

Evaluating Multimodal Large Language Models across Distribution Shifts and Augmentations

*Aayush Atul Verma *Amir Saeidi *Shamanthak Hegde *Ajay Therala
*Fenil Denish Bardoliya *Nagaraju Machavarapu *Shri Ajay Kumar Ravindhiran *Srija Malyala
*Agneet Chatterjee Yezhou Yang Chitta Baral

Arizona State University

Abstract

Foundational models such as Multimodal Large Language Models (MLLMs) with their ability to interpret images and generate intricate responses has led to their widespread adoption across multiple computer vision and natural language processing tasks. However, they suffer from hallucinations and struggle to reason over complex reasoning tasks. In this work, we evaluate the performance of MLLMs across multiple multimodal augmentations and evaluate their performance in out-of-distribution settings. We benchmark 3 models, across 2 vision-language datasets, VQA_{v2} and CLEVR, and assess their performance across adversarial transformations in both the vision and language modalities. We introduce image perturbations using various augmentations, including noise addition, blurring, and median filtering and generate adversarial questions which contain conjunctions, disjunctions and negations. Additionally, we conduct a detailed fine-grained analysis to assess the model's performance on particular question categories, such as those related to shape and color, across images featuring identical or varying objects. Our findings indicate a notable decrease in the performance of current MLLMs for synthetic images, with a gradual decline observed across both vision and language augmentations. Specifically, Gaussian Noise Addition emerges as the most detrimental augmentation, and we observe a significant drop in performance with complex questions containing multiple connectives. In these times of rapid development and deployment of MLLMs in real-world settings, we believe our findings are a first step towards benchmarking the robustness and out-of-distribution behavior of such models.

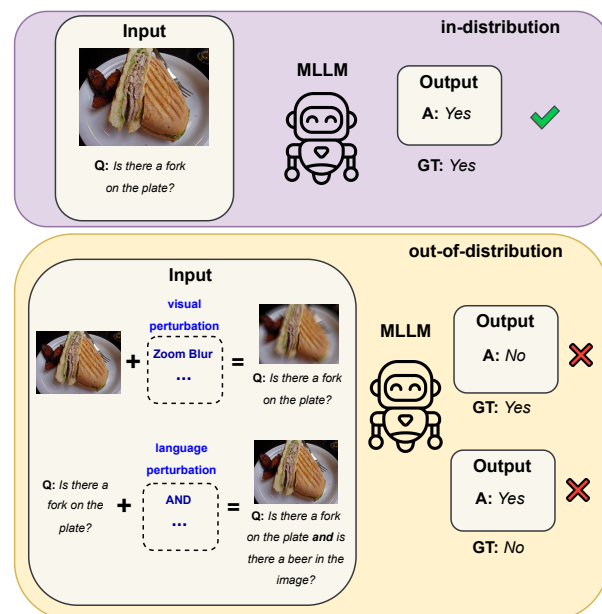


Figure 1. In this work, we evaluate the robustness of Multimodal Large Language Models on out-of-distribution settings. We create adversarial attacks both in the vision and language modality and find that existing models show a steady decline in performance under these scenarios.

1. Introduction

With the continuous expansion of data availability and computational capabilities [27], large language models such as GPT-4 [24], LLaMA-2 [27], Mistral [14], and PaLM2 [2] have demonstrated exceptional performance in both natural language understanding (NLU) and generation (NLG). The emergence of multimodal large language models signifies a shift towards integrating LLMs with additional modalities such as images and audio, enabling them to produce appropriate outputs for diverse input types [35]. Several studies

*Equal contribution. Correspondence to agneet@asu.edu

have explored the fusion of LLMs with visual understanding, exemplified by models like GPT-4V [32], Gemini [26], Flamingo [1], and Qwen [3]. The impressive achievements of Multimodal Large Language Models (MLLMs) in multiple vision tasks has inspired their application in complex multimodal challenges such as robotics [18] and common-sense reasoning [31].

Despite their impressive performance across various tasks, MLLMs face challenges such as hallucinations and reasoning over complex text-image pairs. MLLM hallucinations have been well studied in LURE [36] and Woodpecker [33], and their reasoning abilities are benchmarked in MM-VET [34], MME [11] and MMBench [23]. However, their evaluation and performance under distribution shifts and adversarial attacks remain under-explored.

In this paper, we explore this research question and investigate state-of-the-art MLLMs and benchmark their performance on out-of-distribution settings. We conduct our experiments on 2 widely used vision-language datasets: VQAv2 [13] and CLEVR [15], which consist of real-world and synthetically generated images, respectively. We generate a total of 10 adversarial samples for a given instance, both in the vision and language modality to conduct a holistic study, across multiple out-of-distribution scenarios. Furthermore, we examine how model performance varies based on factors such as the number of objects in an image, the nature of questions (numerical, color-related, binary, shape-related), and the presence of multiple object types within an image. An overview of our workflow is presented in Figure 1.

