VLM for the subsequent study of multi-agent chat setup to discover its failure modes automatically. Our utilization of LLM reveals their potential in automatically discovering the failure modes of state-of-the-art VLMs. By generating challenging OOD questions, LLMs help expose the weaknesses that might not be evident through standard testing. Additionally, our comparative analysis between few-shot prompting and Proximal Policy Optimization (PPO) methods has shown that PPO offers a more effective mechanism for training LLMs. Unlike few-shot learning, which provides limited adaptability, PPO leverages

direct feedback from VLM responses to optimize question generation, aiming specifically to challenge the VLMs effectively. Overall, our approach enhances the methodology for testing and understanding the failure modes of VLMs. By integrating techniques such as clustering of incorrectly answered questions, we provide an automated structured way to categorize and analyze the failure modes, and paving the way for the development of more fault-tolerant vision-language models.

## References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Kumar Ashutosh, Bhattacharya Satwik, Bhandari Manik, and Talukdar Partha. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, page 3609–3619.
- Thompson Brian and Post Matt. 2020a. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 90–121.
- Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. 2023. Benchmarking robustness of adaptation methods on pre-trained vision-language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual

- question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1574–1583.
- Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 875–886.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jindong Gu, Ahmad Beirami, Xuezhi Wang, Alex Beutel, Philip Torr, and Yao Qin. 2023. Towards robust prompts on vision-language models. *arXiv preprint arXiv:2304.08479*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillem Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. **Mistral 7b**.
- Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. 2024. **Negative label guided ood detection with pretrained vision-language models**. In *International Conference on Learning Representations*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.
- Ziyang Luo, Yadong Xi, Rongsheng Zhang, GongZheng Li, Zeng Zhao, and Jing Ma. 2022. **Conditioned masked language and image modeling for image-text dense retrieval**. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 130–140, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. **Improving automatic vqa evaluation using large language models**.
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. **Red teaming language models with language models**.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Seraj Ramtin Mehdizadeh, Siahbani Maryam, and Sarkar Anoop. 2015. Improving statistical machine translation with a multilingual paraphrase database. In *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, page 1379–1390.
- Cao Ruisheng, Zhu Su, Yang Chenyu, Liu Chen, Ma Rao, Zhao Yanbin, Chen Lu, and Yu Kai. 2020. Unsupervised dual paraphrasing for two-stage semantic parsing. In *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 6806–6817.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. **Toolformer: Language models can teach themselves to use tools**.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.
- Zhu Shuguang, Cheng Xiang, Su Sen, and Lang Shuang. 2017. Knowledge-based question answering by jointly generating, copying, and paraphrasing. In *In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, page 2439–2442.
- Gao Silin, Zhang Yichi, Ou Zhijian, and Yu Zhou. 2020. Paraphrase augmented task-oriented dialog generation. In *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 639–649.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. **Finetuned language models are zero-shot learners**. In *International Conference on Learning Representations*.
- Kim Wonjae, Son Bokyung, and Kim Ildoo. 2021. **Vilt: Vision-and-language transformer without convolution or region supervision**.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Andy Zhou, Jindong Wang, Yu-Xiong Wang, and Haohan Wang. 2023. *Distilling out-of-distribution robustness from vision-language foundation models*. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086.

## A Contributions

This section describes the work contribution of each team member,

- **Kanishk Jain:** Worked on idea formulation for the project. Contributed to project proposal, midway report, final presentation and final report. Worked on the setup of multi-agent chat framework between VLM and LLM, question generation using few-shot and PPO based approaches and answer verification specified in the method section.
- **Nathan Cormerais:** Contributed to the project proposal, midway report, final presentation, final report and the code for evaluating the performance of models such as ViLT or LLaVA. Also contributed to the Out-Of-Distribution (OOD) question clustering using the *OpenHermes-2.5-Mistral-7B* model and an approach based on prompt engineering.
- **Youry Macius:** Discussed about the idea (subject) formulation for the project. Contributed to project proposal, midway report, final presentation and final report. Worked on Google colab to run and fine-tune the code for evaluating the performance of the Out-Of-Distribution (OOD) question clustering using the *OpenHermes-2.5-Mistral-7B* model and an approach based on prompt engineering.