



# SCITUNE: Aligning Large Language Models with Human-Curated Scientific Multimodal Instructions

Sameera Horawalavithan, Sai Munikoti, Ian Stewart, Henry Kvinge, Karl Pazdernik  
Pacific Northwest National Laboratory, Richland, WA

## Abstract

Instruction finetuning is a popular paradigm to align large language models (LLM) with human intent. Despite its popularity, this idea is less explored in improving LLMs to align existing foundation models with scientific disciplines, concepts and goals. In this work, we present *SciTune* as a tuning framework to improve the ability of LLMs to follow multimodal instructions generated from scientific publications. To test our methodology, we train a large multimodal model LLaMA-SciTune that connects a vision encoder and LLM for science-focused visual and language understanding. LLaMA-SciTune significantly outperforms the state-of-the-art models in the generated figure types and captions in SciCap and VisText benchmarks. In comparison to the models that are finetuned with synthetic data only, LLaMA-SciTune surpasses human performance on average and in many sub-categories on the ScienceQA benchmark. Our results demonstrate that human-generated scientific multimodal instructions remain highly valuable in tuning LLMs to perform well on science tasks, despite their lower volume and relative scarcity compared to synthetic data. We publicly release the SciTune codebase<sup>1</sup>.

## 1 Introduction

Instruction finetuning has gained significant traction in the NLP community as a means of enhancing the capabilities of large language models (LLMs), allowing them to accurately balance desired outcomes, context, and human preferences, leading to more relevant and coherent responses. More recently, AI assistants have been trained to comprehend and execute multimodal vision-and-language instructions, aligned with human intent, to accomplish diverse real-world tasks in complex multimodal environments. In one of the latest developments, MiniGPT-4 (Zhu et al.,

2023), LLaVA (Liu et al., 2023) and LLaMA-Adapter (Gao et al., 2023) have focused on expanding language-only instruction models to incorporate multimodal capabilities, thereby granting LLMs the ability to perform visual grounded reasoning tasks.

Recent research suggests that high-quality multimodal pretraining data and instructions, such as high-resolution images and diverse multimodal data are important for effective model performance (McKinzie et al., 2024). However, one of the primary bottlenecks is the scarcity of high-quality data for multimodal pretraining and instruction tuning. To mitigate the challenges associated with data scarcity, many recent multimodal models rely on synthetically generated instructions (training data distilled from responses from other models) for fine-tuning instead of human annotations (Liu et al., 2023; Cascante-Bonilla et al., 2023; Bai et al., 2022).

However, using synthetic data to align AI models can create confusion and uncertainty, since synthetic data, being artificially generated, often fails to capture the complexities of human values (Liu et al., 2024; Zhou et al., 2024). This can cause AI models to learn from biased (Feng et al., 2023; Liu et al., 2021), ungrounded (Liu et al., 2022; Patel and Pavlick, 2022) or inaccurate data (Ji et al., 2023; Weidinger et al., 2021). Additionally, training models on recursively generated data can lead to a loss of true data distribution, resulting in less varied and misaligned outputs (Shumailov et al., 2024). Furthermore, models tuned with synthetic data often fail to meet the standards required by certain scientific subdomains, such as medicine (Li et al., 2023; Xia et al., 2024). For example, several open-source medical vision-language models such as LLaVA-Med (Li et al., 2023), MedFlamingo (Moor et al., 2023), MedVInT (Zhang et al., 2023b), and RadFM (Wu et al., 2023) failed to satisfy the trustfulness, fairness, safety, privacy,

<sup>1</sup><https://github.com/pnnl/scitune>

and robustness conditions in the recently introduced CARES (Xia et al., 2024) benchmark. As a result, AI systems relying on synthetic data might behave unpredictably and could potentially cause unintended or harmful outcomes (Anderljung et al., 2023; Zou et al., 2023).

We believe this is mainly due to the absence of alignment methods designed to synchronize existing foundation models with scientific disciplines, concepts, and goals and to ensure that the generated content meets the standards and expectations of the scientific community. Our hypothesis is that scientifically aligned multimodal models can learn from unique patterns and structures present in scientific language generated by human scientists, thus would be able to follow precise instructions about complex procedures, protocols, and guidelines in the scientific environments. This leads to the question: *To what extent, can LLMs align solely with human-curated scientific multimodal instructions?*

To this end, we build on top of the LLaVA (Liu et al., 2023) model architecture to perform scientific multimodal instruction tuning (*SciTune*) on top of a decoder-based pretrained LLM and vision encoder. The *SciTune* training method includes two stages for *scientific multimodal instruction tuning*, i) *scientific concept alignment* to learn across various scientific visual signals (e.g., plots, charts, equation, diagram, etc.), and textual signals (e.g., captions, optical character recognition (OCR) and paragraph mentions), ii) *scientific instruction tuning* to fine-tune on a multimodal scientific reasoning task. To validate our approach, we train our models on top of LLaMA (Touvron et al., 2023a) and the CLIP (Radford et al., 2021) vision encoder model. We show that our model surpasses human performance on the ScienceQA multimodal reasoning benchmark and performs significantly better than state-of-the-art vision-language models in a variety of scientific image understanding tasks. Our results demonstrate that human-curated scientific multimodal instructions remain highly valuable in tuning LLMs to perform well on science tasks, despite their lower volume and relative scarcity compared to synthetic data.

## 2 Methodology

In this section, we describe the SciTune framework in two stages of Scientific Multimodal Concept Alignment and Multimodal Task-specific Instruction Tuning and the design choices and multimodal

architecture used for the experiments.

### 2.1 Scientific Multimodal Instructions

We use an *early-fusion* strategy (Liu et al., 2023) to jointly reason over the text, images, and other modalities with a shared multifaceted representation as presented as SciTune instructions. The SciTune instruction template  $x = (s_D, s_I, s_T)$  includes a system message  $s_D$  to help the model to understand the role and context, instruction  $s_I$  randomly sampled from the visual-grounded questions, and  $s_T$  to encode the multimodal data.

**Human-curated Scientific Instructions** This work solely focuses on multimodal instructions curated by humans instead of machine generated content used in other visual instruction tuned models (Liu et al., 2023; Gao et al., 2023). Our goal is to align the pretrained foundation models with natural scientific concepts and the intentions of humans (scientists). To this end, we chose scientific publications (PDFs) as the medium of scientific instructions that demonstrate various stages of scientific discovery.

