Sight Beyond Text: Multi-Modal Training Enhances LLMs in Truthfulness and Ethics

Anonymous authors
Paper under double-blind review

Abstract

Multi-modal large language models (MLLMs) are trained based on large language models (LLM), with an enhanced capability to comprehend multi-modal inputs and generate textual responses. While they excel in multi-modal tasks, the conventional view within the machine learning community has often undervalued/overlooked their capabilities in pure natural language processing. This paper aims to get out of the box and showcase an intriguing characteristic of multi-modal trained LLMs — our preliminary results suggest that visual instruction tuning, a prevailing strategy to integrate vision knowledge into the LLMs, unexpectedly and interestingly helps models attain both improved truthfulness and ethical alignment in the pure NLP context. For example, a visual-instruction-tuned LLaMA2 7B model surpasses the performance of the LLaMA2-chat 7B model, fine-tuned with over one million human annotations, on `TruthfulQA` and `Ethics` benchmarks. Another example is that the proprietary model GPT-4V, which incorporates visual information, surpasses its LLM-only counterpart GPT-4-turbo by around 0.4% on both aspects. Further analysis reveals that the improved alignment can be attributed to the superior instruction quality inherent to visual-text data. By presenting those findings, we advocate for a broader exploration into visual-text synergies, positing that such multi-modal interactions could be pivotal in advancing alignment research.

1 Introduction

Enhancing truthfulness and reducing hallucinations of Large Language Models (LLMs) is one the paramount challenges in the domain of artificial intelligence. This paper introduces a new perspective on this research topic, advocating for the integration of multi-modal data into LLM training as a strategy to significantly improve their truthfulness and alignment with human values.

Our stance is informed by empirical evidence demonstrating the beneficial impact of diverse data sources on LLM capabilities. For example, the inclusion of code data has been shown to improve the reasoning ability of LLMs (Ma et al., 2024). Building upon this premise, this paper aims to explore the potential benefits of an even more diverse data source – multi-modal data, particularly images, in enhancing the capabilities of LLMs.

Our claim is firmly grounded in our experimental evidence. In our preliminary explorations, we tune LLaMA series models (Touvron et al., 2023a;b) with the visual instruction data from LLaVA (Liu et al., 2023b;a). The results of these experiments are intriguing: for a vanilla LLaMA2 7B model, visual instruction tuning can register impressive scores of 46.0% on `TruthfulQA-mc` (+7.1%) (Lin et al., 2022) and 65.4% on `Ethics` (+19.6%) (Hendrycks et al., 2020), depending on the specific tuning approach. It is particularly noteworthy that, even without engineering efforts that explicitly elicit ethical or truthful behaviors, the performance of the visual instruction-tuned model already outperforms that of the LLaMA2-chat 7B variant, which is heavily tuned with over a million human annotations (Touvron et al., 2023b).

In proposing this novel perspective, we aim to spur a possible paradigm upgrade or even a complete shift to the ongoing dialogue within the machine learning community. We contend that broadening the data diversity for LLMs, beyond traditional text-based inputs, is a pivotal step towards developing models that more accurately reflect, interpret, and respond to the complexities of real-world information (Ma et al., 2023).

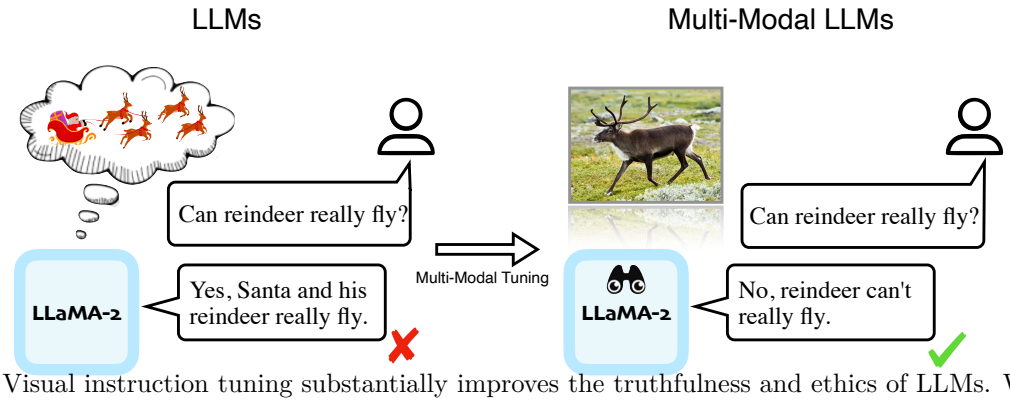

Figure 1: Visual instruction tuning substantially improves the truthfulness and ethics of LLMs. We observe tuning LLMs with only 80k multi-modal data can yield stronger results on truthfulness and ethics than those with over one million human-annotated RLHF data. Note that these LLMs employ images only during the visual instruction tuning and are tested without images for NLP tasks.

This paper seeks to engage the community in a discussion about this evolving approach, underlining its potential impact on the ethical and responsible aspects of AI development.

In summary, our insights accentuate the promise of visual instruction tuning in fostering the ethical and truthful alignment of LLMs. It is our hope that this paper will serve as a catalyst for a new wave of research, one that embraces the rich possibilities offered by multi-modal data and paves the way for more aligned and responsible AI systems.

2 Tuning LLMs with Multi-Modal Data

This section introduces our strategies to tune LLMs using multi-modal datasets. A standard MLLM typically contains three key components: 1) a vision encoder tasked with encoding visual inputs, 2) a vision-language connector that translates visual tokens into the linguistic space, and 3) an LLM for decoding the transcribed visual information. We strictly adhere to the setups in LLaVA (Liu et al., 2023b) for fine-tuning LLMs on visual instruction tuning data.