Our investigation reveals several interesting findings. *First*, we find that MLLMs perform significantly worse on synthetic images in comparison to real-world images. *Second*, across both real-world and synthetic images, model performance declines with the addition of image augmentations, with the Gaussian Low Pass Filter causing the largest drop. *Third*, for questions with adversarial augmentations, models consistently provide affirmative ('Yes') responses. *Fourth*, we perform a detailed study and categorize model responses based on number of objects in an image and by question types; and reveal observations such as a significant decrease in performance on 'color'-related questions with an increasing number of objects in an image. To summarize, our contributions are as follows :

- We study 3 Multimodal Large Language Models, on 2 datasets, VQAv2 and CLEVR, and develop an evaluation benchmark that creates adversarial samples through 6 different image and 4 reasoning-based language augmentations.
- Our evaluations reveal that incorporating these augmentations leads to a decrease in model performance on both natural and synthetic images, with synthetic images exhibiting lower performance overall.

- Additionally, we find that a) the Gaussian Low Pass Filter augmentation leads to the largest drop in performance, b) models struggle when an image contains objects of multiple kinds, c) color-based questions in complex scenes leads to multiple errors and d) models tend to answer in the affirmative when presented with complex logical connectives.

2. Related Works

Large Language Models (LLMs) have demonstrated remarkable proficiency across various tasks, showcasing their adaptability to solve diverse problems across different modalities [16, 30]. This has led to the emergence of Multimodal Large Language Models (MLLMs), specifically designed to tackle challenges with multi-modal inputs and exhibit promising performance, particularly in zero-shot generalization scenarios [6]. Broadly, MLLMs can be categorized based on their modeling into three main types [36]. The first category involves integrating a frozen vision encoder with a large language model, utilizing cross-attention mechanisms to handle cross-modalities with Flamingo [1] being a pioneering work in this regard. In the second category, features extracted from the vision encoder are connected to the pre-trained model through linear layers. PaLM-E [10], which combines PaLM [7] with a vision encoder, exemplifies this approach. Additionally, this method has been adopted in various works, including LLaVA [22] and Shikra [5]. However, a limitation of this approach lies in the creation of the visual sequence length, which can impact the performance of MLLMs. To address this limitation, the third category proposes employing transformer models to reduce the sequence length of visual features efficiently. BLIP-2 [17], inspired by DETR [4], stands as a significant contribution in this category.

Despite their impressive performance in tackling various modalities, MLLMs are often not robust and can produce hallucinations with slight changes in images or text. Recent efforts [21, 29] have focused on addressing this issue and proposing mitigating methods. For example, LVT-Instruction [21] tackles hallucinations by reducing the length of textual instructions. However, this approach comes with the trade-off of limiting the learning capabilities of MLLMs. In contrast, LLaVA RLHF [25] employs synthetic data to modify the robustness of the model. This involves training a reward model using human-annotated data to generate signals, aligning a supervised fine-tuned model with human preferences. Yet, this method requires a significant amount of annotated data, posing a limitation. Meanwhile, HACL [36] introduces hallucination captions as complex negative samples in contrastive learning. Additionally, some studies focus on detecting object hallucinations within different domains of MLLMs that have a high effect on the robustness of the model. For instance, POPE [19] addresses

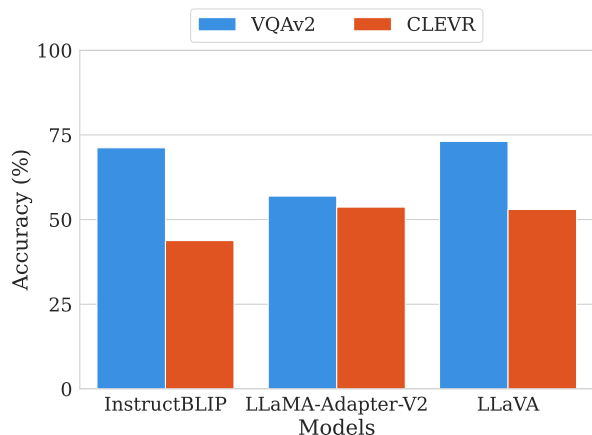


Figure 2. Baseline Performance of InstructBLIP, LLaMA-Adapter-V2, and LLaVA models on the VQAv2 and CLEVR dataset.

this by transforming hallucinations into binary classification problems, thereby informing the model about the presence of objects in an image.

