

AgroGPT 🌱: Efficient Agricultural Vision-Language Model with Expert Tuning

Muhammad Awais^{1,2}, Ali Husain Salem Abdulla Alharthi¹, Amandeep Kumar^{1,3},
 Hisham Cholakkal¹, Rao Muhammad Anwer¹

¹Mohamed bin Zayed University of AI, UAE, ²Georgia Institute of Technology, USA, ³Johns Hopkins University, USA
 aawais@gatech.edu, rao.anwer@mbzuai.ac.ae

Abstract

Significant progress has been made in advancing large multimodal conversational models (LMMs), capitalizing on vast repositories of image-text data available online. Despite this progress, these models often encounter substantial domain gaps, hindering their ability to engage in complex conversations across new domains. Recent efforts have aimed to mitigate this issue, albeit relying on domain-specific image-text data to curate instruction-tuning data. However, many domains, such as agriculture, lack such vision-language data. In this work, we propose an approach to construct instruction-tuning data that harnesses vision-only data for the agriculture domain. We utilize diverse agricultural datasets spanning multiple domains, curate class-specific information, and employ large language models (LLMs) to construct an expert-tuning set, resulting in a 70k AgroInstruct. Subsequently, we expert-tuned and created AgroGPT, an efficient LMM that can hold complex agriculture-related conversations and provide useful insights. We also develop AgroEvals for evaluation and compare AgroGPT's performance with large open and closed-source models. AgroGPT excels at identifying fine-grained agricultural concepts, can act as an agriculture expert, and provides helpful information for multimodal agriculture questions. The code, datasets, and models are available at <https://github.com/awaisrauf/agroGPT>.

1. Introduction

The widespread availability of joint image-text data across the internet has catalyzed significant advancements in large multimodal models (LMMs), such as CLIP [33], GPT-4 [2], Bard [28], and LLaVA [21]. These models, when instruction-tuned to align with human intents, have demonstrated remarkable zero-shot performance and multi-turn conversational capabilities [30]. However, their efficacy often falters in domains requiring specialized expertise, such as biomedicine, climate science, and geo-sensing, resulting in incorrect responses and hallucinations. To ad-

Question

Bard

AgroGPT

Question: What disease is affecting my plant?

Ground Truth: Late Blight Disease

I cannot determine what disease is affecting the leaf in the image you sent. However, some of the possible diseases that could cause similar symptoms include:

- **Anthracnose:** This is a fungal disease that can cause leaf spots, dieback, and cankers on a variety of plants.

The image shows a tomato plant leaf with brown spots and a yellowish tint. The symptoms suggest that the plant is affected by **late blight disease**.

What are some (chemical, biological) ways to control this disease in my crop?

There are several ways to control late blight disease in tomato plants. Chemical methods include the use of fungicides, such as copper-based products, which can be applied at the first sign of disease or at regular intervals throughout the growing season. Biological methods include the use of beneficial fungi, such as *Trichoderma* spp., which can help to suppress late blight disease by competing with the pathogen and producing enzymes that break down the cell walls of the pathogen.

Figure 1. Comparison of AgroGPT with Bard to identify and provide information about a disease from leaf image. Our model can correctly identify the disease and effect of the disease on the plant.

dress this limitation, several fields have successfully leveraged existing domain-specific vision-language datasets to create domain-specific instruction-tuning data, enhancing model performance in specialized areas [17, 27, 40]. For instance, BioMedLLaVA utilized large vision-language data PMC-15M [46] to curate biomedical instruction-tuning data resulting in LLaVA-Med [19]. Despite these advancements, many critical domains, including agriculture, have vision-

only datasets but lack high-quality joint vision-language data suitable for curation of domain-specific instruction-tuning data.

In this paper, we introduce AgroGPT, the first approach to extending multi-modal instruction tuning to the agricultural domain without requiring joint visual language data. Our work is a comprehensive attempt to overcome this limitation, comprising three key components. First, a pipeline for synthesizing expert-level instruction-tuning data from vision-only datasets, effectively bridging the multimodal data gap in agriculture. This resulted in *AgroInstruct*, a 70k expert-tuning dataset for agriculture. Second, *AgroGPT*, an efficient, open-source expert agricultural conversational model, trained on the *AgroInstruct* to provide domain-specific insights and assistance. Finally, we develop *AgroEvals*, a visual question-answering framework to evaluate model performance in agricultural contexts.