Model Architecture. We incorporate the pre-trained visual branch of CLIP ViT-L/14 (Radford et al., 2021) as our vision encoder. Additionally, a trainable linear layer is employed to project visual tokens into the language embedding space. Regarding the choice of LLM, we take the widely recognized open-sourced LLaMA models (Touvron et al., 2023a;b; Geng & Liu, 2023) for this study. Specifically, our investigation focuses on the following six models, containing three latest LLMs and their corresponding instruction-tuned variants:

- Pre-trained LLM: OpenLLaMA-3B (Geng & Liu, 2023), LLaMA-7B (Touvron et al., 2023a), LLaMA2-7B (Touvron et al., 2023b).
- Instruction-tuned LLM: OpenAlpaca-3B (Su et al., 2023), LLaMA2-chat-7B (Touvron et al., 2023b), the Vicuna family (Vicuna-7B, Vicuna-v1.5-7B, Vicuna-v1.5-13B) (Zheng et al., 2023).

As listed above, our study is centered on two model scales: 3B and 7B. While the 3B LLaMA model is sourced from the OpenLM project (Geng & Liu, 2023), the 7B LLaMA models are directly released by Meta (Touvron et al., 2023a;b); additionally, our investigation extends to the instruction-tuned variants of these base LLMs. Concretely, OpenAlpaca-3B is fine-tuned on the Alpaca data (Taori et al., 2023) using OpenLLaMA-3B as its backbone; Vicuna-7B is the v1.1 model from FastChat (Zheng et al., 2023), which is crafted upon LLaMA-7B and employs 125K conversational data from ShareGPT (ShareGPT) during tuning; LLaMA2-chat-7B is well-engineered for human alignment, undergoing its training on publicly available instruction datasets and one million human-annotated examples using RLHF techniques. Note that we test 7B LLM variants by default, and indicate 3B models by the suffix "-3B".

Models	`Ethics` Acc.	`TruthfulQA-gen` Rouge Acc.	`TruthfulQA-mc1` Acc.	`TruthfulQA-mc2` Acc.
LLaMA	50.4%	27.5%	22.0%	34.1%
MM-ft	59.1% (+8.7%)	29.4% (+1.8%)	23.6% (+1.6%)	35.8% (+1.7%)
Vicuna	66.6%	45.4%	32.0%	47.0%
MM-ft	60.4% (-6.2%)	29.3% (-16.1%)	23.8% (-8.1%)	35.8% (-11.2%)
LLaMA-3B	45.6%	25.3%	21.3%	34.6%
MM-ft	58.1% (+12.5%)	26.4% (+1.1%)	21.4% (+0.1%)	32.9% (-1.7%)
MM-lora	45.7% (+0.1%)	25.2% (-0.1%)	23.0% (+1.7%)	35.6% (+1.0%)
Alpaca-3B	44.0%	28.6%	22.4%	34.2%
MM-ft	46.8% (+2.8%)	28.2% (-0.4%)	23.1% (+0.7%)	34.2% (+0.0%)
MM-lora	44.0% (+0.0%)	28.6% (+0.0%)	24.6% (+2.2%)	38.0% (+3.8%)
LLaMA2	45.8%	32.3%	25.2%	38.9%
MM-ft	65.4% (+19.6%)	31.5% (-0.9%)	27.8% (+2.6%)	40.2% (+1.3%)
MM-lora	46.1% (+0.3%)	37.9% (+5.6%)	32.1% (+6.9%)	46.0% (+7.1%)
LLaMA2-chat	58.5%	43.3%	29.5%	44.6%
MM-ft	65.2% (+6.7%)	35.5% (-7.8%)	27.7% (-1.8%)	41.0% (-3.6%)
MM-lora	58.6% (+0.1%)	44.6% (+1.2%)	29.4% (-0.1%)	44.6% (+0.0%)
Vicuna-v1.5-7B	62.1%	40.6%	28.9%	45.4%
MM-ft	69.1% (+7.0%)	40.8% (+0.2%)	30.8% (+1.9%)	45.9% (+0.5%)
Vicuna-v1.5-13B	69.1%	39.0%	30.1%	45.9%
MM-ft	73.9% (+4.8%)	38.8% (-0.2%)	29.4% (-0.7%)	42.5% (-3.4%)

Table 1: Comparison on the original LLMs and the multi-modal fine-tuned ones on `Ethics` (Hendrycks et al., 2020) and `TruthfulQA` (Lin et al., 2022). '-ft' represents full parameter fine-tuning and '-lora' indicates LoRA tuning. We report Rouge-L accuracy for `TruthfullQA-gen` and accuracy for the rest.

Training Procedure. The MLLM training unfolds in two stages. First, we exclusively tune the weight of the vision-language connector, with both the visual encoder and the LLM remaining frozen. In the second phase, we fine-tune the weights of both the connector and the LLM. Data-wise, we adhere to the protocols set by LLaVA (Liu et al., 2023b): the connector is initially trained using 595k image-text pairings filtered from CC3M (Changpinyo et al., 2021); the subsequent stage utilizes 80k instructions-following data from LLaVA, containing image-grounded conversation, image descriptions, and image-based complex reasoning tasks. Note that in the second stage, we probe the effects of both full fine-tuning and LoRA fine-tuning (Hu et al., 2021).

