

Training-Free Mitigation of Language Reasoning Degradation After Multimodal Instruction Tuning

Neale Ratzlaff¹, Man Luo¹, Xin Su¹, Vasudev Lal¹, Phillip Howard¹

¹Intel Labs

neale.ratzlaff@intel.com, man.luo@intel.com, xin.su@intel.com, vasudev.lal@intel.com, phillip.r.howard@intel.com

Abstract

Multimodal models typically combine a powerful large language model (LLM) with a vision encoder and are then trained on multimodal data via instruction tuning. While this process adapts LLMs to multimodal settings, it remains unclear whether this adaptation compromises their original language reasoning capabilities. In this work, we explore the effects of multimodal instruction tuning on language reasoning performance. We focus on LLaVA, a leading multimodal framework that integrates LLMs such as Vicuna or Mistral with the CLIP vision encoder. We compare the performance of the original LLMs with their multimodal-adapted counterparts across eight language reasoning tasks. Our experiments yield several key insights. First, the impact of multimodal learning varies between Vicuna and Mistral: we observe a degradation in language reasoning for Mistral but improvements for Vicuna across most tasks. Second, while multimodal instruction learning consistently degrades performance on mathematical reasoning tasks (e.g., GSM8K), it enhances performance on commonsense reasoning tasks (e.g., CommonsenseQA). Finally, we demonstrate that a training-free model merging technique can effectively mitigate the language reasoning degradation observed in multimodal-adapted Mistral and even improve performance on visual tasks.

Introduction

Multimodal LLMs (MLLMs) have gained significant attention due to their ability to integrate various forms of data, allowing them to perform tasks that require both image and language understanding (Li et al. 2023). One common approach to building MLLMs is to connect a powerful LLM with a vision encoder (Radford et al. 2021) through an intermediate module, followed by multimodal instruction tuning (Wang et al. 2024). This has enabled MLLMs to excel in tasks such as visual question answering (Antol et al. 2015) and image captioning (Lin et al. 2014) by integrating and interpreting both visual and textual inputs (Liu et al. 2024a). While this process equips the model with multimodal capabilities, it may also impact language performance (Huang et al. 2023).

In this work, we study the behavior of MLLMs on language reasoning tasks and aim to answer the question: “*How does multimodal instruction learning affect language reasoning performance?*”. The investigation of this question can

lead to practical guidelines for the deployment of MLLMs in real-world applications such as chatbots, where the user can ask a question purely in language or optionally upload an image to accompany their query. Few prior studies have explored this question, with only a limited number of methods proposed for mitigating language degradation (Zhang et al. 2024). Our work extends this line of research by exploring how the choice of the base LLM affects the degree of language reasoning degradation during MLLM training, and whether this phenomenon can be effectively mitigated without the need for additional model training.

We evaluate MLLMs on a broad range of vision and language reasoning tasks, leading to two major observations. First, the impact of multimodal instruction tuning varies greatly depending on the choice of base LLM. We observe a significant erosion of language reasoning capabilities for LLaVA-Mistral, while LLaVA-Vicuna largely retains its performance and even outperforms the base Vicuna LLM on some tasks. Second, the impact of multimodal instruction tuning on a single model’s performance is not uniform across tasks. In particular, we find that common-sense reasoning improves and mathematical reasoning degrades after visual instruction tuning. In light of these findings, we propose a simple training-free method based on model merging (Iiharco et al. 2022) to mitigate degradation of MLLM language reasoning while preserving or even improving multimodal capabilities. Our experiments show that model merging techniques can effectively prevent language reasoning degradation while also improving performance on multimodal tasks.

Background & Related Work

Liu et al. (2024b) used a synthetic dataset of multimodal language-image instructions generated by GPT-4 to train LLaVA, an MLLM which combines the CLIP (Radford et al. 2021) vision encoder with a pre-trained Vicuna (Zheng et al. 2024) LLM. Using a projection layer to encode image representations in the word embedding space of the LLM, LLaVA learns via its visual instruction tuning to integrate information across both modalities. This can be viewed as a form of domain adaptation, as the weights of the pre-trained LLM are updated as it learns to integrate representations from the vision encoder. Liu et al. (2024a) further extended the LLaVA visual instruction tuning dataset to incorporate other academic task-oriented data. A variety of datasets have been

developed for visual instruction tuning of other MLLMs (Bai et al. 2023)