First, we create *AgroInstruct* by our proposed pipeline for curating expert-level instruction-tuning data solely from vision-only datasets, specifically designed to enhance LMMs in the agricultural domain. To this end, we aggregate six diverse image-only datasets spanning four agricultural domains: plant diseases, weeds, farm insects, and fruits. This curated collection forms the foundation of our approach, ensuring broad coverage of agricultural visual data. Next, to transform the vision-only data into rich expert-tuning examples, we synthesize context-grounded image descriptions using state-of-the-art general-purpose LMM, extract relevant image attributes, and collect class-specific background knowledge from agricultural sources. Finally, we develop domain-specific instruction-following examples and system prompts tailored to agricultural contexts. These elements are fed into language-only LLMs to generate diverse and contextually relevant conversations. Further, we implement a rule-based system leveraging dataset attributes to generate concise question-answer pairs, enhancing the variety and specificity of the instruction-tuning data. Through this pipeline, we bootstrapped *AgroInstruct*, an extensive and diverse multimodal instruction-tuning dataset comprising 70k high-quality conversations.

Furthermore, we extend the two-step visual instruction tuning [21] to include expert tuning and train efficient, conversational models (*AgroGPT-3B* and *AgroGPT-7B*) on *AgroInstruct*. To evaluate the model’s capabilities, we develop six visual question-answering (VQA) tasks (*AgroEvals*) across three different groups. To demonstrate the effectiveness of our method, we present comprehensive quantitative and qualitative results, comparing our models with several open and closed-source models, such as Bard and LLaVA-34B. On VQA tasks, our model significantly outperforms the baselines in identifying fine-grained agricultural concepts. Additionally, *AgroGPT* excels at multi-turn, complex, domain-specific conversations, providing better



Dataset	Num of Imgs	Classes
Cotton Leaf Disease [1]	4800	6
PlantVillage [13]	54303	38
PlantDoc [37]	2598	30
Fruits100 [24]	40000	100
Early Weed Detection [8]	1000	4
Farm Insects [10]	6000	12

Table 1. Four agricultural domains (disease, weeds, insects, and fruits) and statistics of vision-only datasets used to create our expert-tuning *AgroInstruct* data.

guidance than several state-of-the-art general models. A demonstration of our model is shown in Fig. 1.

2. Related Work

General Large Language Models (LLMs) and Large Multimodal Models (LMMs). Large Language Models (LLMs) possess remarkable capabilities to understand and generate text that resembles human language. This enables them to tackle various tasks effectively, with GPT-4 [31], Llama-2 [42], and Gemini [38]. Recently, these LLMs are extended to multi-modal inputs by aligning visual information in the LLM embedding space for which a pre-trained visual backbone [9] is used to process visual data, a large language model [7] for understanding user instructions and generating responses, and a cross-modal connection MLP [20] to integrate visual information with language models. The impressive improvements in tasks related to language instruction following and visual reasoning in natural scenes are recently shown in LLaVA [21], MiniGPT-4 [48] [6], InstructBLIP [20], Otter [18].

Expert Large Language Models. Recent large language models such as ChatGPT [22]/GPT-4 [31] have significantly expanded their domain expertise, specialized areas such as medicine, science, literature, and technology. Due to their training on a large corpus of data, these models have achieved remarkable few-shot and zero-shot learning capabilities [5, 16, 43], leading to having impressive performance even in domains where the model is not specifically trained. However, these large expertise models are not particularly designed for specific domain expertise. Therefore, recently, a large number of effort has been made on open-sourced instruction-tuned LLMs to train a domain-specific language models like Mini-ClimateGPT [27], GeoChat [12, 17], XrayGPT [40], ClimateGPT [41]