3 Evaluations

3.1 Truthfulness and Ethics of MLLMs

We report the evaluation results on the `TruthfulQA` and `Ethics` benchmarks, designed for measuring LLMs' truthfulness and ethical alignment. During this evaluation, we utilize the weights exclusively from the visual-instruction-tuned LLMs, intentionally omitting the visual encoders and vision-language connectors introduced during the fine-tuning process. The results are presented in table 1.

Visual Instruction Tuning Improves Truthfulness and Ethics. Our observations suggest that, rather unexpectedly, visual instruction tuning tends to enhance the truthfulness of LLMs. A compelling observation emerges when comparing LLaMA2 variants: visual-instruction-tuned models, especially LLaMA2 with MM-lora, surpass the LLaMA2-chat model in performance metrics on both `TruthfulQA-mc1` (32.1% *vs.* 29.5%) and `TruthfulQA-mc2` (46.0% *vs.* 44.6%).

Data	Ethics Acc.	TruthfulQA BLEU Acc.	TruthfulQA Rouge Acc.
GPT-4	87.0%	55.0%	55.3%
GPT-4V	87.4% (+0.4%)	55.5% (+0.5%)	54.9% (-0.4%)

Table 2: Comparison on GPT-4-turbo and its multi-modal variant GPT-4V on Ethics and TruthfulQA generation. We report BLEU and Rouge-L accuracy for TruthfulQA generation task.

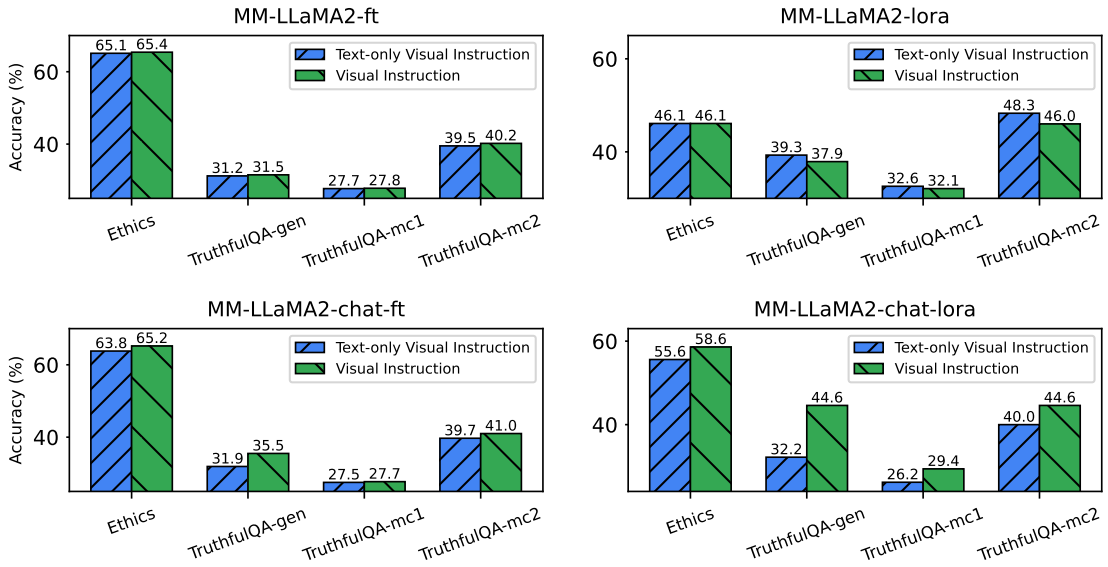

Figure 2: Performance of visual-instruction-tuned LLaMA2 models and text instruction tuned ones on Ethics and TruthfulQA benchmarks. The text-only visual instruction data is taken directly from LLaVA, but without the paired images.

From table 1, we also observe visual instruction tuning leads to substantial improvements on the Ethics task. Echoing the trend in the TruthfulQA evaluations, visual-instruction-tuned models, specifically the MM-ft versions of both LLaMA2 and LLaMA-3B, consistently outpace their instruction-tuned counterparts, such as LLaMA2-chat and Alpaca-3B. For example, the performance enhancements observed for LLaMA2 and LLaMA-3B on the Ethics task amounted to increments of 19.6% and 12.5% respectively, outperforming LLaMA2-chat and Alpaca-3B by margins of 6.9% and 11.3%. Another straightforward observation is that, models with larger parameter scale generally perform better in these two aspects (*e.g.*, 13B *vs.* 7B *vs.* 3B LLMs), as stronger base LLMs are more capable during the multi-modal tuning process.

It should be noted that the employed visual instruction tuning data is the 80k dataset derived from LLaVA (Liu et al., 2023b), which does not contain special designs for aligning models to human preferences. Remarkably, despite this, visual instruction tuning is able to yield empirical advantages that surpass those from RLHF, which heavily utilizes a substantial corpus of human-annotated data dedicated to LLM alignment. This observation strongly attests to the potential that visual instruction tuning holds in addressing AI alignment challenges. However, it is not a silver bullet — our experiments also show that visual instruction tuning is limited at enhancing the alignment of models previously fine-tuned via instruction tuning (*e.g.*, models like Vicuna, LLaMA2-chat), indicating a variability in its efficacy.

We also report model performance of proprietary GPT-4-turbo series on these two NLP tasks in table 2. The GPT-4V model is regarded as an upgrade of GPT-4-turbo, with the visual understanding ability. The GPT-4V demonstrates improved performance on Ethics by 0.4%, as well as on the TruthfulQA generation task under BLEU accuracy, further supporting our claim in bringing visual knowledge to enhance LLMs' ethical and truthful awareness.

