CS2916 Spring 2024 Project: What matters when aligning multimodal models?

Yanheng He
Student ID: 522120910073
Shanghai Jiao Tong University

Jiahe Jin
Student ID: 522031910287
Shanghai Jiao Tong University

Yuxuan Zhang
Student ID: 522030910005
Shanghai Jiao Tong University

Abstract

Alignment is crucial for multimodal models to ensure their outputs meet human needs and expectations. This study explores the relationship between the performance of the multimodal model LLaVA on the Science QA dataset and the quality of alignment data, and provides some conjectures on how fine-tuning affects model performance. By replicating the fine-tuning and evaluation of LLaVA, we assessed the model's performance after vision-language pre-training and subsequent alignment using different fine-tuning datasets. The reformatted dataset led to improved model performance, while the shorter data showed that fine-tuning not only shapes the model's adherence to response formats, but also influences and its answering abilities. The study also reveals some limitations of the Science QA dataset that affect generalization. Code and data are accessible at GitHub.[1]

1 Introduction

Alignment is essential for multimodal models to ensure their outputs meet human needs and expectations. This requires fine-tuning the models to interpret accurately to diverse data types like text, images, and audio, and finally use its knowledge to make satisfactory response. High-quality alignment depends on high-quality training data that reflects human intentions, demands, and guidance on models.

Our work explores the relationship between the performance of the multimodal model LLaVA on the Science QA dataset and the quality of alignment data. We first replicates the fine-tuning and evaluation of LLaVA on the Science QA multiple-choice question dataset, and evaluates the model's performance.

Before fine-tuning, LLaVA underwent a vision-language alignment pre-training stage, leveraging image-text pairs to align visual features with the language model's word embedding space, thereby acquiring extensive multimodal knowledge.

We aim to explore the effects of different fine-tuning datasets on the changes in the model's capabilities after alignment. We conducted two experiments based on the Science QA dataset: one using reformatted training data with rewritten chain-of-thought (CoT) process, and another using short training data that only provided the final choice. We realigned the pre-trained model using these two types of training data and performed subsequent evaluations and analyses.

In the end, the model aligned on our reformatted training set achieved performance improvement compared to the model aligned on the original training set, and the experiments with short alignment data revealed that the fine-tuning process not only teaches the pre-trained model to "follow the response format" but also significantly affects its answering capabilities.

Through this process, we seek to deepen our understanding of multimodal model alignment. Additionally, we identified some limitations of the Science QA dataset, such as the high similarity between its training and test sets, which hinders the model's ability to generalize effectively.

2 Literature Survey

2.1 Visual Instruction Tuning

LLaVA (Liu et al., 2023) is an outstanding end-to-end trained large multimodal model that connects a vision encoder and a large language model (LLM) for general-purpose visual and language understanding. The feature alignment

[1]The code repository can be found at https://github.com/LLM-class-group/Multimodal-Learning.

process is crucial within LLaVA's framework, utilizing a projection matrix to map visual features into the word embedding space of the language model. This mapping ensures that multimodal inputs are coordinated within the model's processing architecture, allowing for consistent understanding of the input data.

LLaVA demonstrates impressive multimodal chat abilities, sometimes exhibiting the behaviors of multimodal GPT-4 on unseen images and instructions. When fine-tuned on Science QA, the synergy of LLaVA and text-only GPT-4 achieves a new state-of-the-art accuracy of 92.53%.

Our work aims to further enhance LLaVA's performance on Science QA by improving the quality of aligned data.

2.2 Reformatted Alignment

Reformatted Alignment (Fan et al., 2024) is a simple and effective method for improving the quality of instruction data while minimizing human annotation and reducing the risk of factual errors. This approach begins with a standard definition stage, where humans express their preferences for response formats in natural language. It then employs LLM and retrieval augmented generation (RAG) to reformat instruction data, ensuring that the responses are both structured and well-founded. Its success across various benchmarks underscores the importance of careful data formatting and the potential of systematic reformatting methods to better align LLMs with human values.