Degradation of language reasoning performance in MLLMs has been observed in a limited number of prior studies. MLLMs such as DeepSeek-VL (Lu et al. 2024) and Kosmos-1 (Huang et al. 2023) have been compared to their corresponding base LLMs on text-only tasks, with mixed results showing that MLLMs can perform better or worse depending upon the benchmark. Zhang et al. (2024) compare Vicuna (Zheng et al. 2024) and Qwen (?) LLMs to their MLLM counterparts trained with different vision encoders, finding varying degrees of language reasoning degradation. The use of interleaved image-text data as well as text-only examples when training MLLMs has been shown to help mitigate performance degradation on language tasks (McKinzie et al. 2024). In contrast to these prior studies, our work investigates how the choice of the base LLM influences the phenomenon of language reasoning degradation when MLLMs are trained. Whereas previously proposed methods for mitigating this effect rely on introducing new data our modules during training, we propose a simple model merging approach which can effectively recover performance in text-only tasks without requiring any additional training.

Model merging is a popular and promising technique to combine the strengths of different models through aggregation of their parameters, allowing for improved generalization and performance. The task arithmetic framework (Ilharco et al. 2022) attempts to combine the capabilities of multiple (possibly fine-tuned) models without catastrophic forgetting. This is done by computing task vectors – the differences in weights between models, then adding or subtracting these task vectors to an initial set of parameters to induce learning or forgetting with respect to a given direction in weight space. The TIES approach (Yadav et al. 2023) builds upon this approach by only considering the largest entries in the task vector as candidates for merging, and uses a sign-consensus algorithm to reduce task interference. In our setup, we compute the task vector representing natural language proficiency (a Mistral LLM), and add it to the instruction tuned LLM to recover any degraded natural language performance caused by visual instruction tuning. In our experiments, we investigate to what degree model merging is effective for Mistral-based MLLMs by scaling the contribution of the task vector to the merged model.

Experiments

Analyzing Language Reasoning Degradation in LLaVA Models

Experimental Details. We focus our analysis on three MLLMs sharing a common architecture: LLaVA-1.5, LLaVA-1.6, and LLaVA-1.6-Mistral. LLaVA-1.5 and LLaVA-1.6 are both derived from the Vicuna-1.5 LLM, but LLaVA-1.6 supports higher resolution images and was trained on an improved visual instruction tuning dataset. LLaVA-1.6 and LLaVA-1.6-Mistral are identical except that the latter was derived from the Mistral LLM (Jiang et al. 2023). Our choice of these MLLMs is motivated by the desire to determine whether differences in language degradation can be attributed

to the choice of the base LLM or the datasets used for visual instruction tuning. We utilize the language model evaluation harness framework (Gao et al. 2024) to evaluate LLM and MLLM performance on 8 language datasets: ARC-E (Clark et al. 2018), Race-H (Lai et al. 2017), OpenBookQA (Mihaylov et al. 2018), GSM8k (Cobbe et al. 2021), Hellaswag (Zellers et al. 2019), MMLU (Hendrycks et al. 2021), CommonsenseQA (Talmor et al. 2018), and Winogrande (Sakaguchi et al. 2021). We keep the evaluation strategy fixed across models for each dataset, where we prompt each model in a zero-shot fashion without chain of thought. The only exception is GSM8K, which is evaluated as 8-shot with chain of thought.

For vision tasks, we evaluate each MLLM on five different visual tasks: GQA (Hudson and Manning 2019), MMBench (Liu et al. 2023), VizWizVQA (Gurari et al. 2018), and the Perception and Cognition tasks from the MME benchmark (Fu et al. 2023). For MMBench, we use only its English subset as the test set. We evaluate model performance using the official metrics for each dataset. To perform the evaluation, we make use of the open-source `lmms.eval` library. We keep the evaluation strategy fixed across models for each dataset. For merged models, we reattach the standard vision encoder to recreate a complete VLM before evaluation.

We utilize the corresponding evaluation metrics as implemented in the language model evaluation harness and `lmms.eval` for language and vision tasks respectively. To evaluate the language reasoning performance of MLLMs in the absence of visual input, we extract the LLM weights and evaluate the resulting model in an identical fashion to the Vicuna and Mistral base LLMs.