We use the SciCap (Hsu et al., 2021) dataset with more than 400,000 scientific figure images extracted from various arXiv papers, including their respective captions and relevant paragraphs. This dataset is composed of arXiv papers from January 2010 to October 2020. It consists eight distinct categories: Computer Science, Economics, Electrical Engineering and Systems Science, Mathematics, Physics, Quantitative Biology, Quantitative Finance, and Statistics. We use the 333,472 examples provided in the SciCap training split for pretraining and use the validation split to evaluate the performance.

We introduce scientific captions ( $s_c$ ), figure types ( $s_t$ ), optical character recognition (OCR) ( $s_o$ ) and paragraph mentions ( $s_m$ ) in the instruction template ( $s_T = \{s_c, s_t, s_o, s_m\}$ ) to convert the SciCap dataset into a multimodal instruction-tuning dataset. Figure-captioning ( $s_c$ ) data typically includes brief text that is highly specific to the associated figure. In contrast, interleaved data such as paragraph mentions ( $s_m$ ) generally features longer and more varied text, which is broadly relevant to the figures it associates. Please see the Appendix (Table 6) for a SciTune instruction sample.

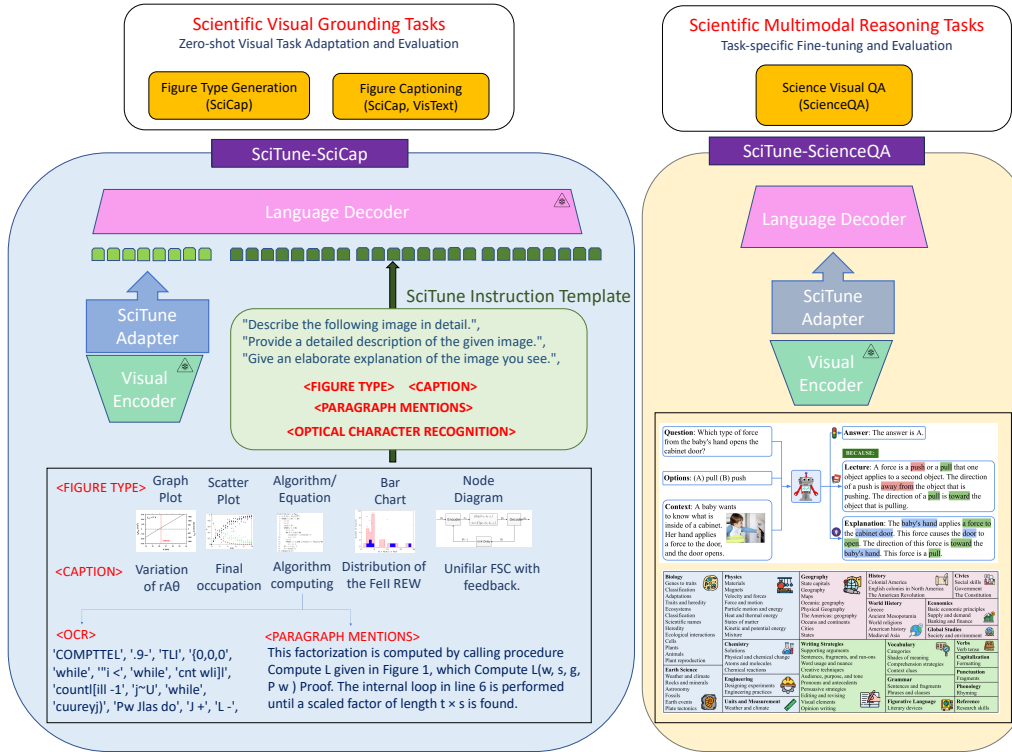


Figure 1: SciTune enables models to learn *scientific concept alignment* across various visual signals (e.g., plots, charts, equation, diagram), and textual signals (e.g., captions, OCR and paragraph mentions); shown in the left graphic in the figure. After alignment, the model can be used to perform various scientific visual grounding tasks (e.g., figure type generation, captioning) with zero demonstrations at inference time (shown in the upper-left graphic). The pretrained model can be further finetuned on a multimodal scientific reasoning task (e.g., ScienceQA; shown in the upper-right graphic).

## 2.2 Multimodal Architecture

**Architecture** We build on top of the most recent multimodal architectures (e.g., LLaVA (Liu et al., 2023), LLaMA-Adapter (Zhang et al., 2023a)) that guide LLMs to follow multimodal instructions. We noticed that adapter-based multimodal training serves as the most efficient technique for injecting multimodal knowledge to a pretrained LLM decoder model. Our goal was to improve the existing LLMs to perform better on science-focused multimodal reasoning and visual grounded tasks. To this end, we chose LLaMA-1 (Touvron et al., 2023a) as the LLM decoder, and CLIP visual encoder (Radford et al., 2021) to experiment with multimodal adapter training as shown in Figure 1.

The SciTune adapter transforms the output of the visual encoder model as inputs to the language decoder with a linear projection layer. While we keep the language decoder and the visual encoder models frozen, the multimodal adapter is updated during the pretraining stage. This modular architecture can be filled by any language decoder and a visual encoder model. We conduct the experiments

with LLaMA 7B and 13B model variants for better comparison with other baseline models. It is worthwhile to note that we chose LLaMA due to its superior performance in the public benchmarks and its open-source accessibility.

We do not use any instruction-tuned LLaMA variants (e.g., Vicuna, Guanaco) in our experiments due to two main reasons. First, we mainly focus on improving the base LLM decoder models with multimodal instructions generated by humans in order to eliminate all confounding factors such as machine generated instruction tuning. Since a majority of instruction-tuned models developed on top of LLaMA are knowledge-distilled from closed-source, proprietary models like GPT-4, we want to avoid any unexpected performance advantages. Second, we want to make a fair comparison with other baseline models proposed in this area developed on top of the base LLaMA model, and test whether the multimodal instruction tuning proposed in this work could lead into better scientific concept understanding compared to those models.