In our work, we employed a method similar to reformatted alignment for multimodal instruction data. By reformatting the Science QA training set according to human-specified format preferences, we successfully improved model performance.

2.3 Science QA

Science QA (Lu et al., 2022) is a scientific question dataset designed to enhance machine learning models' abilities to understand complex scientific questions and generate explanatory responses. It is a large and diverse dataset, containing 21,208 instances and 9,122 unique questions. Science QA extends beyond traditional natural sciences, uniquely encompassing social sciences and linguistics, covering 26 topics, 127 categories, and 379 knowledge skills.

The dataset provides contextual information for many questions to enhance model understanding. Specifically, 48.7% of the questions integrate visual context (i.e., relevant images), 48.2% incorporate textual context, and 30.8% include both visual and textual contexts. This multimodal approach enhances the dataset's applicability in real-world scenarios, as scientific understanding often requires synthesizing various types of information.

A notable contribution of Science QA is its detailed annotations for answers, with 83.9% of questions having background knowledge annotations (lectures), and 90.5% of questions accompanied by detailed explanations of the answering process. These comprehensive responses provide rich background knowledge and problem-solving processes, enabling models to learn to articulate their reasoning, thereby improving their answer quality.

Question:
Identify the question that Kurt's experiment can best answer.
Choices:
A.Do more bacteria grow in liquid with cinnamon than in liquid without cinnamon?
B.Does temperature affect how much bacteria can grow in liquid?
Answer:
A
Hint:
The passage below describes an experiment. Read the passage and then follow the instructions below.
Kurt mixed bacteria into a nutrient-rich liquid where the bacteria could grow. He poured four ounces of the mixture into each of ten glass flasks. In five of the ten flasks, he also added one teaspoon of cinnamon. He allowed the bacteria in the flasks to grow overnight in a 37 ° C room. Then, Kurt used a microscope to count the number of bacteria in a small sample from each flask. He compared the amount of bacteria in the liquid with cinnamon to the amount of bacteria in the liquid without cinnamon.
Figure: flasks of liquid for growing bacteria.
Category:
Designing experiments
Skill:
Identify the experimental question
Lecture:
Experiments can be designed to answer specific questions. How can you identify the questions that a certain experiment can answer? In order to do this, you need to figure out what was tested and what was measured during the experiment.
Imagine an experiment with two groups of daffodil plants. One group of plants was grown in sandy soil, and the other was grown in clay soil. Then, the height of each plant was measured.
First, identify the part of the experiment that was tested. The part of an experiment that is tested usually involves the part of the experimental setup that is different or changed. In the experiment described above, each group of plants was grown in a different type of
...
Examples of questions that this experiment can answer include:
Does soil type affect the height of daffodil plants?
Do daffodil plants in sandy soil grow taller than daffodil plants in clay soil?
Are daffodil plants grown in sandy soil shorter than daffodil plants grown in clay soil?
Solution:

Figure 1: Example of Science QA

3 Baseline Reproduction

3.1 Instruction fine-tuning data generation

We replicated the method used in the LLaVA paper to construct the fine-tuning data (*full-origin* dataset) for Science QA. As shown in 2, "question", "hint", and "choice" from the original Science QA dataset was concatenated to form the user's query, while the "lecture", "solution", and "answer" are concatenated to form the model's response. The "lecture" section is intended to include concepts and background knowledge relevant to the question, sometimes supplemented with examples, with the aim of promoting a coherent chain-of-thought process in the model's response generation, thereby improving the accuracy of the answers.

However, we observed that in the Science QA dataset, the "lecture" is only associated with the task category and required skills, which means that the "lecture" may be the same for all questions sharing the same "skill". This can result in the inclusion of a substantial amount of knowledge that is not directly related to the specific question. Such a design might mislead the model's response, as it could rely on background information that is irrelevant to the question itself, potentially reducing the accuracy and relevance of the model's answers.