Several studies have proposed benchmarks to assess the MLLMs across various reasoning tasks. MM-VET [34] demonstrated MLLMs’ proficiency in handling complex tasks yet highlighted persistent challenges such as their limited generalization capabilities. To address this, they introduced new benchmarks aimed at enhancing MLLMs’ ability to generalize. Similarly, MME [11] introduced a comprehensive benchmark evaluating MLLMs’ performance across 14 subtasks related to perception and cognition. Another noteworthy study, MMBench [23], comprises two key components. Firstly, it offers a curated dataset that surpasses existing benchmarks in terms of question variety and evaluation abilities. Secondly, it presents a cyclic evaluation strategy integrating ChatGPT to assess MLLMs’ performance. Despite these efforts revealing areas of MLLMs’ underperformance across diverse tasks, none have extensively evaluated the robustness of foundational MLLMs. A contemporary work has been proposed that explores image-based attacks on MLLMs [8], however they focus majorly on gradient-based attacks compared to our pixel level attacks and do not take into consideration the impact of language level perturbations.

While previous studies have explored reasoning and hallucinations in MLLMs to show their robustness, the impact of distribution shift and adversarial perturbation on these models remains underexplored. We study this critical problem and the evaluate the generalization patterns associated with multiple models, identifying key points of failure for safer deployment of these models in real-world settings.

3. Experiments and Analysis

In this section, we delve into the comprehensive evaluation of the robustness of foundation models through a series of carefully designed experiments. The experiments primarily focus on two distinct Visual Question Answering (VQA) datasets, VQAv2 [13], a real-world dataset, and CLEVR [15], a synthetic dataset. VQAv2 is based on the images from the COCO dataset [20], which has rich and complex visual content that is highly relevant for the models to answer the questions. CLEVR serves as a diagnostic tool for assessing various visual reasoning abilities with minimal biases. Its detailed annotations provide valuable insights into the specific types of reasoning required for each question, enabling a thorough analysis of model strengths and weaknesses. Moreover, CLEVR images serve as an *out-of-distribution* dataset, as most MLLMs are pre-trained/fine-tuned on real-world images such as COCO.

We begin by establishing baselines using multiple MLLMs such as InstructBLIP [9], LLaMA-Adapter-V2 [12], and LLaVA [22] on the original dataset. Subsequently, we introduce perturbations in both vision and language modalities, aiming to analyze the behavior of these models under various changes. For evaluation, we use the Zephyr-7B [28] language model due to its impressive Natural Language Understanding and Question-answering capabilities. We prompt the Zephyr-7B model to act as an evaluator in our experimental setup and present it with a ground truth answer and predicted answer pair. We then task it with responding to whether the two answers are consistent with each other or not. To maintain consistency and mitigate randomness in the inferences, we use a temperature setting of zero. The insights gained through these experiments help us better understand the models’ adaptability and robustness when faced with multimodal distribution shifts and augmentations.

3.1. Baseline

For the baseline analysis, we randomly sample 15K images from both the VQAv2 and CLEVR datasets, totaling 30K images. The sampling was conducted using a uniform distribution method, ensuring an equal likelihood of selecting any image from the datasets. To ensure diversity and richness in the evaluation, we pair each image with five distinct questions, amounting to approximately 75K Question-Answer (QA) pairs per dataset. This approach enables a robust understanding of the models’ performance across a varied set of scenarios. An examination of the performance of our multimodal foundation models on the two datasets can be seen in Figure. 2. This analysis serves as a crucial reference point for evaluating the subsequent perturbations introduced in the following subsections.

We observed that all models tend to perform noticeably better on the real-world VQAv2 dataset compared to the

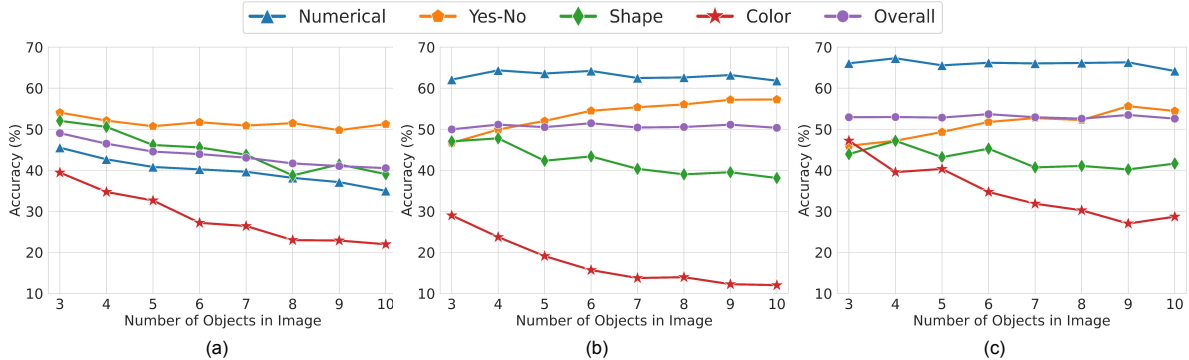


Figure 3. Comprehensive Evaluation of (a) InstructBLIP, (b) LLaMA-Adapter-V2, and (c) LLaVA on CLEVR: Variation in performance (accuracy) with increasing scene complexity, defined by the number of objects in the image.