**Training** We model the language distribution  $p(x)$  from a set of SciTune instructions  $(x_1, x_2, \dots, x_m)$  as the product of conditional multimodal token probabilities as shown in Equation 1.

$$p(x) = \prod_{j=1}^n p(s_{T>j} | s_V, s_I, s_{T<j}) \quad (1)$$

We model  $s_V$  with the multimodal tokens after projection from the respective plot visuals  $V$ . We sample the instruction  $s_I$  from the list of questions presented in the Appendix (Table 7). Note that we skip the token descriptors in  $s_T$  for brevity, unless the model is trained autoregressively to generate exact tokens across all textual modalities in  $s_T = \{s_c, s_t, s_o, s_m\}$ . More importantly, the model is able to jointly generate all modality tokens in a single-turn conversation. For example, given a scientific plot and an instruction, the model first generates the figure type (e.g., Graph Plot, Scatterplot, Node Diagram, Equation, Bar Chart), then the visual content through captioning and OCR, and finally the cited paragraph.

As presented in Figure 1, *LLaMA-SciTune-Scicap* is tuned to align the LLM towards scientific concepts. *LLaMA-SciTune-Scicap* can be further finetuned on a multimodal scientific reasoning task. In our experiments, we name the task finetuned model variant as *LLaMA-SciTune-ScienceQA*.

### 3 Experiments

In this section, we report the performance of *LLaMA-SciTune* models across a variety of science-focused downstream tasks. Our goal is to assess the performance of the models in visual grounded language understanding and multimodal reasoning tasks. For example, we want to show how much difference the training stages contribute to the model performance, or whether adding various scientific modalities in the instruction template improves the overall performance. Note that our objective is not to introduce a model that tops the leaderboard across various downstream tasks. Instead, we aim to demonstrate the extent to how human-curated scientific multimodal instructions can be useful in aligning vision-language models.

To this end, we trained three *LLaMA-SciTune-SciCap* models and finetuned them with scientific multimodal reasoning dataset (e.g., ScienceQA) for the corresponding *LLaMA-SciTune-ScienceQA* models. Three *LLaMA-SciTune-SciCap* models differ on the text input types (e.g., Caption, Figure

Type, OCR, and Figure Mentions) and the scale of the LLM (e.g., 7B and 13B) used in the model training. For example, *LLaMA-SciTune-SciCap-13B (CTOM)* model uses the base LLaMA-1 13B checkpoint and figure caption, figure type, OCR, and figure mentions in the training. Similarly, *LLaMA-SciTune-ScienceQA-13B (CTOM)* model is finetuned on the *LLaMA-SciTune-SciCap-13B (CTOM)* with the ScienceQA (Lu et al., 2022) training split. While *LLaMA-SciTune-SciCap* model tunes the LLM to better understand scientific concepts, *LLaMA-SciTune-ScienceQA* is further finetuned for scientific reasoning.

First, we report the performance of *LLaMA-SciTune-SciCap* in two science-focused visual grounded tasks to assess the scientific concept alignment training stage (Section 3.1). Finally, we use the ScienceQA benchmark to test the multimodal reasoning abilities of *LLaMA-SciTune-ScienceQA* across three scientific subject areas (Section 3.2).

#### 3.1 Vision Grounded Tasks Performance

In this section, we report the performance of the *LLaMA-SciTune-SciCap* model for two zero-shot downstream tasks. Note that, we reference the *LLaMA-SciTune-SciCap-13B (CTOM)* model in this performance analysis. In the first task, we evaluate how well the *LLaMA-SciTune-SciCap* model is able to align the associated figure types with the actual image. In the second task, we evaluate the performance of the *LLaMA-SciTune-SciCap* model in generating the figure captions.

##### 3.1.1 Scientific Figure Type Generation

In the scientific concept alignment stage, one of the learning tasks is to align the scientific visuals with the correct figure type. For example, the model should be able to distinguish a graph plot from a scatter plot. We compare the performance of our model of generating the figure types with a standalone vision encoder. For example, we use the CLIP model (Radford et al., 2021) to perform figure type classification in the zero-shot manner given five candidate types (e.g., Graph Plot, Scatterplot, Node Diagram, Equation, Bar Chart). We use the validation data released by the SciCap challenge to perform our experiments. This validation dataset includes plots and the associated figure types. We locate the figure types in the generated SciTune outputs, and compare it with the ground truth. As shown in Table 1, *LLaMA-*

*SciTune-SciCap* shows 57% performance improvement over the standalone CLIP model used in the figure type classification.

It is important to note that the *LLaMA-SciTune-SciCap* used the same CLIP model as the visual encoder, but the additional multimodal adapter was optimized towards aligning figure types with the plots during the pretraining stage. This multimodal adapter is able to project the outputs of vision encoder into the LLM to improve its understanding on the scientific plots. One could argue that a more ideal comparison would be between the *LLaMA-SciTune-SciCap* model and a version of the CLIP model that is specifically tuned with the same dataset, rather than comparing it with the vanilla (untuned) CLIP model. In this experiment, our objective was to evaluate how much the SciTune adapter (Figure 1) contributes to improving the LLM’s performance in understanding scientific plots, as opposed to improvements gained merely through the use of the tuned visual encoder.

Table 1: Accuracy of Generating the Figure Types. We also report the zero-shot figure type classification performance of the CLIP model.

Figure Type	CLIP	SciTune-SciCap
Graph Plot	52.58	<b>93.63</b>
Scatterplot	52.20	<b>70.14</b>
Node Diagram	77.67	<b>95.40</b>
Equation	60.47	<b>89.54</b>
Bar Chart	32.67	<b>80.33</b>
All	55.11	<b>85.81</b>

### 3.1.2 Scientific Figure Captioning

In this section, we test the model performance of generating scientific figure captions given only the scientific plot. Previous works show that scientific figure captioning is an extremely challenging task due to complex image understanding required in vision-to-language modeling (Huang et al., 2023). We take the first sentence in the generated SciTune output as the generated caption. We compare *LLaMA-SciTune-SciCap* model performance with the SOTA image captioning model, BLIP (Li et al., 2022), trained with more than 14M image-text pairs. We use two text evaluation metrics, BLEU and ROUGE, to measure the quality of generated captions with respect to the ground truth captions. We evaluate the models in two scientific image captioning benchmarks, SciCap and

VisText (Tang et al., 2023). We used the validation split with 47639 and 1202 images in two benchmarks, respectively.

As shown in Table 2, the *LLaMA-SciTune-SciCap* model outperforms the BLIP model in both automated text evaluation metrics. This suggests that *LLaMA-SciTune-SciCap* may have a better understanding of the scientific plot in comparison to the BLIP model finetuned towards image captions. Table 8 (see Appendix) shows a few generated captions in comparison to the baseline and ground truth image captions.