Results. Figure 1 (a) compares the performance of MLLMs to their corresponding LLMs on the eight language reasoning tasks. We can see that LLaVA-1.6-Mistral scores lower than Mistral across all tasks except CommonsenseQA, whereas LLaVA-1.5 and LLaVA-1.6 perform similar to or better than Vicuna on most tasks except GSM8k. Notably, Mistral outperforms Vicuna, suggesting that stronger LLMs may experience more language degradation after visual instruction tuning. LLaVA-1.5 and LLaVA-1.6 performed similarly, indicating that choice of the base LLM has a stronger influence on language degradation than differences in visual instruction tuning datasets. Two additional commonalities are evident in these results. First, all MLLMs exhibit significant performance degradation on the GSM8K math reasoning dataset. Second, all MLLMs significantly outperform their corresponding LLM on CommonsenseQA. This indicates language reasoning degradation is not uniform across tasks, and that performance in some domains (e.g., commonsense reasoning) may actually improve after visual instruction tuning. We posit that some text-only tasks benefit from visual instruction tuning when the required reasoning type requires visual understanding of the world. In contrast, other tasks which are unrelated to visual comprehension are more likely to experience degradation.

Human Evaluation of CommonsenseQA. To the best of our knowledge, prior studies which have observed language degradation in MLLMs have not investigated the inverse

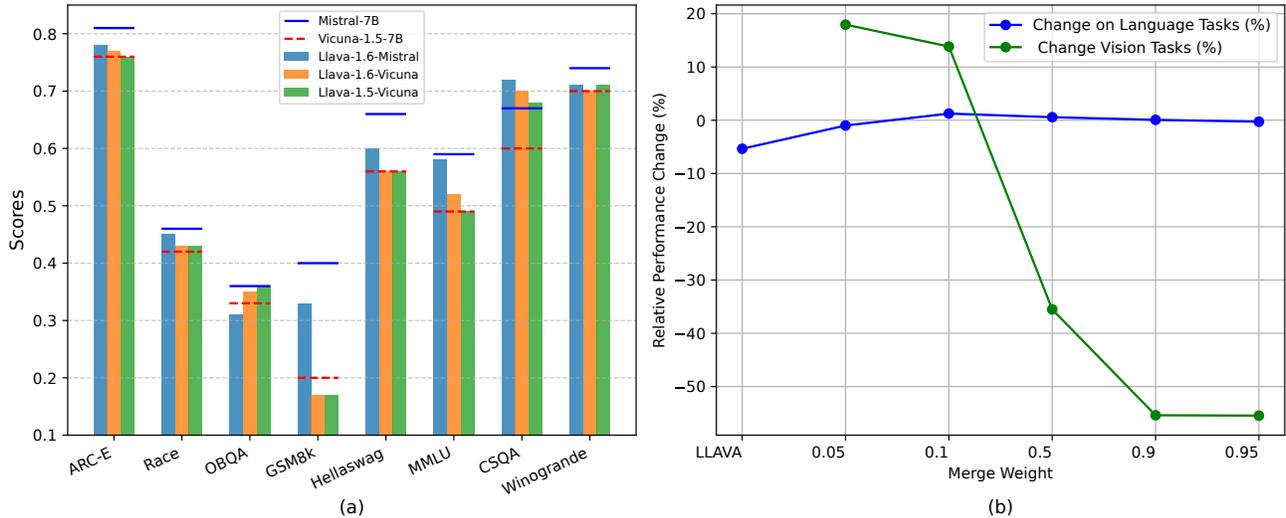


Figure 1: (a) Comparison of the performance of Mistral and Vicuna LLMs on eight language reasoning tasks with their corresponding LVLMs. (b) Average percentage of performance changes across 8 language tasks and 5 vision tasks using different merging weights.