synthetic CLEVR dataset. This can be attributed to multiple factors: As previously mentioned, images from CLEVR are inherently OOD in comparison to the images from VQAv2; VQAv2 consists of diverse and natural real-world scenarios with various objects, scenes and lighting conditions that require commonsense reasoning and contextual understanding. In contrast, the questions in CLEVR focus on spatial reasoning and demand an understanding of object properties such as color, shape, material, and size. Another interesting observation is that although CLEVR is composed of simple geometric shapes and objects, reasoning by foundation models’ is not on par with the images from COCO, which are composed of complex real-world scenes and objects. These results show that current MLLMs do *not* generalize well to out-of-distribution settings. Additionally, there is a lack of grounding towards fine-grained object properties such as material and texture, along with shortcomings in understanding geometrical properties such as spatial relationships.

Next, we split the CLEVR dataset based on the type of expected answers into 4 categories: shape, color, yes-no, and numerical, and observed the variation in performance with varying scene complexity and object distribution.

Scene Complexity Analysis: We define increasing scene complexity as an increase in the number of distinct objects in the image. These categorizations are enabled by the well-designed CLEVR dataset, which provides ground truth annotations. We present our analysis in Figure 3. We find that InstructBLIP shows a large degradation in performance as scene complexity increases, whereas LLaMA-Adapter-V2 and LLaVA exhibit more consistent results. Questions expecting a color and shape-type answer showcase a steep decline across all models with an increase in the number of objects. On the other hand, other questions maintain a performance similar to the entire dataset, with LLaMA-Adapter-V2 and LLaVA performing the best on numerical-

type questions. One potential explanation for the observed shortcomings related to shape and color could stem from the lack of diversity in the datasets used to train these models. For instance, it’s possible that real-world datasets don’t contain sufficient representations of certain features, such as “cubes” or objects colored “cyan”.

Object Distribution Analysis: Next, we split the CLEVR dataset into 2 subsets based on the type of objects present in the image. We define the Intra subset with images containing objects of the same type, specifically, identical shapes. The Inter subset comprises images with multiple object types. We summarise our results in Table 1. The models consistently perform better on images containing objects with the same shape, indicating that as the diversity in object shapes increases, the performance degrades.

Model	Type	Question Type Split				
		Shape	Numerical	Color	Yes-No	Overall
InstructBLIP	Inter	44.07	39.87	28.34	51.63	43.75
	Intra	71.69	43.47	34.75	45.81	46.29
LLaMA-Adapter-V2	Inter	41.79	63.03	17.06	53.74	50.65
	Intra	65.40	63.32	31.01	47.74	53.25
LLaVA	Inter	41.77	65.56	34.89	51.61	52.71
	Intra	52.2	67.79	42.78	53.7	53.37

Table 1. Performance Comparison of InstructBLIP, LLaMA-Adapter-V2, and LLaVA (Accuracy) on the CLEVR Dataset. Results are Categorized Across Object Distribution: Intra (Images with Same Object Shapes) and Inter (Multiple Object Shapes), Split by Question Types.

3.2. Visual Perturbation

Building upon the baseline analysis, we investigate the impact of introducing visual perturbations on these models. Using various image augmentation techniques, we assess the models’ performance in handling changes in the visual aspects of the input data. We consider 6 different image augmentation techniques: 1) Median Filter, 2) Gaussian

Augmentations	InstructBLIP				LLaMA-Adapter-V2				LLaVA			
	VQAv2		CLEVR		VQAv2		CLEVR		VQAv2		CLEVR	
	w/o Aug	w/ Aug	w/o Aug	w/ Aug	w/o Aug	w/ Aug	w/o Aug	w/ Aug	w/o Aug	w/ Aug	w/o Aug	w/ Aug
Median Filter	71.09%	57.45%	43.64%	40.39%	56.78%	55.11%	49.99%	49.52%	72.83%	58.07%	53.34%	48.97%
Gaussian Noise Addition	71.1%	63.16%	44.14%	40.84%	57.16%	59.84%	50.16%	50.24%	73.17%	65.18%	52.49%	48.56%
Multiplicative Noise Addition	71.33%	63.72%	43.55%	41.06%	57.01%	59.52%	50.9%	50.79%	73.02%	65.03%	53.12%	49.05%
Gaussian LowPass Filter	71%	55%	43.93%	39.16%	56.96%	53.05%	51.18%	49.61%	73.26%	54.13%	53.46%	46.71%
Zoom Blur	70.96%	55.12%	43.44%	40.6%	56.66%	54.1%	50.27%	50.18%	72.98%	57.33%	52.47%	47.41%
ISO Noise Addition	71.2%	63.6%	44.08%	42.02%	57.04%	59.5%	51.52%	51.53%	73.22%	65.03%	52.94%	49.3%