Table 2: In-distribution (SciCap) and out-of-distribution (VisText) Evaluation of Generated Figure Captions

Benchmark	Model	BLEU	ROUGE
SciCap	BLIP	0.02±0.02	0.11±0.07
	SciTune-SciCap	<b>0.05±0.03</b>	<b>0.13±0.08</b>
VisText	BLIP	0.06±0.05	0.23±0.11
	SciTune-SciCap	<b>0.10±0.07</b>	<b>0.23±0.12</b>

## 3.2 Scientific Multimodal Reasoning Task Performance

In this section, we evaluate the model performance on science-focused multimodal reasoning question and answering (QA). We report the *LLaMA-SciTune-ScienceQA* model performance in the ScienceQA benchmark (Lu et al., 2022) that includes 21k multimodal multiple choice questions with rich domain diversity across 3 subjects, 26 topics, 127 categories, and 379 skills. We use the ScienceQA training split (12726 examples) to tune the *LLaMA-SciTune-SciCap* model further as shown in Figure 1. Table 3 reports the performance of the models on the ScienceQA test split (4241 test questions). While lectures are shared between training and test splits, there are new questions associated with multimodal contexts, and explanations in the test split. We have three main observations from this table.

First, *LLaMA-SciTune-ScienceQA-13B (CTOM)* model outperforms the human performance on average and in four other sub-groupings. For example, this model records 90.03% accuracy in correctly answering the multimodal reasoning questions in the ScienceQA benchmark, where humans record only 88.40% accuracy. This performance benefit is consistent across social science questions, questions with text or no contexts, and higher-grade questions. More importantly, we noticed that this

Table 3: Results (accuracy %) on ScienceQA dataset. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. We present two variants, *LLaMA-SciTune-ScienceQA (C)* and *LLaMA-SciTune-ScienceQA (CTOM)*. Acronyms inside the parenthesis represent the text inputs used in the SciTune instruction template. E.g., **C**aption, **F**igure **T**ype, **O**CR, and **F**igure **M**entions. We use the notation ♠ to denote the models finetuned with GPT-3.5/4 synthetic instructions, or use GPT-3.5/4 for any support during the inference time. We bold the accuracy values that are greater than what humans achieved.

Method	#Params	Avg	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12
Random Chance	-	39.83	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67
Human Average	-	88.40	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42
UnifiedQA	223M	70.12	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00
UnifiedQA (CoT)	223M	74.11	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82
♠ GPT-3 (Zero Shot)	175B	74.04	75.04	66.59	78.00	74.24	65.74	79.58	76.36	69.87
♠ GPT-3 (CoT) (ALE)	175B	75.17	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68
♠ ChatGPT CoT	175B+	78.31	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03
♠ GPT-4 CoT	1T+	83.99	85.48	72.44	<b>90.27</b>	82.65	71.49	<b>92.89</b>	86.66	79.04
Multimodal-CoT	223M	84.91	87.52	77.17	85.82	87.88	82.90	86.83	84.65	<b>85.37</b>
Multimodal-CoT	770M	<b>91.68</b>	<b>95.91</b>	82.00	<b>90.82</b>	<b>95.26</b>	<b>88.80</b>	<b>92.89</b>	<b>92.44</b>	<b>90.31</b>
♠ LLaMA-Adapter	13B	85.19	84.37	<b>88.30</b>	84.36	83.72	80.32	86.90	85.83	<b>84.05</b>
♠ LLaVa	13B	<b>90.92</b>	<b>90.36</b>	<b>95.95</b>	88.00	89.49	<b>88.00</b>	<b>90.66</b>	90.93	<b>90.90</b>
♠ LLaVa + GPT-4 (judge)	13B	<b>92.53</b>	<b>91.56</b>	<b>96.74</b>	<b>91.09</b>	<b>90.62</b>	<b>88.99</b>	<b>93.52</b>	<b>92.73</b>	<b>92.16</b>
♠ Chameleon (ChatGPT)	175B+	79.93	81.62	70.64	84.00	79.77	70.80	86.62	81.86	76.53
♠ Chameleon (GPT-4)	1T+	86.54	89.83	74.13	<b>89.82</b>	88.27	77.64	<b>92.13</b>	88.03	<b>83.72</b>
SciTune-ScienceQA (C)	7B	85.61	84.36	<b>92.23</b>	82.81	89.56	81.26	<b>88.29</b>	81.28	<b>86.03</b>
SciTune-ScienceQA (CTOM)	7B	86.11	84.50	<b>94.15</b>	82.91	88.35	83.64	<b>88.74</b>	85.05	<b>85.60</b>
SciTune-ScienceQA (CTOM)	13B	<b>90.03</b>	89.30	<b>95.61</b>	87.00	<b>93.08</b>	86.67	<b>91.75</b>	84.37	<b>91.30</b>

model reaches a comparable performance with the LLaVA model, which is trained with synthetic data and twice the size of the training data than what the former model has seen, and in some cases has additional support from GPT-4 during inference.

Second, we noticed that *LLaMA-SciTune-ScienceQA-7B (CTOM)* model performs better than *LLaMA-SciTune-ScienceQA-7B (C)* model pretrained only with captions. For example, CTOM variant (86.11) slightly outperforms C variant (85.11) on average performance and across many other sub-groupings. This suggests the importance of interleaved multimodal data in the scientific concept alignment stage which lifts the downstream task performance over the model tuned only with figure-caption data.

Finally, we noticed a significant performance advantage of the models trained with larger language decoder model (13B) compared to the relatively smaller model (7B). For example, the *LLaMA-SciTune-ScienceQA-13B (CTOM)* model has nearly 5% performance advantage over the 7B model variant. This advantage is 5x bigger than what reported by the LLaVA model when scaled from 7B to 13B (Liu et al., 2023). While this observation suggests that the larger language decoder model helps

to improve the multimodal reasoning performance, we believe it could lead to huge performance benefit with even larger models (LLaMA-65B) when trained with highly-curated scientific multimodal instruction tuning datasets.