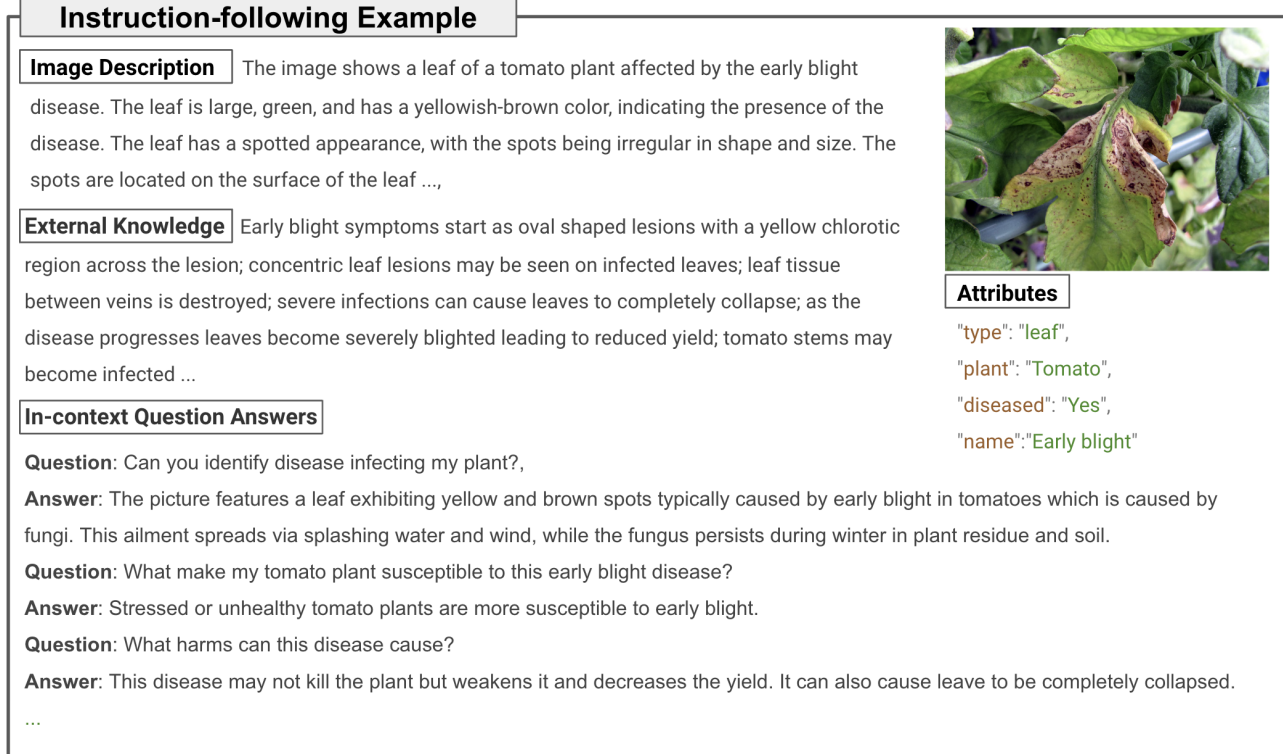


Figure 2. An illustration of instruction-following examples used to generate expert-tuning from vision-only agriculture data.

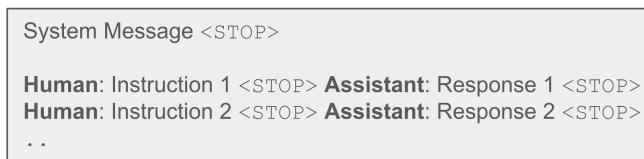


Figure 3. An illustration of sequencing of conversation before being fed to LLM, following LLaVA and LLaVA-Med [19,21].

and VideoGPT [26, 35]. However, unlike these models, we introduce a pipeline to construct expert-tuning data from existing image-only datasets and developed efficient AgroGPT (3 and 7B), which can hold complex conversations about fine-grained agricultural concepts.

Deep learning in Agriculture. Supervised learning-based approaches have mainly been explored for agriculture, such as vision-based datasets for important tasks in agriculture (plant diseases [13, 32, 34, 36, 37], harmful insects [25], fruits [29], and weeds identification [8]). However, the agriculture field lacks any comprehensive image-text datasets. Similarly, deep learning models have also been proposed to solve these tasks [3, 11, 39]. Recently, PLLaVA [44] was introduced for plants. However, it is limited to text-based questions. We proposes an AgroInstruct pipeline to convert vision-only datasets into expert-tuning datasets and train a large model capable of identifying fine-grained agricultural concepts and holding complex conversations.

3. Agricultural Expert-tuning

In the following sections, we first explain our efficient pipeline for curating expert tuning data to instill agricultural knowledge in Large Multimodal Models (LMMs). We utilize classification-only image datasets and class-specific information to synthesize the data. Next, we describe the architecture of our LMMs, followed by the training process using the expert tuning dataset. This tuning is performed after the vision-language alignment and general visual instruction tuning stages [21].

3.1. Curation of AgroInstruct for Expert Tuning

The agricultural domain has several vision-only datasets (as depicted in Fig. 2), but it lacks multi-modal data, presenting a challenge for developing instruction-tuning datasets. To address this issue, we propose a solution that leverages vision-only datasets from four different domains (diseases, weeds, insects, fruits) as shown in Tab. 1 to synthesize expert tuning data called AgroInstruct following visual instruction tuning [21] and LLaVA-med [19].

AgroInstruct consists of three types of questions: complex multi-turn conversations, simple questions, and image descriptions. First, we generate an image description with generic LMMs by providing it with the context of the dataset and image-specific attributes. Next, to generate multi-turn complex conversations, we utilize an open-

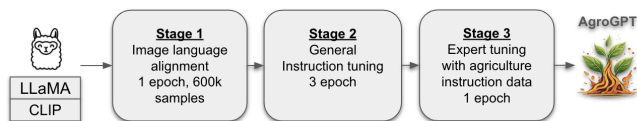


Figure 4. An overview of AgroGPT three-stage training process to convert an image encoder and language-only LLM to a domain expert. The first two stages consist of vision-language concept alignment and general instruction-tuning, which is based on LLaVA [21].

source LLM to generate expert-tuning data by providing it with context-conditioned descriptions, class-specific external background information from reputable agricultural resources, and in-context examples of conversations as shown in Fig. 2. This is provided as input into a Large Language Model (LLM) alongside appropriate system prompts. Finally, we utilize attributes of image-only datasets to generate simple questions, such as identifying plants and diseases. Our expert-tuning data, AgroInstruct, consists of 70k conversations belonging to three types (10K descriptions, 35 complex questions, and 35k simple questions). An example of these three types of questions is provided in Fig. 5. In the following, we elaborate on each step.