Table 2. The Impact of 6 Image Augmentations on Model Performance (accuracy) over VQAv2 and CLEVR Datasets: Drop in performance observed for every image augmentation compared to baseline performance, indicating a lack of robustness when exposed to out-of-distribution settings.

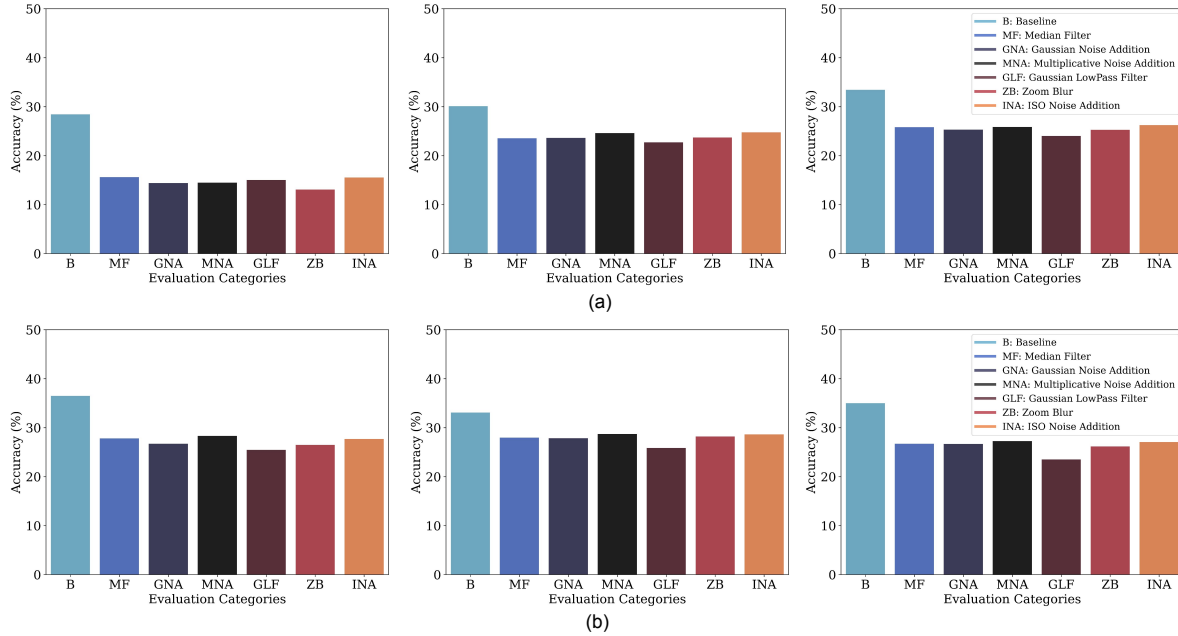


Figure 4. The Impact of Six Image Augmentations on Model Performance (accuracy) over CLEVR Dataset: Variation observed in (a) color-based questions, and (b) shape-based questions. (**B**: Baseline, **MF**: Median Filter, **GNA**: Gaussian Noise Addition, **MNA**: Multiplicative Noise Addition, **GLF**: Gaussian Low-Pass Filter, **ZB**: Zoom Blur, **INA**: ISO Noise Addition)

Noise Addition, 3) Multiplicative Noise Addition, 4) Gaussian Low Pass Filter, 5) Zoom Blur, and 6) ISO Noise Addition. We choose these augmentations as they entail multiple aspects: image smoothing (median filtering), sampling from varying distributions (Gaussian, Multiplicative, and ISO Noises), and change in the frequency domain (Gaussian Low Pass).

We randomly select 30K questions out of 75K in the datasets and perform augmentation on their corresponding images. Using these specific image and question pairs, we evaluate the performance of the models with and without performing augmentations whose accuracies are shown in Table 2. It is apparent from the table that all models exhibit a decline in performance when subjected to visual perturbations across the VQAv2 and CLEVR datasets. This drop underscores the challenge of maintaining robustness in multimodal foundation models when faced with alterations in the

visual aspects of input data. However, LLaMA-Adapter-V2 consistently demonstrates greater resilience to visual perturbations compared to the other models across both datasets. This suggests a higher level of adaptability to multiple variations in image characteristics.