**Explanation Performance Analysis** In addition to generating the specific answers to the questions asked, *LLaMA-SciTune-ScienceQA* models also generate a corresponding lecture and explanation for the answers. Please see Figures 3 and 4 and in the Appendix for several examples of generated lectures and explanations. In order to better understand the behavior of generated solution, we manually investigate a few random test examples. Specifically, we picked 50 samples from both the correct and incorrect predictions. We observe that even the correct samples contain a certain amount of incorrect solutions, i.e., around 8% in C and 2% in CTOM version of the 7B models. These results indicate that solution may not always benefit the final answer, and the model is robust to some extent, i.e., it can predict the correct answer even with incorrect rationales. The incorrect solutions are further divided into two major categories, namely commonsense that requires commonsense knowledge such as factual information and count-

ing numbers in the images, and the logical mistakes which shows contradictions in the reasoning. In our experiment, commonsense mistakes are dominant compared to logical, which aligns with previous work (Zhang et al., 2023d). Furthermore, there are cases where solutions are correct in an absolute sense but their final answers are wrong. We also noticed that solutions generated by the CTOM version of the model are more accurate compared to the C version of the model, further emphasizing the importance of multi-modal training with additional scientific modalities. There are certain task categories where our model performs extremely well compared to baselines. In our manual analysis, we found the model is very good with numerical questions, including temperatures and distances, and can answer all topological/map related questions such as "which ocean is highlighted" in the image.

While we observe high performance in aggregate, it is also important to determine whether this performance persists in cases with minimal training examples. We evaluate the performance of the model for questions whose accompanying lectures are only observed a few times in the training data. For these few-shot examples, the model will be less likely to have the exact lecture memorized and ready to use in its generation of the answer, which could lead to lower performance.

We show the model performance on questions for which the lectures were viewed in 5, 10, 25, and 50 times during training, in Table 4. The model performance drops substantially for questions with only 5 or fewer lectures in the training data but quickly recovers after the lecture is viewed at least 10 times. This suggests that the *LLaMA-SciTune-ScienceQA* model doesn't require substantial exposure to a particular type of knowledge to achieve adequate performance. Furthermore, this performance drop is worse for the 7B model as compared to the 13B model, which means that the 13B model is able to learn more quickly from fewer examples or may have more knowledge "baked in" from pretraining that can be leveraged for few-shot examples. Future extensions of the model to other datasets should test performance on completely unseen data, e.g. a more standard VQA dataset not used during training, to determine whether the model is similarly robust in other domains.

### Chain of Thought Reasoning Performance

Outside of the coarse-grained accuracy metric (did the model get the answer right?), we also need to

Table 4: Few-shot performance analysis. We report the number of times lectures seen during the training in frequency, and the number of test questions with the lecture.

Frequency	#Questions	Accuracy (7B)	Accuracy (13B)
5	36	75.00	83.33
10	125	81.60	85.60
25	412	80.34	85.92
50	1140	81.05	86.14

determine whether the model's overall process of reasoning was correct (did the model accurately explain the reasoning that supports the answer?). We investigate the accuracy of the generated text, outside of the answer alone, assessing if the model is able to accurately recover the lecture and the solution that it was trained to generate and to help its reasoning toward the final answer. We report the BLEU and ROUGE scores over all the generated text, separated into the lecture and solution components and compared with the corresponding ground-truth data, e.g. compare the generated lecture component with the ground-truth lecture.

The aggregate results for the generation metrics are shown in Table 5. When considering all the questions, the model generates the solution text with higher accuracy than the lecture text. However, in cases where the model answers incorrectly, the trend reverses and the model has a higher accuracy in generating the lecture text as compared to the solution text. Therefore, the model may be failing to answer these questions due to a failure to reason in the "solution stage" of its generation. Furthermore, for the 13B model we see that the lecture generation performance is higher for incorrect answers than correct answers (ROUGE score of 0.924 for incorrect vs. 0.861 for correct). This could indicate overfitting, where the model "memorizes" lectures that apply to the problem but fails to apply the lectures to the actual solution.

This problem is apparent with an example question about object properties, where the model must determine the property shared by an icicle, a fish bowl, a glass, and a tea cup. The model correctly generates the lecture about object properties required to reason through the problem ("An object has different properties. A property of an object can tell you how it looks, feels, tastes, or smells."). However, in the solution stage the model incorrectly reasons that all the objects were transparent

instead of fragile, based on a failure to infer the properties of the objects from the image (“You can see clearly through a transparent object. All four objects are transparent.”).

Incorrect reasoning can be attributed to two factors, i.e., linguistic and visual features. In a manual analysis of 100 test samples, we found that linguistic features are a weakness for mainly two use cases, namely retrieving commonsense facts (e.g. characteristics of bird song) and semantic understanding of words in terms of figure of speech and relative position of words in the dictionary. In contrast, visual features appear to be strong in use-cases such as identifying geographic areas but it lags in counting numbers in images and retrieving properties of objects such as color, texture and states.

These observations suggest that to improve model training, we need a wide variety of human-curated instructions, especially datasets that include both text and visuals explained by humans. Such diverse explanations would help the model understand various scenarios that require different types of reasoning.

Table 5: Evaluation of generated lectures and solutions.

	7B Model		13B Model	
	BLEU	ROUGE	BLEU	ROUGE
All answers				
Lecture	0.763	0.778	0.854	0.868
Solution	0.791	0.838	0.872	0.921
Correct answers				
Lecture	0.765	0.780	0.847	0.861
Solution	0.829	0.873	0.893	0.937
Incorrect answers				
Lecture	0.751	0.767	0.909	0.924
Solution	0.565	0.631	0.694	0.778

## 4 Related Work

Zhang et al. (2023a) proposed LLaMA-Adapter to guide the LLaMA model to follow multimodal instructions. Specifically, they proposed a zero-init attention with gating as a Parameter-Efficient Fine-Tuning (PEFT) technique to prepend learnable multimodal adaptation prompts to the input text tokens at higher transformer layers in the LLaMA model. The same authors proposed LLaMA-Adapter-V2 (Gao et al., 2023) that distributes the learnable parameters across all layers in the LLaMA model to improve performance in multimodal reasoning. MiniGPT-4 (Zhu et al., 2023) combined the frozen LLM (Vicuna) and a vision

encoder with a single projection layer and fine-tuned with a highly-curated visual conversation dataset. More recently, Liu et al. (2023) introduce *visual instruction tuning* to develop general-purpose visual assistant (LLaVA) that follows multimodal instructions. They present several data reformation techniques to construct multimodal instruction-following data from the standard image-text pairs. For example, the LLaVA model was trained with 595K image-text pairs filtered from the CC3M dataset (Sharma et al., 2018), and 158K unique language-image instruction-following data generated from ChatGPT/GPT-4 (Liu et al., 2023). This multimodal instruction set includes image-based *conversations* and *detailed descriptions* and *complex reasoning* questions. LLaVA (Liu et al., 2023) reaches the best performance in the ScienceQA benchmark with support from GPT-4 that acts as a judge to evaluate the generated answers. LLaVAR (Zhang et al., 2023c) extends the LLaVA for text-rich images by training with additional 422K image-OCR and 16K conversations generated from GPT-4.