Context-grounded Image Descriptions. First, we generate context-grounded image descriptions by prompting a vision-language model to describe the input based on dataset-specific attributes such as plant name, disease name, and whether it’s a plant or a leaf. We construct a system of prompts to instruct models to utilize context and generate specific and field-specific information. We devised 10 questions to describe the image following MedLLaVA [19]. This process resulted in the construction of approximately 35K descriptions spanning four domains. We only utilize 10k descriptions in our AgroInstruct data, while the rest are utilized as context for the generation of complex conversations explained next.

Complex, Multi-turn, and Open-ended Conversations. We need a dataset containing multi-turn conversations to equip our model with the capability to answer complex, open-ended, and agriculture-related questions about an image. To this end, we craft a diverse set of conversations featuring multi-round conversations using a language-only Large Language Model (LLM) named Mistral [15]. To guide LLM in generating field-specific and to-the-point conversations, we utilize the following four as context for each input image.

- **Image Attributes.** Image-only datasets (Tab. 1) have attributes for each image, such as plant name, disease name, fruit name, etc. These attributes are added in the context of LLM to facilitate its understanding of the nature of the input.

Complex Conversation

Question: What plant is shown in the image?
Answer: The image appears to depict a leaf from a pepper bell plant.
Question: What condition does the leaf in the image seem to be in?
Answer: The leaf in the image shows signs of bacterial spot disease, as indicated by brown spots and a curled appearance.
Question: What symptoms might be associated with the bacterial spot on a pepper plant?
Answer: The image reveals a leaf of a pepper plant with signs of bacterial spot disease, such as brown spots and a curled appearance. Symptoms of bacterial spots on pepper leaves include water-soaked lesions that dry out and turn brown, raised brown cankers on stems, and cracked brown lesions on fruits. The disease can be favored by warm, wet condition and can spread within a field by wind-driven rain, clipping plants, and aerosols.
Question: What can be done to manage bacterial spots in pepper plants?
Answer: Management practices for bacterial spot in pepper plants include using disease-free planting material and removing destroying all crop debris after harvest. The material can also be plowed deeply under soil to prevent the disease’s spread.

Short Questions

Question: Name the plant in the photo using only one word.
Answer: Pepper bell
Question: Is the plant in this image sick? Respond in one word
Answer: Yes
Question: Identify the disease in the plant shown here. Just name the disease.
Answer: Pepper bell bacterial spot

Image Description

Question: Describe the image
Answer: The image features a leaf of a pepper plant, which appears to be affected by a bacterial spot disease. The leaf is covered in numerous small, round, brown spots, indicating the presence of the disease. The spots are scattered across the surface of the leaf, and their size and distribution suggest that the infection is relatively severe. This type of disease can negatively impact the plant’s health and yield, so it is essential to monitor and address the issue promptly to prevent further damage.

Figure 5. An example of three types of conversations in the AgroInstruct dataset: complex, multi-turn, short questions, and image descriptions. Zoom in for a better view.

- **External Information.** To further enable the model to understand nuanced information about the topic of the input image, we curated class-specific information from reputable agricultural external resources for each class. Our dataset consists of 202 classes, and we curate, on average, a page-long external information for each class. We add external resources in the Appendix.
- **In-context Examples.** We created a few in-context examples for each domain to create a diverse set of multi-turn question-answers. These in-context examples are meant to guide LLM in the generation of question-and-answer pairs. Further, we also want the model to use the information provided in the attributes, descriptions, and external information. These in-context examples are designed in such a way that they correlate information with answers, hence guiding LLM to look into the provided information.
- **System Prompt.** Finally, we generated a system prompt for each domain to instruct the model to generate comprehensive conversations for that domain. Overall, we follow MedLLaVA’s [19] technique to create these prompts.

We provide an example in Fig. 2. We provided this context to an LLM (Mistral) and system prompt to create 3-