Additionally, certain augmentations have a more pronounced impact on model performance. For instance, Gaussian, multiplicative, and ISO noise addition appear to have a lesser impact on all three models compared to other augmentation techniques. This observation indicates variations in sensitivity to different types of visual alterations among the models and that *noise*, across different distributions, does not impact model performance significantly. It's worth noting that in some cases, especially with VQAv2, certain augmentations result in significant performance drops, exceeding 20%. Similarly, in the CLEVR dataset, performance dips of over 10% are observed for some augmenta-

QT	QF	Question	Answer
Original	Q_1	Are there any gray balls to the left of the yellow block?	No
	Q_2	Is there an object that has the same material as the yellow sphere?	Yes
NOT	$\neg Q_1$	Are there not any gray balls to the left of the yellow block?	Yes
	$\neg Q_2$	Is there not an object that has the same material as the yellow sphere?	No
AND	$Q_1 \wedge Q_2$	Are there any gray balls to the left of the yellow block and is there an object that has the same material as the yellow sphere?	No
OR	$Q_1 \vee Q_2$	Are there any gray balls to the left of the yellow block or is there an object that has the same material as the yellow sphere?	Yes
COMPLEX	$Q_1 \vee \neg Q_2$	Are there any gray balls to the left of the yellow block or is there not an object that has the same material as the yellow sphere?	No
	$Q_1 \wedge \neg Q_2$	Are there any gray balls to the left of the yellow block and is there not an object that has the same material as the yellow sphere?	No
	$\neg Q_1 \vee Q_2$	Are there not any gray balls to the left of the yellow block or is there an object that has the same material as the yellow sphere?	Yes
	$\neg Q_1 \wedge Q_2$	Are there not any gray balls to the left of the yellow block and is there an object that has the same material as the yellow sphere?	Yes

Table 3. Illustration of question composition using conjunction and disjunction operations, used for experiments involving language perturbations. (QT: Question Type, QF: Question Formula)

Augmentation	InstructBLIP				LLaMA-Adapter-V2				LLaVA			
	VQAv2		CLEVR		VQAv2		CLEVR		VQAv2		CLEVR	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Original	80.9%	81.43%	49.31%	53.14%	71.62%	40.69%	53.71%	53.83%	42.76%	57.84%	65.76%	21.88%
NOT	43.71%	57.96%	40.61%	53.41%	66.63%	10.74%	65.54%	23.75%	39.71%	22.24%	64.03%	6.09%
AND	77.78%	68.46%	61.95%	28.48%	69.24%	28.12%	57.74%	41.84%	81.47%	77.24%	66.92%	7.23%
OR	71.82%	57.32%	72.07%	48.88%	81.73%	75.76%	75.95%	66.73%	72.24%	54.76%	75.44%	56.07%
COMPLEX	57.63%	57.32%	57.36%	36.61%	66.67%	2.95%	44.98%	42.98%	69.97%	59.41%	66.06%	4.63%

Table 4. Impact of Four Language Augmentations on F1-scores of ‘Yes’ and ‘No’ classes over VQAv2 and CLEVR Datasets with higher scores indicated in bold: An increased disparity between the two classes leading to reduced fairness in the models as complex connectives are introduced. We notice an increased bias with higher scores for the ‘Yes’ class, indicating an inclination towards affirmative responses.

tions. As mentioned in the previous section, we can observe that the performance of the models on the CLEVR dataset is worse compared to its on the VQAv2 dataset due to the CLEVR dataset having images that are synthetic in nature and require a more in-depth understanding of color, shape and spatial reasoning compared VQAv2.

To gain deeper insights into the varying degrees of sensitivity observed among different models and augmentations, we perform experiments focused on questions related to the physical aspects of objects in images from the CLEVR dataset - color and shape, as summarized in Figure 4.

Color-based Questions All models show a decline in performance on color-based questions. We hypothesize that the diversity of colors present in the CLEVR dataset leads to the models generating incorrect responses. InstructBLIP exhibits the most substantial decline for color-related questions across the six augmentation methods, indicating a higher sensitivity to color-centric changes. On the other hand, LLaMA-Adapter-V2 and LLaVA display a more resilient performance.

Shape-based Questions Unlike color-based questions, all three models display a similar dip in performance, with LLaMA-Adapter-V2 being the best-performing model.

These results demonstrate the differing degrees of robustness among the three models in addressing color-related questions while showing a similar decline when dealing with shape-related questions. Notably, among all the augmentations applied, the Gaussian Low Pass Filter consistently proved to be the most disruptive, noticeably affecting the models’ understanding of both color and shape properties in the images.