For example, within the dataset, we find two problems that, despite their distinct subjects, share the same categorization under the "skill" of "Identify the experimental question". One problem delves into the correlation between the landing time of a ping pong ball and the angle of its projection, while the other examines how the material of a snowboard affects sliding duration. Due to their shared skill classification, both questions are paired with an identical "lecture" that, notably, includes an extensive example concerning the relationship between soil conditions and the growth of daffodils. This example, although it may enrich the dataset with general scientific inquiry methods, does not directly relate to the specifics of the physical experiments being inquired about. The inclusion of such an example could potentially lead the model astray, causing it to incorporate irrelevant information into its reasoning process, which may ultimately affect the precision and relevance of the responses generated for the given tasks.

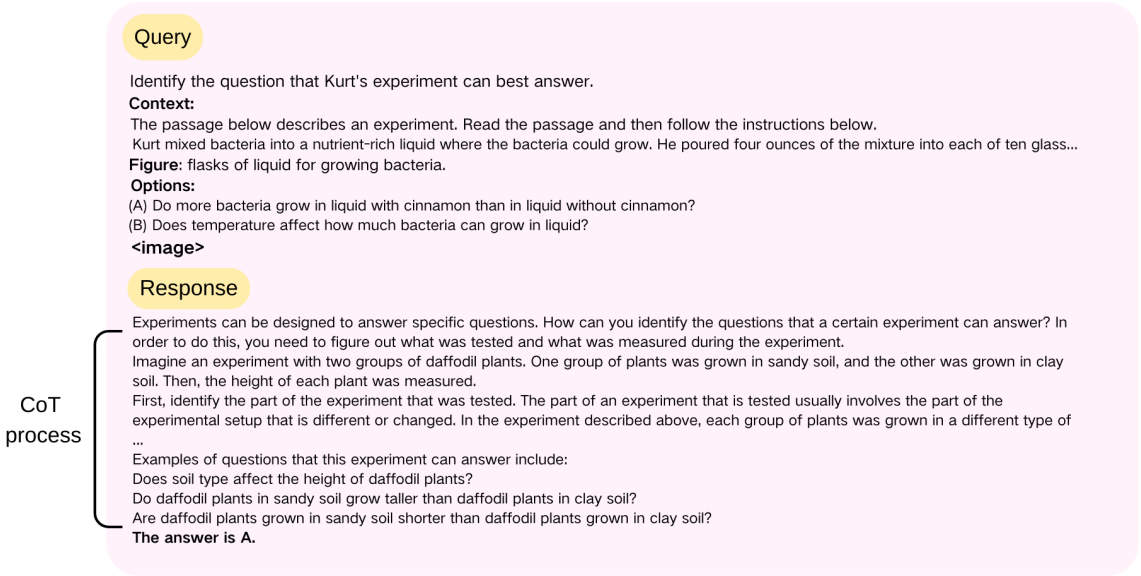

Figure 2: Example of *full-origin* dataset

3.2 Fine-tuning and Evaluation

We used the same hyperparameters and fine-tuning method (freezing the visual encoder) as in the original paper to fine-tune the pre-trained model provided by LLaVA. Additionally, we applied DeepSpeed's ZeRO-3 offloading technique to reduce GPU memory usage. The fine-tuning process took approximately 10 hours on four A800 GPUs. The fine-tuned model achieved an accuracy of 91.15% on the Science QA test set, which is similar to the results reported in the original paper. Minor differences might be attributed to the temperature settings during the answer generation process.

4 Exploring Reformatted Alignment Method

4.1 Preliminary Attempt

As an initial attempt, we employed the Reformatted Alignment method to reformat the *full-origin* dataset. First, we utilized the Google Search API for retrieval augmentation. Subsequently, GPT-4o was tasked with rewriting the evidence obtained from Google Search, along with the responses in *full-origin* dataset, following the specified template we provided. Through this process, we reformatted all of the 12,726 responses in the *full-origin* dataset and built the *full-realign-preliminary* dataset.