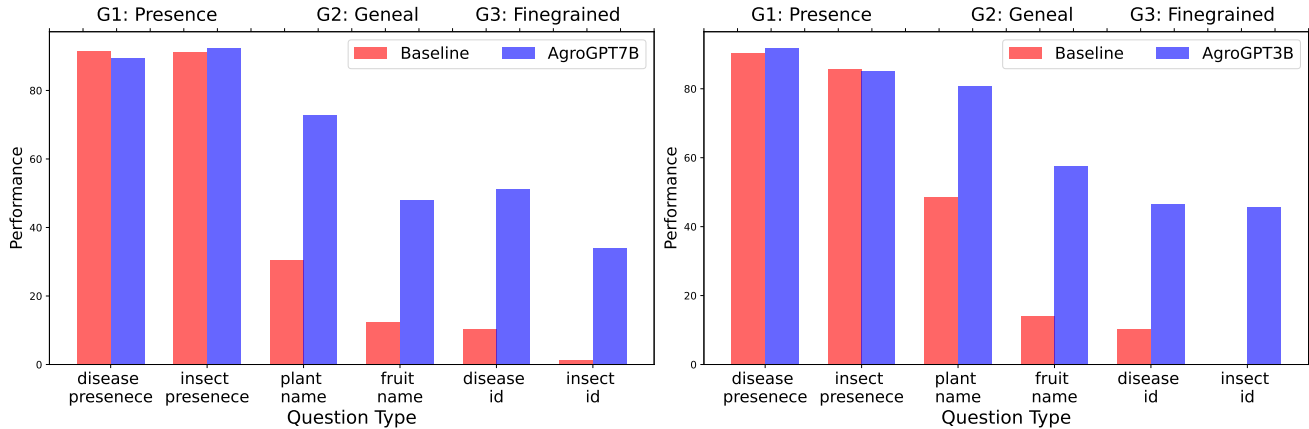


Figure 6. Comparison of our method’s ability to improve agricultural knowledge of baseline on two versions of AgroGPT. General instruction-tuned models work well for general knowledge (e.g., is this leaf diseased?) but fail on fine-grained concepts such as disease names or insect names. Expert tuning significantly improves this.



User:	What disease is affecting my plant?
AgroGPT:	The plant in the image is affected by early blight , a fungal disease.
User:	How can I control this disease?
AgroGPT:	The disease can be controlled by applying appropriate fungicides at the first sign of disease, removing and destroying infected plant debris, and practicing crop rotation. It is also important to maintain good plant hygiene and avoid over-application of nitrogen fertilizer.
Bard:	Based on the image you sent, the plant is most likely affected by a bacterial disease. There are a few possibilities , including: Bacterial leaf spot, Xanthomonas spot, and Erwinia blight.
LLaVA-34B:	The plant in the image appears to be affected by a disease that causes leaf spots and necrotic tissue. The spots are brown and have a concentric ring pattern, which is characteristic of a red fungal disease .

Table 2. Comparison of our efficient AgroGPT-3B’s abilities with Bard, ChatGPT4o, and LLaVA-34B on open-ended questions for an image taken from web. AgroGPT-3B correctly identifies the disease (early blight) and suggests control methods, while Bard and ChatGPT-4o decline to answer, and LLaVA-34B provides an incorrect response.

5 questions per image. These conversations include questions to identify specific elements within the image and answer domain-specific questions about the identified object. We produced approximately 35k multi-round, complex, agriculture-related conversations using this method.

Simple Question Answers In addition to the complex, multi-round conversations, we have augmented our dataset

with rule-based simple question-answer pairs. These questions are designed to assist the model in accurately identifying specific elements within the image and providing concise answers. To create these questions, we leverage dataset attributes such as the plant name, health status (diseased or healthy), disease name, insect names, fruit names, etc. Subsequently, we generate several variations for each type of question, prompting the model to respond with a single word regarding the specified attribute. For example, for the disease attribute, one question is: “What disease does the plant in the image have? Provide the name of the disease only.” We have generated 35k such question-answer pairs.

3.2. Architecture of AgroGPT

For our baseline model, we utilized two types of models from the LLaVA family, comprising a visual encoder, a projection layer, and an LLM. For our AgroGPT, we employed two different models: a 7B LLaVA model [21] and an efficient 3B Mipha model [49]. These models are chosen to suit different computational requirements and performance considerations for downstream applications. The LLaVA and Mipha utilize pre-trained CLIP ViT-L/14 [33] and SigLIP [45] as visual encoders. In our experiments, we found SigLIP better than the traditional CLIP model, even when a smaller LLM is used. In conjunction with the projection layer, this visual encoder transforms an image into the language embedding space, producing visual tokens. These image tokens are combined with language instruction tokens and fed into the LLM. Subsequently, the LLM generates text conditioned on the language instructions and the image.

3.3. Expert Tuning AgroGPT with AgroInstruct

Our main goal is to instruction-tune a vision-language chat model that is expert in the agriculture field. To achieve this, we use the first two stages of visual instruction tun-



User:	What is the disease affecting the leaves?
AgroGPT:	The leaf in the image appears to be affected by tomato leaf curl virus .
User:	How can I control this disease?
AgroGPT:	Tomato Leaf Curl Virus is a viral disease that can be controlled planting resistant varieties, removing infected plants and burning them, avoiding the use of infected seedlings, and practicing crop rotation.
Bard:	Based on the image you sent, it is difficult to say for sure what disease is affecting the leaf. However, there are a few possibilities, including: Early blight(..) Septoria leaf spot(..), Bacterial leaf spot(..). It is also possible that the leaf is damaged by insects or other pests.
LLaVA-34B:	The plant in the image appears to be affected by a disease that is causing the leaves to become discolored and possibly wilted. The symptoms include yellowing and browning of the leaf edges, which could indicate a variety of issues such as nutrient deficiency, water stress, or a fungal or bacterial infection.