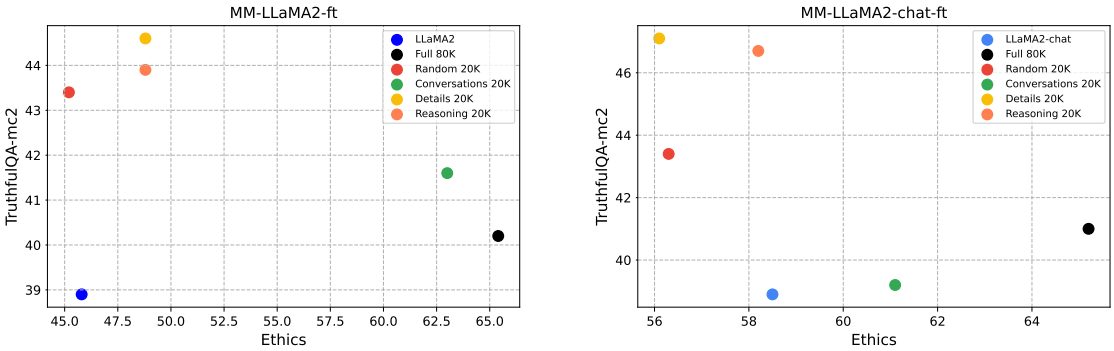

Figure 3: Results of different data components on `Ethics` and `TruthfulQA` of visual-instruction-tuned LLMs. We utilize 20K of different forms of data (Conversation, Details, Reasoning), and additionally sample 20K data out of the original 80K training instances (Random 20K) for comparison.

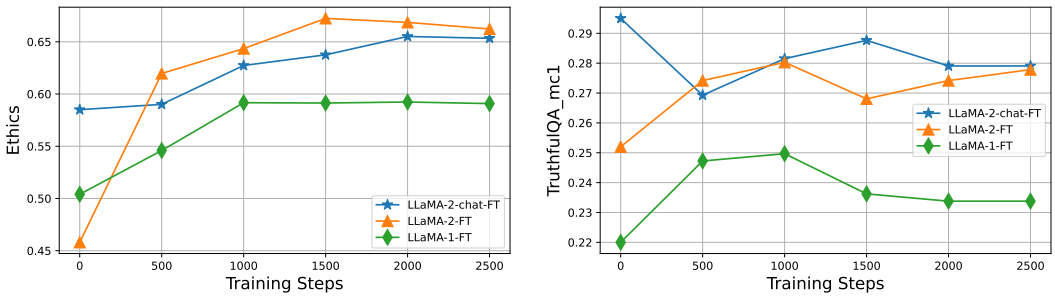

Figure 4: Results of three MLLMs on `Ethics` (*left*) and `TruthfulQA-mc1` (*right*) during MLLM visual instruction tuning.

Effects of Modalities in Visual Instruction-Tuning Data on LLM Alignment. Next, we seek to understand how different modalities in the visual instruction data contribute to the alignment of LLMs. Specifically, we design a set of ablations where we only utilize the text part of the visual instruction tuning data to tune the LLMs, and draw a comparison with the models tuned with both the visual inputs and the corresponding texts.

As shown in fig. 2, we observe that models with text-only visual instruction tuning can largely attain comparable alignment performance with the vanilla visual instruction tuning baseline where both images and texts are used. While additionally including visual inputs yields seemingly "modest" alignment improvements, we stress that these gains are consistent across different LLMs, tuning methods, and alignment tasks. For example, this can be verified across three model variants, resulting in an average accuracy improvement of 2.5% across three sub-tasks presented in fig. 2.

This observation leads to our hypothesis that there exists a promising avenue in leveraging visual data to construct enhanced instruction-tuning datasets. Although textual information plays a significant role in alignment, it is crucial to recognize that this text is inherently grounded in its corresponding real-world visual content; therefore, utilizing such paired information is integral to ensuring strong alignment in LLMs. These findings underscore the multifaceted benefits of visual data: it not only enhances alignment quality but also contributes significantly to the creation of more accurate instruction-tuning datasets.

Types of Visual Instruction Data Matters. We further extend our investigation to understand how varying types of visual instruction-tuning data affect LLM alignment. Specifically, we utilize data from LLaVA (Liu et al., 2023b), which categorizes visual instruction tuning data into three groups: `Conversation`, `Details`, and `Reasoning`. Each group comprises 20k data points, sampled from the original training splits. For a fair comparison, we also take a uniform sample of 20k from the full 80k visual instructions to form the

