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

Haoqin Tu*[1] **Bingchen Zhao***[2] **Chen Wei**[3] **Cihang Xie**[4]

*equal contribution

[1] University of Chinese Academy of Sciences [2] University of Edinburgh
[3] Johns Hopkins University [4] UC Santa Cruz

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 pure NLP abilities of MLLMs are often underestimated and left untested. In this study, we get out of the box and unveil an intriguing characteristic of MLLMs — our preliminary results suggest that visual instruction tuning, a prevailing strategy for transitioning LLMs into MLLMs, 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. Further analysis reveals that the improved alignment can be attributed to the superior instruction quality inherent to visual-text data. In releasing our code at `github.com/UCSC-VLAA/Sight-Beyond-Text`, we aspire to foster further exploration into the intrinsic value of visual-text synergies and, in a broader scope, multi-modal interactions in alignment research.

1 Introduction

As large language models (LLMs) have already catalyzed significant transformations across various fields, their evolution to multi-modal capabilities, facilitating interactions with inputs from various domains in a human-like way, has garnered substantial interest. While prior research on multimodal LLMs (MLLMs) largely focuses on their prowess in tasks such as visual reasoning and visual-grounded generation [21, 41], there remains a limited investigation for reassessing their performance on pure NLP tasks.

In this paper, we are interested in probing how multi-modal training can affect the "hallucination" phenomenon observed in LLMs, where they generate misleading or factually incorrect content. Our focus is primarily on visual instruction tuning, a technique that turns LLMs into MLLMs by tuning with visual-text inputs. Our findings are compelling — visual instruction tuning intriguingly presents a promising pathway to enhance the truthfulness and ethics of LLMs. For example, a vanilla LLaMA2 7B model, post visual instruction tuning, is able to register impressive scores of 46.0% on `TruthfulQA-mc` (+7.1%) [19] and 65.4% on `Ethics` (+19.6%) [8], depending on the specific tuning approach. It is noteworthy that, even without engineering prompts that explicitly prioritize ethical or truthful behaviors, the performance of our visually-instructed model already outperforms that of the LLaMA2-chat 7B variant, which is heavily tuned with over a million human annotations [28]. Moreover, our analysis suggests that the primary factor in driving such stronger alignments is the quality of visual-text instructions, and our ablations show that the efficacy of different forms of visual instruction data is intricately tied to the specific alignment standard measured.

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

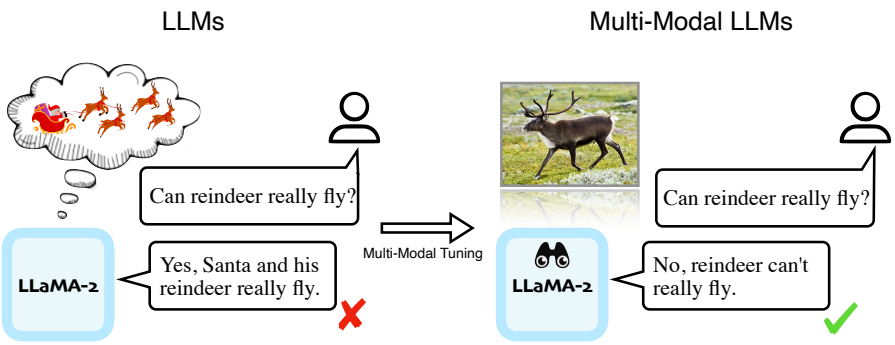

Figure 1: Visual instruction tuning substantially improves the truthfulness and ethics of LLMs. We observe that tuning LLMs with only 80k multi-modal data can yield stronger results on truthfulness and ethics than those with over one million human-annotated RLHF data. Note that these LLMs employ images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

In summary, our insights accentuate the promise of visual instruction tuning in fostering ethical and truthful alignment in LLMs. We present our findings as a cornerstone, inspiring the research community to explore the potential of visual instruction tuning, and, more broadly, multi-modal interactions in rectifying AI alignment conundrums.

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 [21] for fine-tuning LLMs on visual instruction tuning data.

Model architecture. We incorporate the pre-trained visual branch of CLIP ViT-L/14 [23] 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 [27, 28, 6] 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 [6], LLaMA-7B [27], LLaMA2-7B [28].
- Instruction-tuned LLM: OpenAlpaca-3B [25], Vicuna-7B [40], LLaMA2-chat-7B [28].

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 [6], the 7B LLaMA models are directly released by Meta [27, 28]; additionally, our investigation extends to the instruction-tuned variants of these base LLMs. Concretely, OpenAlpaca-3B is fine-tuned on the Alpaca data [26] using OpenLLaMA-3B as its backbone; Vicuna-7B is the v1.1 model from FastChat [40], which is crafted upon LLaMA-7B and employs 125K conversational data from ShareGPT [24] 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".