Table 3. Comparison of efficient AgroGPT-3B with Bard and 34B LLaVA on open-ended questions for an image taken from the held-set of PlantDoc dataset [37]. AgroGPT can correctly identify the disease and suggest several methods to control it in the crop. In contrast, Bard and LLaVA-34B are unable to identify the name.

ing [21] and introduce an expert tuning stage, as shown in Fig. 4. Specifically, the first stage consists of visual-language concept alignment on general images, and the second stage consists of instruction tuning on general vision-language data. We introduce a third expert tuning stage to instill agriculture-related expertise in the model. We kept the visual encoder frozen for expert tuning and trained the projection layer and LLM weights on our AgroInstruct data. For training, we converted all three formats of instructions (descriptions, small questions, multi-turn, and complex questions) into the same conversation format. Then, we randomly sampled conversations, added system messages as shown in Appendix Fig. 3, and trained LLM on its original auto-regressive objective. In the ablations, we also compare the effect of expert tuning after stage 1 and stage 2. Generally, expert tuning after stage 2 results in better results.

4. Experimental Results

Model Training. We use Phi-2-2.7B [14] and Vicuna-7B [47] as our LLMs and CLIP visual encoder ViT-



User:	What insect is in this image?
AgroGPT:	The insect in the image is a fall armyworm .
User:	How can it affect my crop?
AgroGPT:	Fall armyworms are a type of caterpillar that can cause significant damage to crops, particularly corn and soybeans. They feed on the leaves, stems, and fruits of these plants, which can lead to reduced yield and quality. In severe cases, the infestation can cause complete crop failure. The damage caused by fall armyworms can be difficult to control and can result in significant economic losses for farmers.
Bard:	The insect in the image you sent appears to be a caterpillar, but it is difficult to identify the exact species without more. However, based on the visible characteristics, here are some possibilities: Mythimna separata(..), Fall armyworm(..), Prodenia spp(..), Texas armyworm caterpillar(..) ..
LLaVA34B:	The insect in the image is a caterpillar .

Table 4. Comparison of efficient AgroGPT-3B with Bard and 34B LLaVA on open-ended questions for an image taken from the holdout-set of Farm Insects dataset [10]. AgroGPT can correctly identify the name of the insect and assess the potential harm it can do. In contrast, Bard is uncertain and provides several possibilities, and LLaVA34B incorrectly identifies it as a caterpillar.

L/14 [33] and SigLIP [45] with resolution 336 as our image encoders. We used Mipha [49] and LLaVA1.5 [21] provided pre-trained weights, which have been trained with visual instruction tuning. Unless stated otherwise, we applied our expert tuning after image-text alignment (stage 1) and visual instruction tuning (stage 2). We use AdamW [23] to optimize along with the cosine learning rate scheduler to train our model. Unless otherwise stated, we train our model for 1 epoch with a batch size of 8 or 16, depending on computational resources. For our training, we utilize 4 A6000 GPUs.

Expert Tuning Data. To generate our expert tuning, we used two different models. First, we use Mistral7B [15] to generate instruction tuning, following the method proposed in the previous section. Second, we use LLaVA-13B to create descriptions used as in-context information for Mistral7B, which is also explained in the previous section.

4.1. Qualitative Results

To demonstrate the effectiveness of AgroGPT in identifying agricultural issues (diseases, insects, weeds) and guiding users, we hold conversations on several examples. First, we compared AgroGPT’s performance with LLaVA-34B and Bard for crop diseases, farm insects, and weeds on held-

	Category	Baseline	Stage2	Stage3
Group1	Disease Presence	91.52	90.25	89.41
	Insect Presence	91.16	96.59	92.51
Group2	Plant Name	30.51	72.46	72.88
	Fruits Name	12.30	47.38	47.96
Group3	Disease Id	10.27	46.57	51.37
	Insect Id	0.25	35.57	35.89

Table 5. Comparison of performing expert tuning after stage 1 (image-text alignment) vs. stage 2 (general instruction-tuning) for AgroGPT-7B.

out images. The results are shown in Tab. 3, 4. These conversations highlight AgroGPT’s capability to identify fine-grained attributes (disease, insect, and weed names) and answer field-specific questions. For instance, AgroGPT accurately identifies diseases affecting the leaf, weed name, and farm insect in the images, provides suggestions for managing these issues, and explains its effect on the crop. In contrast, both Bard and LLaVA-34B either fail to identify fine-grained concepts or make incorrect identifications.