Models	MMLU (Acc.)	GSM8K (Acc.)	MathQA (Acc.)	sQuAD (F1.)	BoolQ (Acc.)
LLaMA	36.8%	8.0%	27.7%	19.5%	75.1%
MM-ft	27.7% (-9.0%)	0.9% (-7.1%)	28.5% (+0.8%)	9.1% (-10.4%)	47.5% (-27.6%)
Vicuna	47.2%	10.0%	29.0%	19.3%	78.1%
MM-ft	44.0% (-3.2%)	5.4% (-4.6%)	29.4% (+0.4%)	10.1% (-9.2%)	52.5% (-25.6%)
LLaMA-3B	26.7%	2.4%	26.4%	20.7%	65.6%
MM-ft	26.5% (-0.2%)	1.7% (-0.6%)	25.8% (-0.6%)	8.6% (-12.1%)	53.6% (-12.0%)
MM-lora	26.8% (+0.1%)	3.1% (+0.7%)	26.3% (-0.1%)	18.8% (-1.9%)	66.3% (+0.7%)
Alpaca-3B	24.9%	0.1%	24.6%	28.2%	71.1%
MM-ft	24.5% (-0.4%)	0.0% (-0.1%)	25.6% (+1.0%)	12.1% (-16.1%)	69.0% (-2.1%)
MM-lora	24.3% (-0.6%)	0.1% (+0.0%)	25.4% (+0.8%)	24.2% (-4.1%)	71.1% (+0.0%)
LLaMA2	45.9%	13.7%	30.1%	26.3%	77.7%
MM-ft	39.4% (-6.5%)	5.5% (-8.2%)	29.6% (-0.5%)	8.5% (-17.8%)	56.3% (-21.4%)
MM-lora	46.6% (+0.7%)	15.0% (+1.3%)	30.4% (+0.3%)	20.1% (-6.2%)	77.6% (-0.1%)
LLaMA2-chat	45.8%	18.2%	31.1%	20.1%	80.7%
MM-ft	45.2% (-0.6%)	6.2% (-12.0%)	30.0% (-1.1%)	10.2% (-9.3%)	67.0% (-13.7%)
MM-lora	45.9% (+0.2%)	17.1% (-1.1%)	30.8% (-0.3%)	25.5% (+5.4%)	81.5% (+0.8%)
Vicuna-v1.5-7B	50.0%	18.0%	30.0%	16.9%	82.1%
MM-ft	50.6% (+0.5%)	17.4% (-0.6%)	29.2% (-0.8%)	18.2% (+1.3%)	78.3% (-3.8%)
Vicuna-v1.5-13B	55.8%	31.9%	33.8%	16.3%	86.2%
MM-ft	56.0% (+0.2%)	27.5% (-4.4%)	33.0% (-0.8%)	18.6% (2.3%)	85.1% (-1.1%)

Table 3: Performances of both the vanilla LLMs and visual-instruction-tuned LLMs on five NLP capabilities benchmarks.

baseline group. We tune LLaMA2 and LLaMA2-chat with each data group (of 20k data points) separately, and report the results in fig. 3.

Our analysis reveals that, in general, conversational data has a greater impact on improving LLMs' performance on the `Ethics` task, resulting in an improvement of ~15% on MM-LLaMA2-ft and ~3% on MM-LLaMA2-chat-ft. Conversely, reasoning and details data tend to be more effective in improving performance on `TruthfulQA`, yielding gains of more than 2% and 6% on these two models. This suggests that a targeted approach, leveraging the unique strengths of each data type, can facilitate more nuanced and effective instruction tuning for LLM alignment.

Imperfect Match Between Multi-Modal and NLP Objectives. The incorporation of visual information has shown to benefit the ethical and truthful aspects of LLMs. In fig. 4, we present the model performance on the `Ethics` and `TruthfulQA` tasks during the multi-modal LLM finetuning stage. At the initial stage of visual instruction tuning, there is a noticeable improvement in most models for both aspects within the first 1000 training steps. Specifically, the scores on `Ethics` task continue to increase as more visual knowledge is incorporated, indicating a well-aligned training objective between visual instruction training and ethical awareness. However, the alignment between incorporating visual perception into LLMs and enhancing model truthfulness may not be optimal, as the scores for truthfulness degenerate with more training steps considered (*e.g.*, two LLaMA2 models achieve their highest `TruthfulQA-mc1` scores at the $1000th$ training step).

By analyzing the trajectory of model performance on these tasks, we observe that the optimization goals between multi-modal ability and the improved truthfulness and ethics are not perfectly aligned. Though LLMs trained on full visual tuning steps surpass the vanilla LLMs on ethics and truthfulness, the same training budgets designed for multi-modal tasks might not be optimal for models' NLP abilities.

3.2 Standard NLP Abilities

Given these LLMs are further fine-tuned with multi-modal data, it might be intuitively expected that their standard NLP capabilities could degrade. Such a phenomenon is commonly referred to as catastrophic

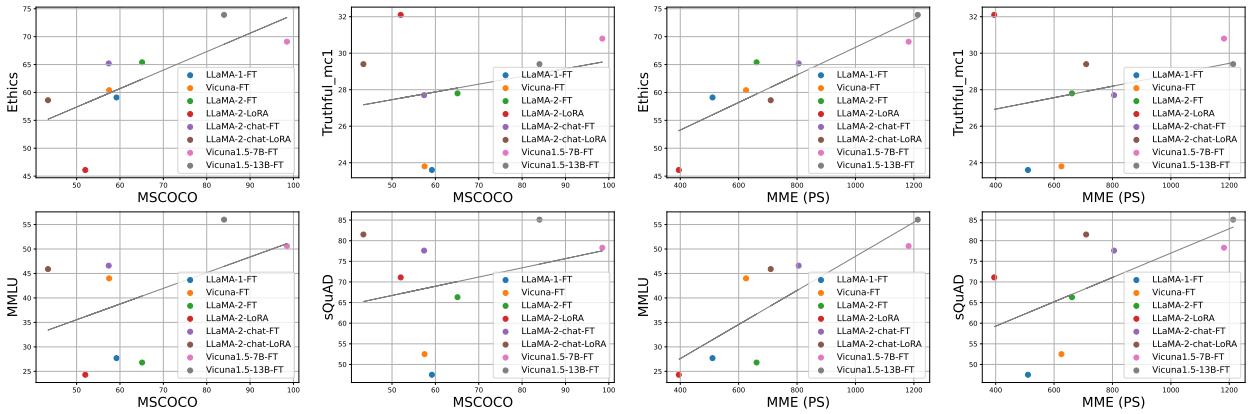

Figure 5: Results of eight different MLLMs on two multi-modal and four NLP tasks.

forgetting (Kirkpatrick et al., 2017) or in the AI alignment community — the alignment tax (Christiano, 2019; Jensen et al., 2023).