## 5 Conclusion

In this work, we present scientific multimodal instruction tuning to align LLMs with scientific concepts and goals. To this end, we use human-generated multimodal instructions curated from visual signals (e.g., plots, charts, equations), and textual signals (e.g., captions, optical character recognition (OCR) and paragraph mentions) found within scientific publications. We train several models built on top of LLaMA language decoder model and CLIP vision encoder model and test the models on science-focused multimodal downstream tasks. In evaluation, we show that the resulting *LLaMA-SciTune-SciCap* models can perform better on classifying scientific visuals and generating figure captions compared with SOTA vision-to-language models. Furthermore, the *LLaMA-SciTune-ScienceQA* model surpasses the human performance in ScienceQA, the standard multimodal science-focused reasoning QA benchmark.

Our results suggest human-curated scientific multimodal data remains highly valuable despite the advancements in synthetic data generation techniques. While it is fast and easy to generate large volume of synthetic training data with closed-source models such as GPT-4, they may contain inaccuracies or biases due to lack of expert review. In



contrast, although existing human-generated scientific multimodal datasets are comparatively smaller, they provide reliable ground truth for tuning LLMs, which leads to better generalization and performance in downstream science applications.

## Acknowledgements

This work was supported by the NNSA Office of Defense Nuclear Nonproliferation Research and Development, U.S. Department of Energy, and Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC05-76RLO1830. This article has been cleared by PNNL for public release as PNNL-SA-186641.

## References

- Markus Anderljung, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O’Keefe, Jess Whittlestone, Shahr Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. 2023. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, et al. 2023. Going beyond nouns with vision & language models using synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20155–20165.
- Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.
- Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. *Llama-adapter v2: Parameter-efficient visual instruction model*. *Preprint*, arXiv:2304.15010.
- Ting-Yao Hsu, C Lee Giles, and Ting-Hao’Kenneth’ Huang. 2021. Scicap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624*.
- Chieh-Yang Huang, Ting-Yao Hsu, Ryan Rossi, Ani Nenkova, Sungchul Kim, Gromit Yeuk-Yin Chan, Eunye Koh, Clyde Lee Giles, and Ting-Hao’Kenneth’ Huang. 2023. Summaries as captions: Generating figure captions for scientific documents with automated text summarization. *arXiv preprint arXiv:2302.12324*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Chunyan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.
- Haotian Liu, Chunyan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.
- Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M Dai. 2022. Mind’s eye: Grounded language model reasoning through simulation. *arXiv preprint arXiv:2210.05359*.
- Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakkka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

- Roma Patel and Ellie Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.
- Ilya Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.
- Benny J Tang, Angie Boggust, and Arvind Satyanarayan. 2023. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.
- Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *arXiv preprint arXiv:2406.06007*.
- Renrui Zhang, Jiaming Han, AoJun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023a. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.
- Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023c. Llavav: Enhanced visual instruction tuning for text-rich image understanding. *Preprint, arXiv:2306.17107*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023d. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.
- Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. *arXiv preprint arXiv:2403.05020*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.
- Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

## A SciTune Multimodal Instructions

Table 6 shows two SciTune instruction examples used in the scientific concept alignment training stage. Table 7 presents the LLaVA questions that used to sample the  $s_I$  instructions.

## B Training Details

We use the LLaVA codebase (Liu et al., 2023) for multimodal adapter training with SciTune instructions. We train the model for 1 epoch with 128 batch size with a 0.002 learning rate and 2048 context length. LLaMA-SciTune 7B and 13B model variants took 6.5 and 11.2 hours to train, respectively with 8 x A100 GPUs. We use Pytorch Fully Sharded Data Parallel (FSDP) to recursively wrap the language models decoder layers in the task-specific instruction finetuning stage. We finetuned the models for 12 epochs with the ScienceQA training example to make a fair comparison with LLaVA.

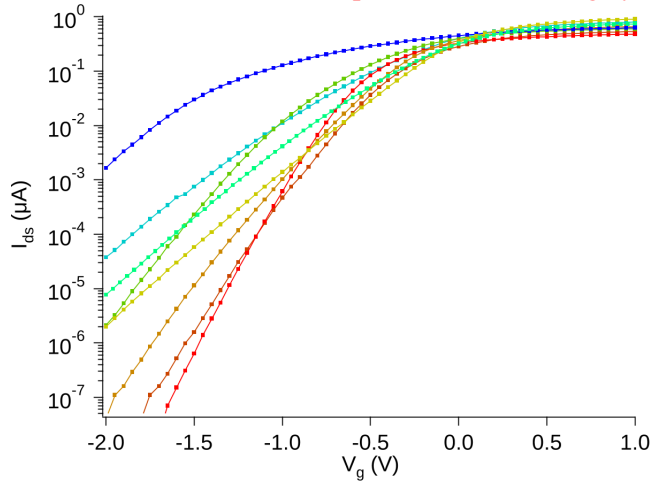
## C LLaMA-1 and LLaMA-2 Base Model Comparison

We also test our methodology with LLaMA-2 (Touvron et al., 2023b) as the base language decoder

Table 6: SciTune Multimodal Instruction Examples. We distinguish the system message  $s_D$ , natural language instruction  $s_I$ , scientific figure type  $s_t$ , caption  $s_c$ , OCR  $s_o$  and paragraph mentions  $s_m$ . The list of instructions used to sample  $s_I$  are presented at Table 7.

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human’s questions.

Human: Give an elaborate explanation of the image you see.