To evaluate the generalizability of AgroGPT, we further collected several images from online agriculture forums where users have asked questions. We compare AgroGPT’s capabilities to answer questions with LLaVA-34B, Bard, and ChatGPT4o. The conversations are presented in Tab. 2, and more results are relegated to the Appendix. AgroGPT can identify diseases correctly, assess potential damages, and suggest possible remedies. In contrast, both Bard and ChatGPT4o are unable to answer it, and LLaVA-34B hallucinates. For more details on quantitative evaluations, please refer to the Appendix.

4.2. Quantitative Evaluations

For quantitative evaluations, we create *AgroEvals* by utilizing the test and a holdout set of datasets utilized in the generation of AgroInstruct (see Tab. 1). Based on these datasets, we create six visual question-answering sets consisting of three groups that test the model’s capabilities at different levels. **Group 1** consists of questions to ask for the presence of a disease or insect in an image. Since these two attributes are general and coarse, a model trained on general instruction tuning (e.g., visual tuning) can easily answer these questions without expert tuning. **Group 2** consists of questions to name general categories such as fruits or plant names. These two attributes are finer-grained than group 1, but they are also present in instruction-tuning data. However, **Group 3** consists of questions that probe the model for fine-grained and field-specific questions, including the name of the disease or insect. Further, for our quantitative evaluations, we use two versions of our models: AgroGPT7B, which is based on LLaVA1.5 [21] and AgroGPT3B, which is based on Mipha3B [49].

	Type	Disease	Insects
Baseline	ZeroShot	10.27	0.00
ChatGPT	Zeroshot	30.82	32.65
AgroGPT	Zeroshot	51.37	45.58

Table 6. Comparison of AgroGPT with baseline and ChatGPT. AgroGPT outperforms ChatGPT in terms of fine-grained agricultural knowledge, demonstrating superior performance in fine-grained knowledge of agriculture (Group 3).

Main Results. In Fig. 6, we compare our method’s ability to add field-specific knowledge with the baseline of visual instruction tuning across six types of questions from three groups. Models fine-tuned with our method perform on par with visual instruction tuning for Group 1 questions, which tests the model’s coarse recognition abilities: presence and absence of disease and insects in a given image. However, our method performs significantly better for Group 2, which tests the model’s more fine-grained recognition abilities (e.g., name of fruit or plant name).

However, baseline models (fine-tuned with visual instruction tuning) completely fail on fine-grained questions from Group 3, which tests more fine-grained and relevant concepts such as identifying diseases or insects. On the other hand, our proposed expert tuning can instill these fine-grained concepts in the model and significantly improve their performance. For instance, both 3B and 7B baseline Mipha and LLaVA models completely fail to identify insects, while our method significantly improves this performance. Interestingly, AgroGPT3B performs better on fine-grained questions compared with AgroGPT7B. We hypothesize that this is because of the Mipha [49] baseline using a more capable image encoder (SigLIP [45]) compared with LLaVA’s CLIP.

Comparison with ChatGPT. We also quantitatively compare AgroGPT with ChatGPT for benchmark from Group 3 of AgroEvals. We chose Group 3 as it represents fine-grained agriculture knowledge (plant disease and farm insect identification). As shown in Tab. 6, AgroGPT outperforms ChatGPT by significant margins, such as on diseases, our model has a performance of 51.37% compared with ChatGPT’s 30.82%. It is also important to note that the efficient baseline model (Mipha3B) shows very poor performance as it lacks expert knowledge about fine-grained agricultural concepts.

Performance on Unseen Datasets. To evaluate our model’s generalization to unseen agricultural concepts, we select three datasets: the Sugarcane Leaf Disease dataset [36], the Potato Leaf Disease dataset [32], and the Plant Disease Recognition dataset [34]. We then assessed AgroGPT and our baseline using the AgroEvals protocol described previously. As illustrated in Figure 7(b), AgroGPT demonstrates significantly higher performance, indicating its ability to generalize beyond the datasets used

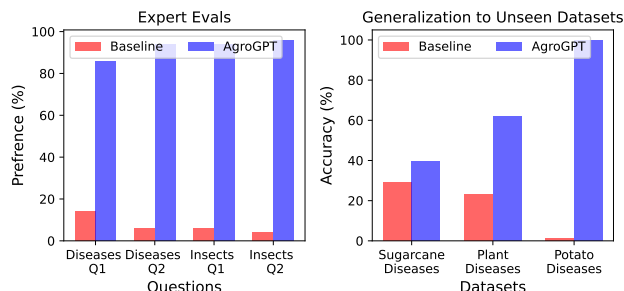


Figure 7. (a) Percentage of responses preferred by an expert for AgroGPT compared to the baseline. (b) Generalization performance of AgroGPT on datasets outside of its training set.