3.3. Language Perturbation

In this subsection, we turn our attention to the language modality to investigate how the visual understanding of these foundation models is impacted when faced with questions that are logical combinations of their original questions. Our objective is to analyze how these models handle questions that entail combinations of logical operations such as NOT, OR, AND, and COMPLEX, thereby introducing complexity to the question. For this, we curate questions




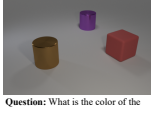
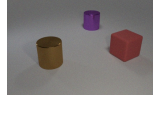
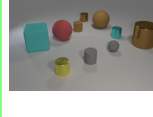
Baseline	Median Filter	Gaussian Noise Addition	Multiplicative Noise Addition	Gaussian Low Pass Filter	Zoom Blur	ISO Noise Addition
 Question: What color is the truck? GE: Red and White Pred: White and Red	 Pred: White and Blue	 Pred: Red and White	 Pred: Red	 Pred: White	 Pred: White	 Pred: White and Red
 Question: How many dogs are there? GE: 1 Pred: 1	 Pred: 1	 Pred: 1	 Pred: 1	 Pred: 2	 Pred: 0	 Pred: 1
 Question: What is the color of the rubber cube? GE: Red Pred: Red	 Pred: Red	 Pred: Red	 Pred: Red	 Pred: Purple	 Pred: Red	 Pred: Red
 Question: Are there fewer tiny yellow cylinders than yellow metal cubes? GE: No Pred: No	 Pred: Yes	 Pred: No	 Pred: No	 Pred: Yes	 Pred: Yes	 Pred: Yes

Figure 5. Illustrative results of LLaVA and the subsequent change in performance with various image augmentations.

from both datasets, each originally having a *yes-no* type answer, and perform logical operations to create compound questions. Table 3 illustrates an example of how the new questions used in the experiments have been created. We then balance the dataset by considering an equal number of samples for both *yes* and *no* answer-type questions.

The initial analysis indicates that the models perform relatively well on simple *yes-no* questions, achieving similar F1-scores for the *yes* and *no* classes. However, on introducing complexity to the questions, a variation is observed, as seen in Table 4. The F1-score of the three models shows an increase across most experiments for the *yes* class, consistently achieving F1-scores over 70% on the VQAv2 dataset and over 60% on the CLEVR dataset. These scores reflect a better performance for *yes*-type answers when compared to the original questions. However, a substantial drop in performance is observed for the *no*-type answers, resulting in an average gap of over 20% when compared to the performance of the *yes*-type answers. This finding suggests that these models have a tendency to lean towards affirmative responses when faced with complex questions involving binary connectives.

4. Qualitative Results

We showcase qualitative analysis of our experiments in Figure 5 and 6, respectively. We present examples of the LLaVA model since we find it to be the best-performing model among the three models. In Figure 5, we illustrated the effect of the multiple visual perturbations on the predictions made by LLaVA. As we find quantitatively, the model does not perform well on the Gaussian Low Pass Filter augmentation, with a similar performance drop on the zoom blur as well. For all other augmentations, the performance delta is less but lower in comparison to the baseline performance. Both in color and count-based questions, we find that the model is *close* to the ground truth answer however, it is not able to make an accurate prediction.

In Figure 6, we present the inability of the model to answer complex reasoning questions. We find that, while the model is able to answer *simpler* questions, on combining them with another question leveraging logical operators, it fails to understand the importance of the connectives. As also described before, the model tends to answer in the *affirmative* when asked questions with multiple connectives and especially when negation is involved.