Interestingly, contrary to these assumptions, our results presented in table 3 show that MM-lora (marked in the gray background) results in only an average 0.17% performance decrease across five NLP capability benchmarks and four models, after applying visual instruction tuning. More notably, in certain instances, MM-lora even modestly improves performance on these benchmarks.

Also, MLLMs with better the LLM component, *i.e.* more advanced model, larger model scale can perform better on these NLP benchmarks — multi-modal tuned Vicuna-v1.5-13B surpasses its 7B variant and the v1.3 counterpart by average 5.3% and 15.8% on five NLP tasks, respectively.

In conjunction with the insights from Section 3.1, these observations altogether highlight the ability of visual-instruction-tuned LLMs in both maintaining the strong capability on standard NLP benchmarks and aligning better with human values, not to mention the additional capability of recognizing visual inputs. Such findings pave new avenues for both academic exploration and practical implementations within multi-modal domains. We believe these insights should catalyze further investigations into the tuning of LLMs with multi-modal interactions.

3.3 Analysis on Multi-Modal Benchmarks

Models	Unicorn oodcv / sketch	MME^{CS}	MME^{PS}	COCO	Flickr30k	POPE R / A / P
MM-LLaMA-ft	45.2 / 80.9	199.3	510.5	59.2	27.1	65.7 / 57.8 / 59.9
MM-Vicuna-ft	58.4 / 82.7	270.7	625.2	57.5	24.6	76.5 / 66.5 / 73.8
MM-LLaMA2-ft	55.0 / 83.6	237.1	661.3	65.1	31.6	65.0 / 55.4 / 56.3
MM-LLaMA2-lora	- / -	200.0	395.0	52.0	26.2	50.8 / 50.4 / 50.6
MM-LLaMA2-chat-ft	54.5 / 80.9	234.6	805.4	57.4	26.7	69.8 / 57.9 / 60.3
MM-LLaMA2-chat-lora	- / -	228.6	709.8	43.4	23.0	65.9 / 56.8 / 59.2
MM-Vicuna-v1.5-7B	58.4 / 80.4	**320.4**	1182.0	**98.5**	**62.8**	**89.3 / 79.7/ 85.5**
MM-Vicuna-v1.5-13B	**59.7 / 87.8**	287.9	**1213.3**	84.0	51.3	87.4 / 78.7 / 84.1

Table 4: Performances of our MLLM family on five widely employed multi-modal benchmarks. We test models on oodcv and sketch sub-tasks in the Unicorn benchmark (Tu et al., 2023a) and Random (R), Adversarial (A), and Popular (P) in POPE (Li et al., 2023c).

In this section, we test the visual-instruction tuned models on recent multi-modal evaluation benchmarks, where five multi-modal benchmarks are deployed: Unicorn benchmark (Tu et al., 2023a) dedicates evaluating the MLLM ability in safety scenarios, we take two OODCV-VQA tasks and Sketchy-VQA tasks for testing whether

Models	COCO (CIDEr)	COCO-C (CIDEr)
MM-LLaMA-ft	59.2	48.6 (-17.9%)
MM-Vicuna-ft	57.5	46.0 (-20.0%)
MM-LLaMA2-ft	**65.1**	**54.6** (-16.1%)
MM-LLaMA2-lora	52.0	43.2 (-16.9%)
MM-LLaMA2-chat-ft	57.4	47.5 (-17.2%)
MM-LLaMA2-chat-lora	43.4	33.8 (-22.1%)

Table 5: Performances of the MLLM family on MSCOCO (Lin et al., 2014) with corrupted visual inputs.

models can well handle OOD visual/text input and sketch images, respectively. MME (Fu et al., 2023) consists of two evaluation aspects, *i.e.*, cognition (CS) and perception (PS) with total 14 VQA tasks;[1] MSCOCO (Lin et al., 2014) and Flickr30k (Young et al., 2014) captioning tasks are commonly used benchmarks in the field of image caption generation. We report the zero-shot CIDEr (Vedantam et al., 2015) scores (with three text-only QA examples) on the test set from the Karpathy split (Karpathy & Fei-Fei, 2015). POPE (Li et al., 2023c) is used to evaluate the level of object hallucinations in MLLMs, which consists of three versions of balanced yes/no VQA tasks considering objects in the given image. It is built upon MSCOCO-2017 dataset (Lin et al., 2014). Additionally, We also make use of the image corruptions proposed in ImageNet-C (Hendrycks & Dietterich, 2019) to measure the performance of the MLLMs on corrupted images for MSCOCO task (denoted as MSCOCO-C).[2].

Enhanced MLLMs Expand NLP Capabilities. *Is a better visual reasoner also a better NLP task solver?* In fig. 5, we illustrate the correlation between model performance in multi-modal and NLP tasks. The results reveal positive correlations across tasks from different domains, suggesting that the visual reasoning abilities of the eight models analyzed contribute to their boosted performance in NLP benchmarks. Unlike the misaligned objectives observed between MLLM multi-modal and NLP abilities during training in Sec. 3.1, this finding might not be surprising, as MLLMs rely heavily on language to reasoning and expression. It is plausible that an improved LLM (*i.e.* better training data, larger model scale) could enhance the expressive abilities of an MLLM, resulting in a close correlation between abilities across different modalities.