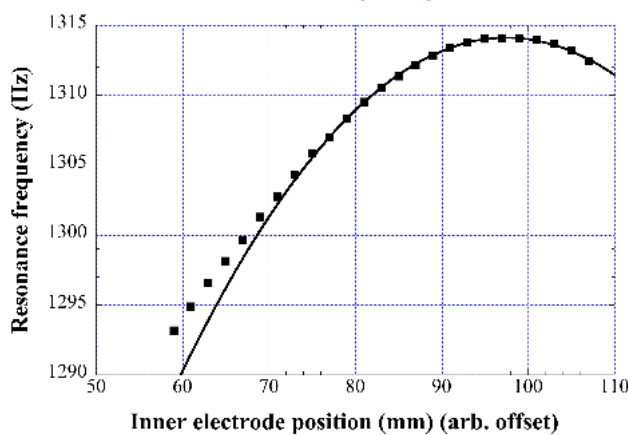


Assistant: Graph Plot Room temperature characteristics (drain-source current versus gate voltage) for a set of nominally identical samples from the same wafer. Sample to sample variations are very important, especially below the threshold voltage (-0.5 V for the best device, in red). In the most extreme case (blue curve) the many decades of excess current compared to the best device are attributed to direct transport through individual arsenic dopants which have diffused into the channel after doping and annealing steps. The source-drain voltage is 10mV for these data.

FIG. 6: figure S1: Room temperature characteristics (drain-source current versus gate voltage) for a set of nominally identical samples from the same wafer.

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human’s questions.

Human: Describe the following image in detail.



Assistant: Graph Plot Resonance frequency of the (1, 0) mode of the  $r_1 = 4$  cm cavity filled with 1 atm of  $N_2$  as a function of the vertical position of the central electrode.

Inner electrode position, (arb , offset)

In Fig. 10 we have an example of this behavior: the resonance frequency shows a maximum which corresponds to a minimum in the eccentricity of the electrode [39].

Table 7: LLaVA Instructions for detailed image descriptions

"Describe the following image in detail."
"Provide a detailed description of the given image."
"Give an elaborate explanation of the image you see."
"Share a comprehensive rundown of the presented image."
"Offer a thorough analysis of the image."
"Explain the various aspects of the image before you."
"Clarify the contents of the displayed image with great detail."
"Characterize the image using a well-detailed description."
"Break down the elements of the image in a detailed manner."
"Walk through the important details of the image."
"Portray the image with a rich, descriptive narrative."
"Narrate the contents of the image with precision."
"Analyze the image in a comprehensive and detailed manner."
"Illustrate the image through a descriptive explanation."
"Examine the image closely and share its details."
"Write an exhaustive depiction of the given image."

model. LLaMA-2 was reported to have superior performance compared to LLaMA-1 with additional pretraining corpus (2x tokens), larger context length (2x), and adopted grouped-query attention. In addition, there were additional steps taken to improve the safety of LLaMA-2 models. We do not use the LLaMA-2 model variants optimized for chat and dialogue use cases to make a fair comparison with LLaMA-1 model. In this experiment, we repeat the entire training (CTOM) pipeline (as shown in Figure 1) with the LLaMA-2 (13B) model. Figure 2 shows a comparison of ScienceQA performance with LLaMA-1 and LLaMA-2 base language models. Despite the reported performance improvements in the LLaMA-2 model over the LLaMA-1, we do not observe any performance advantage of the former model in the ScienceQA benchmark. This may be due to the effect of larger pretraining data or a different dataset mix used to pretrain LLaMA-2. For example, LLaMA-2 authors reported an increase in toxicity of the new models with more than 7B parameters (Touvron et al., 2023b). It remains as a future work to reason this performance difference with more empirical results. We use the LLaMA-1 as the base language decoder model in the rest of the experiments unless explicitly mentioned.

## D Visual Grounded Task Performance

Table 8 shows a few generated captions for the SciCap images used to test the model performance on visual grounded tasks. We report the gold-standard

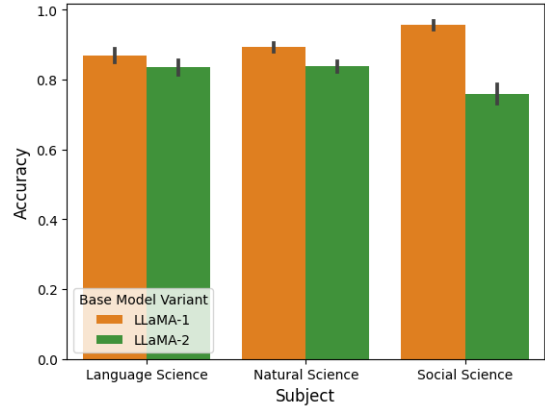


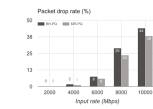
Figure 2: ScienceQA Performance of the LLaMA-SciTune models with LLaMA-1 and LLaMA-2 as the base language decoder models

captions as they appeared in the arXiv articles used to collect SciCap dataset, and the captions generated from the BLIP and LLaMA-SciTune (13B, CTOM) models for the comparisons.

## E ScienceQA Chain of Thought Reasoning Examples

Figures 3 and 4 show a few answers, lectures and solutions generated by LLaMA-SciTune (13B, CTOM) for ScienceQA test instances.

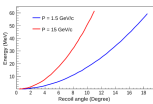
Table 8: A Sample of Generated Captions. We highlight the gold standard caption in red, and generated captions from the BLIP (Li et al., 2022) model in gray. LLaMA-SciTune model first generates the figure types followed with the captions colored in blue.



**Packet drop rate**

a chart of a bar chart with a number of different items

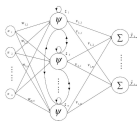
**Bar Chart** Packet drop rate for each method.



**The kinetic energy of the recoil protons as a function of the recoil angle at beam momenta P=1.5 and 15 GeV/c, blue and red, respectively.**

a plot of a curve with a blue line and a red line.

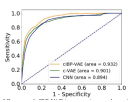
**Graph Plot** The angular distribution of the electron recoil spectrum in the 1.5 GeV/c and 15 GeV/c electron beams.



**Artificial neural network structure.**

a diagram of a network with several different paths.

**Node Diagram** The generative neural network.



**ROC curves of cIBP-VAE in comparison to alternative models on the clinical ECG data set.**

a plot of the average and average time of a cell phone.

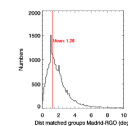
**Graph Plot** ROC curves of c-VAE, CNN, and c-VAE+CNN on the cerebellar atrophy dataset.



**Functional architecture of the developed prototype.**

a diagram of a camera and a person on a phone.