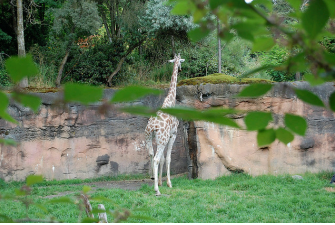

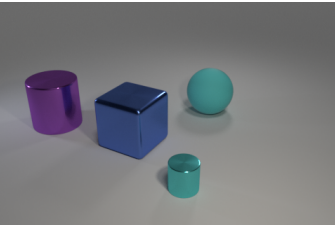
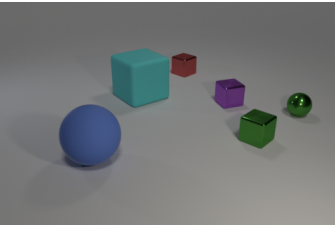
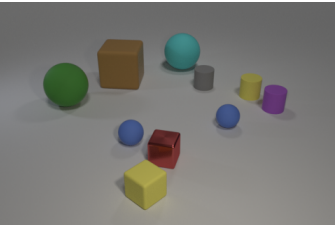
 <p>a. Question: Is there NOT a yellow vest? GT: No Pred: No</p> <p>b. Question: Is there a yellow vest OR are the skiers in a competition? GT: Yes Pred: Yes</p> <p>c. Question: Is there a yellow vest AND are the skiers NOT in a competition? GT: No Pred: No</p>	 <p>a. Question: Is this a zoo? GT: Yes Pred: Yes</p> <p>b. Question: Is this a zoo OR are all tree in this photo alive? GT: Yes Pred: Yes</p> <p>c. Question: Is this NOT a zoo AND are all tree in this photo alive? GT: No Pred: Yes</p>	 <p>a. Question: Is the building behind the motorcycle all brick? GT: No Pred: Yes</p> <p>b. Question: Is the building behind the motorcycle all brick OR is this a classic motorcycle? GT: Yes Pred: No</p> <p>c. Question: Is the building behind the motorcycle NOT all brick AND is this a classic motorcycle? GT: Yes Pred: No</p>
 <p>a. Question: Is the shiny cylinder the same color as the big cylinder? GT: No Pred: No</p> <p>b. Question: Are there lesser tiny cyan cylinders than cyan things OR is the tiny shiny cylinder the same color as the big cylinder? GT: Yes Pred: No</p> <p>c. Question: Are there more tiny cyan cylinders than cyan things AND is the tiny shiny cylinder the same color as the big cylinder? GT: No Pred: Yes</p>	 <p>a. Question: Are any blue balls visible? GT: Yes Pred: Yes</p> <p>b. Question: Are any blue balls visible OR is the material of the green block the same as the sphere right of the big matte sphere? GT: Yes Pred: Yes</p> <p>c. Question: Are any blue balls visible AND is the material of the green block NOT the same as the sphere right of the big matte sphere? GT: No Pred: No</p>	 <p>a. Question: Do the yellow block AND the green object not have the same size? GT: Yes Pred: No</p> <p>b. Question: Is there anything else that is the same material as the cyan thing OR do the yellow block and the green object have the same size? GT: Yes Pred: Yes</p> <p>c. Question: Is there anything else that is the same material as the cyan thing AND do the yellow block and the green object have the same size? GT: No Pred: Yes</p>

Figure 6. Illustrative results of LLaVA and the subsequent change in performance when presented with questions that contain different logical connectives.

5. Conclusion

In this work, we evaluate Multimodal Large Language Models and benchmark their robustness and generalization capabilities across datasets, vision-language augmentations, and reasoning abilities. We find that model performance shows a steady decline under distribution shift and presents a fine-grained analysis across multiple attributes. Our findings are crucial to fostering efficient and safe deployment of foundational models in real-world settings, where distribution shift is a common phenomenon. While our study sheds light on the limitations of these models under various settings, we advocate for rigorous evaluation of robustness through stringent benchmarks. It is imperative to acknowledge the potential limitations of our evaluation model, Zephyr. While it serves as a valuable method in evaluating the prediction, there exists a possibility that it might also be susceptible to errors. This recognition underscores the importance of ongoing scrutiny and refinement

of the models, ensuring that its outputs remain accurate and reliable. Moving forward, efforts should also be directed towards implementing robust validation processes and continuously improving the model to mitigate such risks. As these systems are actively deployed in real-world settings, it is imperative that these failures are better understood and addressed. We believe our findings are a first step towards this and can be used as a means for better utilization of foundational models in vision and language tasks.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan.

- Flamingo: a visual language model for few-shot learning, 2022. 2
- [2] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023. 1
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 2
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 2
- [5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic, 2023. 2
- [6] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023. 2
- [7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. 2
- [8] Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks, 2023. 3
- [9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3
- [10] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. 2
- [11] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2023. 2, 3
- [12] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xianguy Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 3
- [13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3
- [14] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7b, 2023. 1

- [15] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2, 3
- [16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 2
- [17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2
- [18] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation, 2023. 2
- [19] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. 2
- [20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 3
- [21] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2023. 2
- [22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 3
- [23] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2023. 2, 3
- [24] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reichihiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rim-bach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sasstry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Val-lone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welin-der, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 1
- [25] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. 2
- [26] Gemini Team. Gemini: A family of highly capable multi-

- modal models, 2023. [2](#)
- [27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. [1](#)
 - [28] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment. *ArXiv*, abs/2310.16944, 2023. [3](#)
 - [29] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and Conghui He. Vige: Visual instruction generation and correction, 2024. [2](#)
 - [30] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language, 2022. [2](#)
 - [31] Yuqing Wang and Yun Zhao. Gemini in reasoning: Unveiling commonsense in multimodal large language models, 2023. [2](#)
 - [32] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of llms: Preliminary explorations with gpt-4v(ision), 2023. [2](#)
 - [33] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023. [2](#)
 - [34] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023. [2](#), [3](#)
 - [35] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention, 2023. [1](#)
 - [36] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models, 2023. [2](#)