effect we found for the CommonsenseQA dataset, where all MLLMs outperform their LLM counterparts. To better understand why this is the case, we analyzed the types of questions for which the MLLM was correct while its corresponding LLM was incorrect. For each MLLM, we sampled 20 questions where it produced a correct answer on CommonsenseQA while the corresponding base LLM did not, resulting in a total of 60 samples. We then categorized these questions into five groups: *commonsense physical locations*, *object-action associations*, *physical appearance and characteristics*, *situational or event-based commonsense*, and *other*. For *commonsense physical locations*, the questions typically involve scenarios where specific actions or objects commonly occur. *Object-action associations* comprises questions about objects and their associated functions or behaviors. The *physical appearance and characteristics* category involves questions about the external features of objects. *Situational or event-based commonsense* samples are questions about common events or scenarios that can be visually represented. We also include *other*: a category for questions where it is difficult to directly link the knowledge involved to visual information, as these questions are often abstract or conceptual. One possible explanation for questions belonging to *other* is that some abstract concepts are indirectly acquired through visual training, which help improve the model’s performance. Our analysis reveals that in 60% of these cases, the relevant knowledge or context can be presented visually, which may explain why multimodal instruction fine-tuning leads to better performance. In Table 1, we provide an example for each category along with a possible visual representation of the relevant knowledge. Our findings suggest that language models can acquire commonsense knowledge not only through text-based pretraining but also through visual information, which can then be applied to text-only tasks, particularly for knowledge that can be conveyed both textually and visually.

Mitigating Language Reasoning Degradation with Model Merging

Mitigating the phenomenon of language reasoning degradation is important. Prior studies have proposed to mitigate language reasoning degradation through various training strategies, such as interleaved image-text data, text-only examples, and extra modules to compensate for attention shifts. In contrast, we explore a simpler model merging strategy which requires no additional training.

Model Merging Overview There are multiple well-studied ways to combine model parameters, many of which utilize the task-arithmetic framework that specifies a set of fine-tuned models to be merged back into a base model. One straight-forward approach is to linearly interpolate between the parameter vectors, weighted by some hyperparameter α . Another approach is to consider that the differences between a fine-tuned model and a base model can be attributed to knowledge about the fine-tuning task. Thus, the difference between the fine-tuned model and the base model can be treated as a vector that points in the direction of a given task, this is the essence of the task arithmetic framework. In this work we leverage a task-arithmetic inspired method called TIES (Yadav et al. 2023). The TIES merging algorithm consists of four steps: computing the task vector, TRIM, ELECTSIGN, and MERGE. First, a task vector is computed as the difference between the fine-tuned model and the base model. In the TRIM step, redundant parameters are pruned by selecting the top-k entries in the task vector with the highest magnitudes, setting the rest to zero. The ELECTSIGN method resolves any sign conflicts with respect to a specific parameter within the set of task vectors. One option for ELECTSIGN is to perform a majority vote. In our case, we have only one task vector, and the ELECTSIGN step is skipped. Finally, MERGE is performed as a weighted sum of the task vector(s) as the base model. In our setup, we only retain the TRIM and MERGE steps

Category	%	Question	Visual Representation
Commonsense Locations	Physical 33%	Where do cars usually travel at very high speeds?	The inside of a car, showing a highway and the speedometer.
Object-Action	Associa- 10%	Some food can be stored at room temperature until you open it, then you should keep it in what?	Canned food being placed in a refrigerator after opening.
Physical Appearance and Characteristics	7%	A statue that shoots liquid is called a what?	A fountain statue with water flowing from it.
Situational or Event-based Common Sense	10%	Joe’s cat smelled something delicious and jumped into this, causing him to panic and fear for its life. Where might it have jumped?	A kitchen stove or an oven, with the cat in a risky situation.
Others	40%	What does someone have that causes them committing murder?	N/A

Table 1: Examples of question categories with potential visual representations.

from TIES, with ELECTSIGN being a noop. In our case, to merge the parameters of the Mistral base LLM (θ_{llm}) into the LLaVA-Mistral VLM language model (θ_{vlm}), we compute a task vector T defined as $T = \theta_{vlm} - \theta_{llm}$. The task vector captures the relevant differences between the two models, but may also contain information about parameters that have been adapted to handle vision-related tasks, which we want to preserve. We hypothesize that the largest entries of T correspond to parameters that were critical for language modeling but degraded due to visual instruction tuning. Therefore, in our experiments we mask the bottom- $K\%$ of entries of T , taking only the top- K entries as candidates for merging; we denote the pruned task vector as \hat{T} . Finally, the task vector is combined with θ_{vlm} as: $M = \alpha\hat{T} + \theta_{vlm}$. In our experiments we vary both α and K , the results can be seen in figure (fig. 1). For all experiments involving model merging, we utilized the Mergekit library (Goddard et al. 2024) and selected the TIES method (Yadav et al. 2023) to perform the merge. We focus our investigation of model merging on LLaVA-1.6-Mistral, as it exhibited the greatest and most consistent language reasoning degradation relative to its base LLM. Figure 1 (b) shows the result of utilizing increasing amounts of merging between this MLLM and its base LLM. As the weight proportion (x-axis) increases, more weight is being given to the base LLM during the model merging. As expected, we observe that performance on language tasks among merged models approaches that of the base LLM as the weight proportion increases. In contrast, increasing weight proportion decreases the performance of the merged models on visual reasoning tasks, as the MLLM is deviating further from its original state after visual instruction tuning.