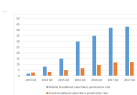
**Node Diagram** An overview of the system architecture of the proposed method.



**Distance between matched groups in Madrid and RGO catalogs (bins of 0.1 degrees). The red line represents the mean value.**

a plot of a line of data with a red line and a white line.

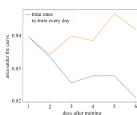
**Graph Plot** Distance correlation between groups matched by Madrid RGO.



**Comparison of penetration rate of mobile broadband subscribers with that of fixed broadband subscribers.**

a chart of the number of people who are using the internet.

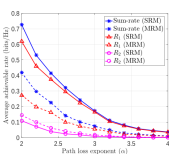
**Bar Chart** The average rate of mobile broadband subscribers and fixed broadband subscribers for each quarter (in thousands).



**Model staleness of the one-off trained model vs. the model retrained every day.**

a diagram of a graph with a line graph and a line graph.

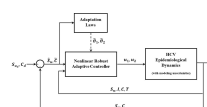
**Graph Plot** Prediction accuracy of model trained once and fine-tuned every day.



**Comparison of the effect of the path loss exponent  $\alpha$  on rates achieved by both transmitters,  $M = 4$ .**

a plot of a line graph with a blue line and red line.

**Graph Plot** The sum-rate and sum-rate of RRM-RRM with respect to the path loss exponent  $\gamma$  for the two cases:  $\gamma = 2$  and  $\gamma = 3$ .



**Conceptual diagram of nonlinear adaptive method developed to control the HCV epidemic in the existence of uncertainties on parameters of the model.**

a diagram of a block diagram of a nuclear system.

**Node Diagram** Block diagram of the proposed non-linear SIR epidemic model with adaptive controllers.

**Problem**

**Question:** Which ocean is highlighted?

**Context:** N/A

**Options:** (A) the Indian Ocean (B) the Atlantic Ocean (C) the Pacific Ocean (D) the Southern Ocean

**Ground-truth:** (A)

**Ground-truth Rationale:** LECTURE: Oceans are huge bodies of salt water. The world has five oceans. All of the oceans are connected, making one world ocean. SOLUTION: This is the Indian Ocean.

**Answer:** (A)

**Explanation:** LECTURE: Oceans are huge bodies of salt water. The world has five oceans. All of the oceans are connected, making one world ocean. SOLUTION: This is the Indian Ocean



(a) An example with right answer and right explanation

**Problem**

**Question:** Which property matches this object ?

**Context:** Select the better answer.

**Options:** (A) transparent (B) bouncy

**Ground-truth:** (A)

**Ground-truth Rationale:** LECTURE/SOLUTION: An object has different properties. A property of an object can tell you how it looks, feels, tastes, or smells. Properties can also tell you how an object will behave when something happens to it.

**Answer:** (B)

**Explanation:** LECTURE: An object has different properties. A property of an object can tell you how it looks, feels, tastes, or smells. Properties can also tell you how an object will behave when something happens to it. SOLUTION: Look at the object.



(b) An example with incorrect answer and incorrect explanation

Figure 3: Two Multimodal QA examples with answer and explanation generated by LLaMA-SciTune

**Problem**

**Question:** How long is a hiking trail?

**Context:** Select the best

**Options:** (A) 4 meters (B) 4 millimeters (C) 4 kilometers (D) 4 centimeters

**Ground-truth:** (c)

**Ground-truth Rationale:** LECTURE: Measurements are written with both a number and a unit. The unit comes after the number. The unit shows what the number means. When you are using metric units, length can be written with units of millimeters, centimeters, meters, or kilometers. One meter contains 100 centimeters or 1,000 millimeters. So, 1 meter is larger than 1 centimeter, and 1 centimeter is larger than 1 millimeter. The tip of the pencil shown here is only 1 millimeter wide, but the pencil is about 16 centimeters long. A red fox is about 1 meter long. The Sydney Harbour Bridge in Australia is about 1,000 meters, or 1 kilometer, in length. SOLUTION: The best estimate for the length of a hiking trail is 4 kilometers. 4 millimeters, 4 centimeters, and 4 meters are all too short.

**Answer:** (c)

**Explanation:** LECTURE: Measurements are written with both a number and a unit. The unit comes after the number. The unit shows what the number means. When you are using metric units, length can be written with units of millimeters, centimeters, meters, or kilometers. One meter contains 100 centimeters or 1,000 millimeters. So, 1 meter is larger than 1 centimeter, and 1 centimeter is larger than 1 millimeter. The tip of the pencil shown here is only 1 millimeter wide, but the pencil is about 16 centimeters long. A red fox is about 1 meter long. The Sydney Harbour Bridge in Australia is about 1,000 meters, or 1 kilometer, in length. SOLUTION: The best estimate for the length of a hiking trail is 4 kilometers. 4 millimeters, 4 centimeters, and 4 meters are all too short.

None

(a) An example with right answer and right explanation

**Problem**

**Question:** Would you find the word bruise on a dictionary page with the following guide words? Blush-buffalo

**Context:** N/A

**Options:** (A) no (B) yes

**Ground-truth:** (B)

**Ground-truth Rationale:** LECTURE: Guide words appear on each page of a dictionary. They tell you the first word and last word on the page. The other words on the page come between the guide words in alphabetical order. To put words in alphabetical order, put them in order by their first letters. If the first letters are the same, look at the second letters. If the second letters are the same, look at the third letters, and so on. If one word is shorter, and there are no more letters to compare, then the shorter word comes first in alphabetical order. For example, becomes before bed. SOLUTION: Put the words in alphabetical order. Since bruise is between the guide words blush - buffalo, it would be found on that page.

**Answer:** (A)

**Explanation:** LECTURE: Guide words appear on each page of a dictionary. They tell you the first word and last word on the page. The other words on the page come between the guide words in alphabetical order. To put words in alphabetical order, put them in order by their first letters. If the first letters are the same, look at the second letters. If the second letters are the same, look at the third letters, and so on. If one word is shorter, and there are no more letters to compare, then the shorter word comes first in alphabetical order. For example, becomes before bed. SOLUTION: Put the words in alphabetical order. Since bruise is not between the guide words blush - buffalo, it would not be found on that page.

None

(b) An example with incorrect answer and incorrect explanation

Figure 4: Two Unimodal QA examples with answer and explanation generated by LLaMA-SciTune