Subsequently, we fine-tuned pre-trained LLaVA on *full-realign-preliminary* dataset and evaluated for Science QA. Unfortunately, the results were not promising, with an accuracy of only 87.83%, which is lower than the original performance of LLaVA (fine-tuned on *full-origin*). Upon examining the model's evaluation results and *full-realign-preliminary* dataset, we identified several deficiencies in *full-realign-preliminary* dataset:

1. The retrieval augmentation using the Google API did not provide further valuable information, as the Science QA lectures already offered sufficient background information, and GPT-4o had an adequate understanding of the topic;

2. For non-inference questions that CoT process is unnecessary, our template led the model to produce unnecessarily lengthy CoT process, which caused the fine-tuned model to perform poorly on such questions;

3. The responses generated by GPT-4o did not strictly follow our template and sometimes failed to stop appropriately.

4.2 Experiment Setup

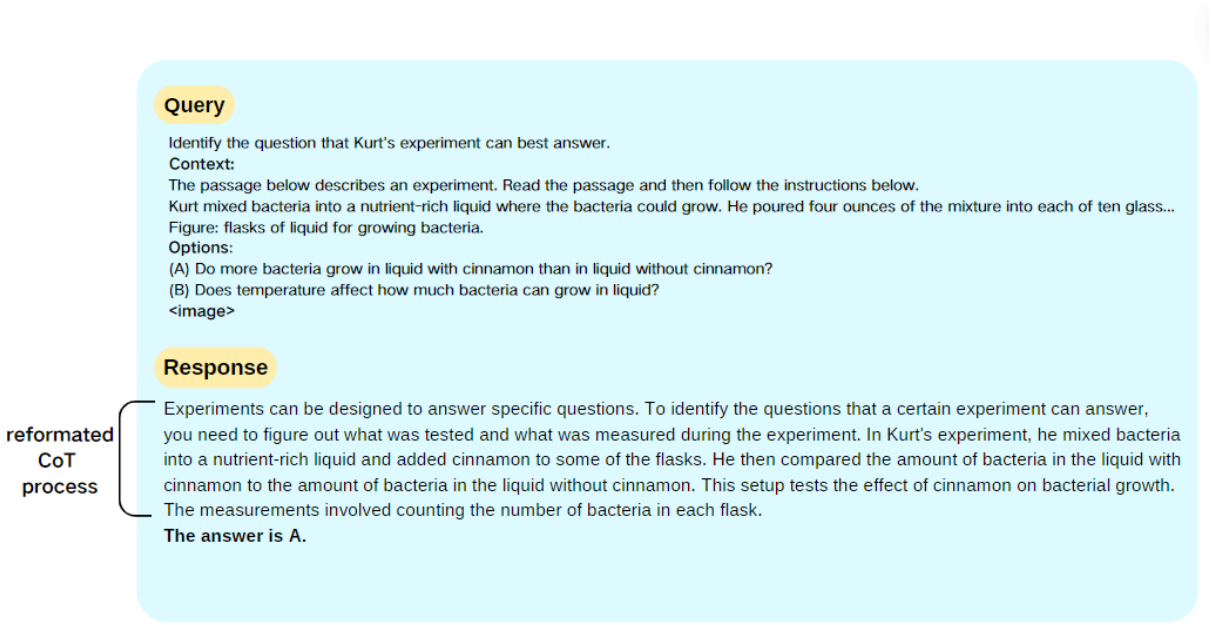

Figure 3: Example of *mini-realign* dataset

Based on the above observations, we made several adjustments and conducted further experiments. We randomly selected 1,000 problems from *full-origin* dataset and construct *mini-origin* dataset, and the reformatting for this stage was conducted on it.

Specifically, we made the following enhancements:

1. We abandoned retrieval augmentation, and asked GPT-4o to assess the reasonableness of the CoT process of original Science QA responses. For unreasonable answers, GPT-4o was tasked with rewriting the solution;

2. We determined that guiding the model to output a CoT process is only meaningful for questions requiring reasoning. Therefore, we used the length of the original Science QA answer to judge whether the question required reasoning simply, and only reformatted answers with lengths exceeding 200 characters;

3. We prompted the model to provide more concise analysis, reducing unnecessary explanations, while strictly adhering to the template format for answering questions.