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 [21]: the connector is initially trained using 595k image-text pairings filtered from CC3M [1]; the subsequent stage utilizes 80k instructions-following data from LLaVA [21], 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 [11].

Models	Ethics	TruthfulQA-gen	TruthfulQA-mc1	TruthfulQA-mc2
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%)

Table 1: Comparison on the original LLMs and the multi-modal fine-tuned ones on `Ethics` [8] and `TruthfulQA` [19]. '-ft' represents full parameter fine-tuning and '-lora' indicates LoRA tuning. We report Rouge-L accuracy for `TruthfullQA-gen` and accuracy for the rest.

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%).

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 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%.

It should be noted that the employed visual instruction tuning data is the 80k dataset derived from LLaVA [21], 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.

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.

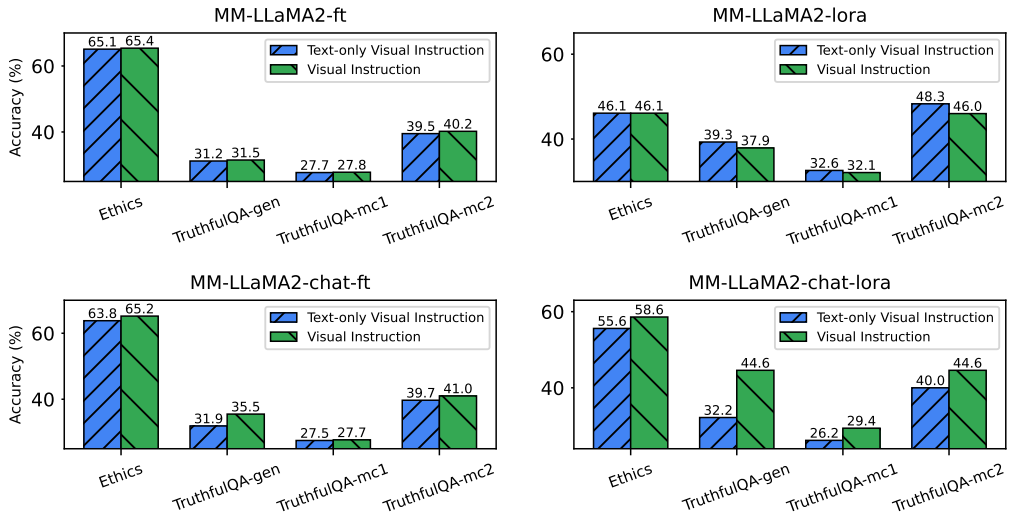

Figure 2: Performance of visual-instruction-tuned LLaMA2 models and text instruction tuned ones on `Ethics` and `TruthfulQA` benchmarks. The text-only visual instruction data is taken directly from LLaVA, but without the paired images.

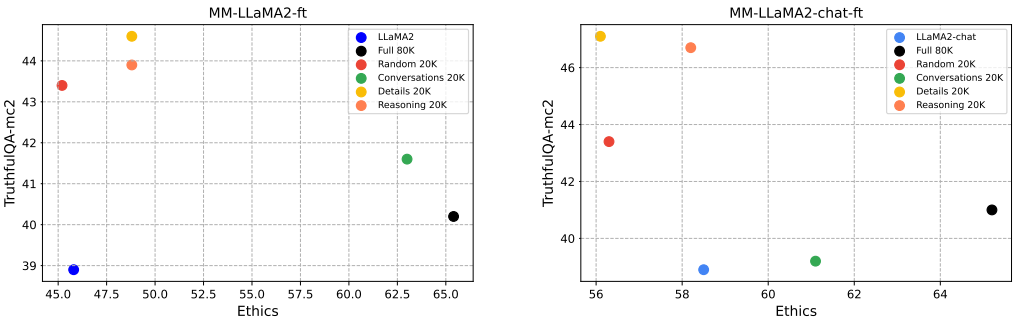

Figure 3: Results of different data components on `Ethics` and `TruthfulQA` of visual-instruction-tuned LLMs. We utilize 20K of different forms of data (Conversation, Details, Reasoning), and additionally sample 20K data out of the original 80K training instances (Random 20K) for comparison.

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

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%)

Table 2: Performances of both the vanilla LLMs and visual-instruction-tuned LLMs on five NLP capabilities benchmarks.

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 the TruthfulQA benchmark, 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.

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 forgetting [14] or in the AI alignment community — the alignment tax [2, 12].

Interestingly, contrary to these assumptions, our results presented in table 2 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.