Potential Inconsistency in Current Multi-Modal Benchmarks. In table 4, MLLMs incorporating *aligned* LLMs have demonstrated superior performance in comprehensive and challenging tasks such as Unicorn, MME and POPE. Specifically, MM-Vicuna-ft and MM-LLaMA-chat-ft outperform their corresponding vanilla MLLM counterparts by an average of 164.9 on MME and 7.5% on POPE. However, despite the incorporation of text-aligned LLMs, MLLMs exhibit unexpected shortcomings in comparison to models leveraging vanilla LLMs when evaluated on three traditional vision-text tasks (*e.g.*, an average 4.2 CIDEr score drop on two captioning tasks). The inconsistent performance across these benchmarks highlights the imperative for improving evaluation techniques within multi-modal benchmarks.

Need for Studying Multi-Modal Alignments. Despite the effectiveness of text-aligned models like Vicuna and LLaMA2-chat, their MLLM variants exhibit poor performance on corrupted images as shown in table 5. These models not only lag behind MLLMs without instruction-tuned LLMs, but also demonstrate performance drops of over 17% when evaluated on corrupted images compared to clean ones, which are higher than drops observed for MM-LLaMA-ft and MM-LLaMA2-ft. This observation indicates that though visual instruction tuning improves the truthfulness and ethics of LLMs in the language domain, these MLLMs still face their unique challenges in the multi-modal domain.

4 Related Work

Alignments. The alignment of AI systems to human values is an important topic for today's advanced AI systems, from testing model robustness to out-of-distribution shifts (Hendrycks & Dietterich, 2019; Hendrycks

[1]We exclude landmark and artwork tasks to accelerate the evaluation process.

[2]For corrupted images, we report the average results of tested models on four noises (gaussian noise, defocus blur, contrast, brightness) across three severity levels (1, 3, 5)

et al., 2021a; Zhao et al., 2022) to adversarial attacks (Hendrycks et al., 2021b; Eykholt et al., 2018; Xie et al., 2020), many works have been proposed. The recent development of LLMs has revolutionized natural language processing and has been widely adopted in various applications. Thus, concerns regarding the honesty and truthfulness of these models have also emerged, prompting alignment researchers to investigate the ethical implications and potential risks associated with their deployment. `TruthfulQA` (Lin et al., 2022) is proposed to measure how LLMs imitate human misconceptions. And `Ethics` (Hendrycks et al., 2020) is used to assess a language model's knowledge of basic concepts of morality. Advanced techniques for aligning language models with human preference are also popular these days, from RLHF (Ouyang et al., 2022) to DPO (Rafailov et al., 2023) optimization, the alignment training paradigms have shifted fast recently. The concept of LLM alignment has also gradually switched from human-supervised (Ouyang et al., 2022) to the paradigm of incorporating other AI model supervisions (Lee et al., 2023), and to most recently the employment of weak signals (Burns et al., 2023). Given the popularity of the use of large language models, adversarial attacks on LLMs have also been explored (Zou et al., 2023). In this work, we present our findings on how visual instruction tuning can help the LLMs align with human values, our results show impressive performance boost on these datasets without explicit prompting such behaviors.

Multi-Modal and Large Language Models. Multi-modality has long been a hot topic, CLIP (Radford et al., 2021) proposes to align representations of both images and text, and later works proposed more techniques for this aim (Yu et al., 2022; Mu et al., 2022; Zhao et al., 2023). In light of the rapid evolvement of large language models (LLMs), recent studies about multi-modal systems have turned their focuses from incorporating fine-grained multi-modal data (Liang et al., 2021; Tu et al., 2023b) to integrating powerful LLMs with few-shot capability. More recently, some instruction-tuned MLLMs have emerged, showing excellent generalization ability in unseen VL tasks (Zhu et al., 2023; Liu et al., 2023b; Ye et al., 2023; Li et al., 2023a; Dai et al., 2023). For example, MiniGPT4 (Zhu et al., 2023) is built upon QFormer (Li et al., 2023a) and Vicuna (Zheng et al., 2023) and only activates the linear layer connecting the vision encoder and LLM. LLaVA (Liu et al., 2023b;a) projects the output of a vision encoder to word tokens and trains both the VL connector and the LLM on synthetic data. mPLUG-owl (Ye et al., 2023) tunes LLaMA with a query-based VL connector using both text-only and vision-language instruction data. InstructBLIP (Dai et al., 2023) uses BLIP2 (Li et al., 2023a) as the backbone but is additionally instruction-tuned on a collection of VL datasets. Other multi-modal LLMs in a vast range of modalities sparkles insights and deployment in real-world applications (Zhang et al., 2023c; Bai et al., 2023; Bavishi et al., 2023; Zhang et al., 2023a; Liu et al., 2024).

Despite the rapid growth in this domain, recent benchmark works have shown that current multi-modal large language models still suffer from problems like being unable to handle counterfactual statements (Zhang et al., 2023b; Wu et al., 2023; Yu et al., 2023), hallucination (Li et al., 2023b; Zhou et al., 2023), and simple answer set permutations (Zong et al., 2023). In our work, we demonstrate a new perspective on these MLLMs – tuning LLMs with multi-modal data greatly helps align them with human values.