These results show that the merging weight proportion can be tuned to optimally balance visual reasoning capabilities and performance in text-only tasks. This hyperparameter can be set based on the targeted use cases for the MLLM to optimize desired performance, without needing to perform any additional training of the model. Our results show that smaller weight proportions (e.g., 0.1) can effectively recover most of the degraded performance across language reasoning tasks without significantly disrupting the MLLM’s visual reasoning capabilities. Surprisingly, we find that this amount

of model merging actually increases performance on three out of the five visual reasoning tasks relative to the original LLaVA-1.6-Mistral model. We also find similar improvements in visual reasoning capabilities when smaller weights are used to merge LLaVA-1.5 and LLaVA-1.6 with Vicuna. We hypothesize that multimodal tasks require language reasoning in addition to visual perception, which explains why performance can improve when merging the MLLM with its corresponding LLM. These results demonstrate the benefits of merging for preserving language reasoning and multimodal capacity.

Conclusion

Our study reveals how multimodal instruction tuning of foundation models can lead to undesired language reasoning performance degradation. We observed that choice of the base LLM prior to visual instruction tuning is more significant than differences in training datasets in influencing the phenomenon of language reasoning degradation, and that stronger LLMs experience a greater degree of degradation. Moreover, language degradation is not uniformly exhibited across datasets, with certain tasks such as commonsense reasoning actually exhibiting the inverse effect. We proposed a simple training-free model merging strategy which can effectively counteract language degradation in MLLMs, offering the ability to customize the balance between language & visual reasoning performance without requiring any additional training. We believe this points to model merging as a promising direction for future research on mitigating undesired performance regressions.

References

- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.
- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved

- question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.
- Gao, L.; Tow, J.; Abbasi, B.; Biderman, S.; Black, S.; DiPofi, A.; Foster, C.; Golding, L.; Hsu, J.; Le Noac'h, A.; Li, H.; McDonell, K.; Muennighoff, N.; Ociepa, C.; Phang, J.; Reynolds, L.; Schoelkopf, H.; Skowron, A.; Sutawika, L.; Tang, E.; Thite, A.; Wang, B.; Wang, K.; and Zou, A. 2024. A framework for few-shot language model evaluation.
- Goddard, C.; Siriwardhana, S.; Ehghaghi, M.; Meyers, L.; Karpukhin, V.; Benedict, B.; McQuade, M.; and Solawetz, J. 2024. Arcee's MergeKit: A Toolkit for Merging Large Language Models. *arXiv preprint arXiv:2403.13257*.
- Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3608–3617.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multi-task Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O. K.; Patra, B.; et al. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36: 72096–72109.
- Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6700–6709.
- Ilharco, G.; Ribeiro, M. T.; Wortsman, M.; Gururangan, S.; Schmidt, L.; Hajishirzi, H.; and Farhadi, A. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. *arXiv preprint arXiv:1704.04683*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, 740–755. Springer.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Sun, Y.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.
- McKinzie, B.; Gan, Z.; Fauconnier, J.-P.; Dodge, S.; Zhang, B.; Dufter, P.; Shah, D.; Du, X.; Peng, F.; Weers, F.; et al. 2024. Mml: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.
- Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.
- Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.
- Wang, Y.; Chen, W.; Han, X.; Lin, X.; Zhao, H.; Liu, Y.; Zhai, B.; Yuan, J.; You, Q.; and Yang, H. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.
- Yadav, P.; Tam, D.; Choshen, L.; Raffel, C.; and Bansal, M. 2023. Resolving interference when merging models. *arXiv preprint arXiv:2306.01708*, 1.
- Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.
- Zhang, Y.-K.; Lu, S.; Li, Y.; Ma, Y.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; Zhan, D.-C.; and Ye, H.-J. 2024. Wings: Learning Multimodal LLMs without Text-only Forgetting. *arXiv preprint arXiv:2406.03496*.
- Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.