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	MME CS	MME PS	MSCOCO	Flickr30k	POPE R / A / P
MM-LLaMA-ft	199.3	510.5	59.2	27.1	65.7 / 57.8 / 59.9
MM-Vicuna-ft	**270.7**	625.2	57.5	24.6	**76.5 / 66.5 / 73.8**
MM-LLaMA2-ft	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	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

Table 3: Performances of our MLLM family on five widely employed multi-modal benchmarks. On POPE, we test all models on three sub-tasks: Random (R), Adversarial (A), and Popular (P).

Models	MSCOCO (CIDEr)	MSCOCO-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 4: Performances of the MLLM family on MSCOCO [20] with corrupted visual inputs.

In this section, we test the visual-instruction tuned models on recent multi-modal evaluation benchmarks, where five multi-modal benchmarks are deployed: MME [5] consists of two evaluation aspects, *i.e.*, cognition (CS) and perception (PS) with total 14 VQA tasks;[1] MSCOCO [20] and Flickr30k [34] captioning tasks are commonly used benchmarks in the field of image caption generation. The former two benchmarks are based on MSCOCO-2017 dataset [20]. For the latter two captioning tasks, we report the zero-shot CIDEr [30] scores (with three text-only QA examples) on the test set from the Karpathy split [13]. Additionally, We also make use of the image corruptions proposed in ImageNet-C [9] to measure the performance of the MLLMs on corrupted images for MSCOCO tasks (denoted as MSCOCO-C).[2]. POPE [16] 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.

Potential inconsistency in current multi-modal benchmarks. In table 3, MLLMs incorporating text-aligned LLMs have demonstrated superior performance in comprehensive and challenging tasks such as 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 comparison across these five 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 4. 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 [9, 7, 39] to adversarial attacks [10, 4, 32], 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 [19] is proposed to measure how LLMs imitate human misconceptions. And Ethics [8] is used to assess a language model's knowledge of basic concepts of morality. Given the popularity of the use of large language models, adversarial attacks on LLMs have also been explored [43]. 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.

[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)

Multi-Modal and Large Language Models. Multi-modality has long been a hot topic, CLIP [23] proposes to align representations of both images and text, and later works proposed more techniques for this aim [35, 22, 38]. 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 [18, 29] to integrating powerful LLMs with few-shot capability. More recently, some instruction-tuned MLLMs have emerged, showing excellent generalization ability in unseen VL tasks [41, 21, 33, 15, 3]. For example, MiniGPT4 [41] is built upon QFormer [15] and Vicuna [40] and only activates the linear layer connecting the vision encoder and LLM. LLaVA [21] projects the output of a vision encoder to word tokens and trains both the VL connector and the LLM on synthetic data. mPLUG-owl [33] tunes LLaMA with a query-based VL connector using both text-only and vision-language instruction data. InstructBLIP [3] uses BLIP2 [15] as the backbone but is additionally instruction-tuned on a collection of VL datasets. 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 [37, 31, 36], Hallucination [17], and simple answer set permutations [42]. 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 Conclusion, Discussion, and Future Work

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.

In light of our findings, we advocate for future research endeavors to focus on devising innovative methodologies for crafting visual instruction tuning 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.

Acknowledge

This work is partially supported by a gift from Open Philanthropy. We thank Center for AI Safety for supporting our computing needs.

References

[1] 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. 2

[2] Paul Christiano. Current work in ai alignment. *Effective Altruism*, 2019. 5

[3] 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. 7

[4] 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*, pages 1625–1634, 2018. 6

[5] 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. 6

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

[7] 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*, 2021. 6

[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, 6

[9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 6

[10] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 6

[11] 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. 2

[12] 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. 5

[13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 6

[14] 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. 5

[15] 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*, 2023. 7

[16] 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*, 2023. 6

[17] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. 2023. 7

[18] 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)*, pages 5596–5611, 2021. 7

[19] 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)*, pages 3214–3252, 2022. 1, 3, 6

[20] 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*, pages 740–755. Springer, 2014. 6

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 3, 4, 7

[22] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, 2022. 7

[23] 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, 7

[24] ShareGPT. https://sharegpt.com. 2

[25] 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

[26] 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

[27] 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*, 2023. 2

[28] 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*, 2023. 1, 2

[29] 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*, 2023. 7

[30] 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*, pages 4566–4575, 2015. 6

[31] 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. 7

[32] 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*, pages 819–828, 2020. 6

[33] 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. 7

[34] 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. 6

[35] 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. 7

[36] 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. 7

[37] 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*, 2023. 7

[38] 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. 7

[39] 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*, pages 163–180. Springer, 2022. 6

[40] 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, 7

[41] 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. 1, 7

[42] 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. 7

[43] 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. 6