5 Discussion, Conclusion, and Future Work

More Aligned Objectives between Multi-Modal and NLP Abilities. Our exploration shows that training on multi-modal instruction tuning data can also benefit the LLMs' factual accuracy and ethics. In fig. 4, we have shown that these alignment-focused metrics improve while training proceeds on multi-modal. However, current multi-modal data is not designed for alignment, the main focus is still eliciting language models with multi-modal perception. Our results demonstrate a promising new avenue for developing models to understand and interact with the world more truthfully, and also suggest the need for exploration to identify appropriate tasks that can effectively improve these two aspects simultaneously. And we hope our paper could inspire discussions on this direction.

Exploring the Training Framework. In our pilot study, we have shown the results of using image-based instruction fine-tuning data to support our findings that leveraging multi-modal interactions could yield more aligned models. Based on our results, it is reasonable to assume that introducing multi-modal data to the pre-training stage could also yield more aligned models. For example, the Gemini family models could be an interesting case for study (Team et al., 2023). Understanding how to instruction fine-tune the base model

for multi-modal ability and alignment is another direction worthy of exploration. Our study explores full parameter fine-tuning as well as LoRA parameter efficient fine-tuning. It can be beneficial to study how varies types of parameter-efficient fine-tuning techniques helps (Kopiczko et al., 2024; Zhao et al., 2024). Besides the training techniques, the training data can also be explored, how should we create a mixture of data for fine-tuning, how to determine the ratio of multi-modal data to text-only data (Ye et al., 2023; Liu et al., 2023a), and how to extend to other modalities other than images. These exploration could help gives a more practical and comprehensive guide.

Conclusion. In this study, we offer preliminary findings that underscore the potential of enhancing the truthfulness and ethical alignment of LLMs through visual instruction tuning. Remarkably, even without prompts tailored for truthfulness or ethical behaviors, our approach to tuning LLM weights using visual instruction datasets yielded significant improvements in both the `TruthfulQA` and `Ethics` benchmarks. Notably, such improvements are even stronger than that of RLHF, which tunes LLMs with a huge corpus of human-aligned data points. The follow-up analysis demonstrates the importance of instruction data quality for improving aligned values in MLLMs, as well as specific types of data models employed for applying to different alignment tasks.

Future Work. In light of our findings, we advocate for future research endeavors to focus on devising innovative methodologies for crafting visual instruction data that can more effectively align LLMs. Exploring novel MLLM architectures could also be a fruitful avenue. We hope fostering LLM interactions with real-world environments may emerge as a pivotal strategy for achieving superior model alignment.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 9

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL https://www.adept.ai/blog/fuyu-8b. 9

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023. 9

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

Paul Christiano. Current work in ai alignment. *Effective Altruism*, 2019. 7

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. *ArXiv*, abs/2305.06500, 2023. URL https://api.semanticscholar.org/CorpusID:258615266. 9

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018. 9

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 8

Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama. https://github.com/openlm-research/open_llama, May 2023. 2

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 8

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*, 2020. 1, 3, 9

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a. 8

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021b. 9

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

Mckay Jensen, Nicholas Emery-Xu, and Robert Trager. Industrial policy for advanced ai: Compute pricing and the safety tax. *arXiv preprint arXiv:2302.11436*, 2023. 7

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 8

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 2017. 7

Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki Markus Asano. Vera: Vector-based random matrix adaptation. In *ICLR*, 2024. 10

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023. 9

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a. 9

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. 2023b. URL https://arxiv.org/pdf/2305.10355. 9

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023c. 7, 8

Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xiubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. Maria: A visual experience powered conversational agent. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5596–5611, 2021. 9

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022. 1, 3, 9

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014. 8

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *ICLR*, 2024. 9

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023a. 1, 9, 10

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b. 1, 2, 3, 4, 5, 9

Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At which training stage does code data help llms reasoning? In *ICLR*, 2024. 1

Yuxi Ma, Chi Zhang, and Song-Chun Zhu. Brain in a vat: On missing pieces towards artificial general intelligence in large language models. *arXiv preprint arXiv:2307.03762*, 2023. 1

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, 2022. 9

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 9

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 2, 9

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023. 9

ShareGPT. https://sharegpt.com. 2

Yixuan Su, Tian Lan, and Deng Cai. Openalpaca: A fully open-source instruction-following model based on openllama. https://github.com/yxuansu/OpenAlpaca, 2023. 2

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 2

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 9

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a. 1, 2

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b. 1, 2

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023a. 7

Haoqin Tu, Yitong Li, Fei Mi, and Zhongliang Yang. Resee: Responding through seeing fine-grained visual knowledge in open-domain dialogue. *arXiv preprint arXiv:2305.13602*, 2023b. 9

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015. 8

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*, 2023. 9

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 819–828, 2020. 9

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 9, 10

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 8

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 9

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. *arXiv preprint arXiv:2305.14010*, 2023. 9

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a. 9

Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Xin Wen, and Bingchen Zhao. What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2023b. 9

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023c. 9

Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: a benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *European Conference on Computer Vision*, pp. 163–180. Springer, 2022. 9

Bingchen Zhao, Quan Cui, Hao Wu, Osamu Yoshie, Cheng Yang, and Oisin Mac Aodha. Vision learners meet web image-text pairs. *arXiv preprint arXiv:2301.07088*, 2023. 9

Bingchen Zhao, Haoqin Tu, Chen Wei, Jieru Mei, and Cihang Xie. Tuning layernorm in attention: Towards efficient multi-modal llm finetuning. In *ICLR*, 2024. 10

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 2, 9

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. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. 9

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 9

Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations. *arXiv preprint arXiv:2310.01651*, 2023. 9

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. 9