   Based on these enhancements, we built *mini-realign* dataset, as shown in figure 3.

4.3 Results

We evaluated the model fine-tuned on *mini-realign* dataset and achieved an accuracy of 75.48%, successfully surpassing the model fine-tuned on *mini-origin* dataset, as shown in Table 1.

Through analyzing the model's performance, we attribute the improvement to the following refinements:

1. For questions requiring complex reasoning, the fine-tuning guided the model to produce a structured and concise CoT process, preventing it from getting bogged down in lengthy analysis and kept repeating.

2. For certain types of questions, Science QA provides a standard response process rather than an CoT process directly related to the specific question. Our reformatted alignment required GPT-4o to rewrite all such responses that were not accurately related to the content of the question. This ensured that for all reasoning questions, there was a structured, concise, and accurately related CoT process that led to the correct result.

Table 1: Accuracy on Science QA before and after reformatted alignment

train set	accuracy	question count
full-origin	**91.15**%	12,726
full-realign-preliminary	87.83%	
mini-origin	74.75%	1000
mini-realign	**75.48**%	

5 Exploring the effect of short alignment data

To further explore the impact of aligning multimodal models, we constructed special short alignment datasets. For each question in origin Science QA train set, we discarded all the CoT process and simply answered with "The answer is [choice]." We anticipate that this shortened dataset will primarily teach the model the answer format while introducing minimal additional knowledge.

5.1 Experiment Setup

As mentioned, we discarded the CoT process in all problems from the *full-origin* and mini-origin datasets to create the *full-short* and *mini-short* datasets, respectively. We simply inserted the options from Science QA into the format "The answer is [choice]." to build short responses. Subsequently, we fine-tuned the pre-trained LLaVA and evaluated the model for Science QA.

5.2 Results

The evaluation results are shown in Table 2. We observed that the *full-short* dataset achieved performance closely matching that of the *full-origin* dataset (the original dataset used to fine-tune LLaVA on Science QA). However, when the dataset size is reduced to 1/12, the training effectiveness of the *mini-short* dataset is disappointing, yielding an accuracy of only 69.91%, which is approximately 5% lower than the *mini-origin* dataset. Our analysis indicates that the *mini-short* dataset alone is sufficient to teach the model to answer questions in correct format. The significant performance difference between *full-short* and *mini-short* suggests that the fine-tuning process is not merely about training the model to answer questions in a specific format. Hence, the outstanding performance of LLaVA on Science QA derives not only from the pre-training process but also significantly from the fine-tuning on the *full-short* dataset.

6 Conclusion

This paper investigated the impact of two different alignment method (reformatted and short) on the performance of the pre-trained LLaVA model on the Science QA dataset. By using the reformatted Science QA training set, we successfully improved model performance, demonstrating the effectiveness of fine-tuning with high-quality data. Through experiments with the short alignment data, we discovered that LLaVA's outstanding performance in Science QA does not primarily stem from the pre-training process. Additionally, we identified a limitation of the Science QA dataset: its training and test sets have is highly similar. This similarity diminishes the significance

Table 2: Accuracy on Science QA of short alignment data

train set	accuracy	question count
full-origin	91.15%	12,726
full-short	90.57%	
mini-origin	74.75%	1000
mini-short	69.91%	

of the rich background information provided by Science QA, as a short alignment dataset constructed solely from question answers is sufficient to achieve excellent results. We hope our work will inspire future research on aligning multimodal models.

Acknowledgements

We extend our sincere thanks to Ethan Chern and Pengfei Liu for valuable discussions on our work. We are also deeply appreciative of the computing resources provided by GAIR.

References

[1] Run-Ze Fan, Xuefeng Li, Haoyang Zou, Junlong Li, Shwai He, Ethan Chern, Jiewen Hu, and Pengfei Liu. Reformatted alignment. *arXiv preprint arXiv:2402.12219*, 2024. URL https://arxiv.org/abs/2402.12219.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

[3] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.