in expert tuning. However, it is crucial to note that AgroGPT may not perform as well on concepts that differ considerably, such as livestock, which are not part of AgroInstruct.

vExpert Evaluations. AgroGPT is an expert model that can answer complex agriculture-related questions in detail, making it difficult to assess its performance solely based on quantitative measures. Hence, to further understand AgroGPT’s performance, we perform expert evaluations by designing an interface, generating detail-oriented questions from Group 3 (plant diseases and insects that require more fine-grained knowledge of the field), and anonymously presenting answers of baseline and AgroGPT to the expert evaluators. For a thorough description, please refer to the Appendix. Results are shown in Tab. 7(a), which clearly illustrate that AgroGPT’s answers are overwhelmingly preferred as they answer with relevant details. On the other hand, the baseline model often gives generic answers that are less relevant to the problem.

Expert Tuning after Stage 1 vs. Stage 2. Another crucial aspect of our method is determining the stage at which expert tuning should be applied. Our method is based on visual instruction-tuning [21], which consists of two stages: image-text alignment (stage 1) and general visual instruction tuning (stage 2). We conducted experiments to evaluate the impact of expert tuning after each stage. The results are detailed in Tab. 5. Overall, AgroGPT achieves better results when expert tuning is performed after Stage 2. Further discussion on this topic can be found in the Appendix.

Low-Rank Adaptation vs. Fine-tuning. In Tab. 7, we compare Low-Rank Adaptation (LoRA) vs. full fine-tuning for expert tuning. The results show that fine-tuning is significantly better than LoRA fine-tuning. This is in line with previous findings [4].

4.3. Effect of Number of Expert-tuning Epochs.

To understand effect of increasing the number of expert-tuning epochs on the model’s performance, we conducted an experiment where we trained AgroGPT-3B for 1, 3, and 5 epochs. Fig. 8 compares the results of different numbers of epochs with the baseline (instruction-tuned Mipha3B [49]).

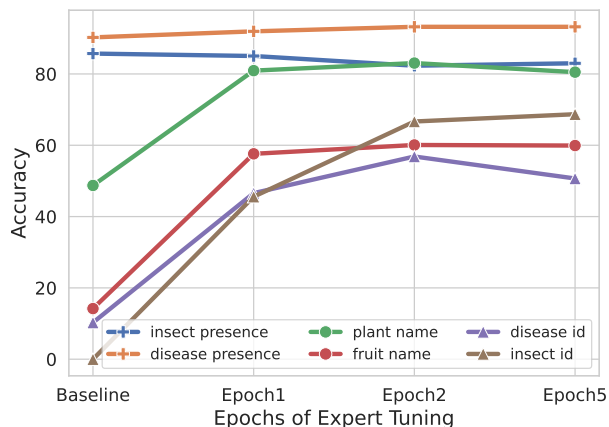


Figure 8. Effect of multiple epochs of expert-tuning on performance on three groups of questions for AgroGPT-3B.

Category		Fully Finetuned	LoRA Finetuned
Group1	Disease Presence	89.41	92.37
	Insect Presence	92.52	87.75
Group2	Fruits Name	47.96	47.38
	Plant Name	72.88	72.88
Group3	Disease Id	51.37	43.83
	Insect Id	34.01	32.65

Table 7. Comparison of full finetuning and LoRA finetuning for the AgroGPT-7B model across three groups of questions. LoRA performs on par with full finetuning for coarse and general questions from Groups 1 and 2. However, full finetuning performs significantly better than LoRA finetuning for fine-grained and field-specific questions from Group 3.

The results demonstrate that model’s performance on Group 1 questions is not affected by expert tuning. Conversely, more epochs of expert tuning enhance results for Group 2 and Group 3 questions, highlighting AgroInstruct’s capability to embed fine-grained and field-specific knowledge into the model. However, the improvement subsides after two epochs, indicating diminishing returns.

5. Conclusion

In this work, we introduce AgroGPT, an efficient large vision-language model tailored for the agricultural domain capable of conducting multi-turn, multimodal, open-ended conversations. To achieve this, we present a pipeline for generating expert-tuning data from vision-only datasets, following a self-instruct approach augmented with class-specific external knowledge. We constructed the AgroInstruct dataset using this pipeline and trained the efficient conversational model called AgroGPT. We also developed AgroEvals to evaluate the domain-specific capabilities of our model. AgroGPT excels at identifying fine-grained, domain-specific concepts and engaging in detailed conversations on agriculture-specific questions.

References

- [1] Cotton disease dataset, version 1. 2020. <https://www.kaggle.com/datasets/janmejybhoy/cottondisease-dataset>. 2
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [3] Ümit Atila, Murat Uçar, Kemal Akyol, and Emine Uçar. Plant leaf disease classification using efficientnet deep learning model. *Ecological Informatics*, 61:101182, 2021. 3
- [4] Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. Lora learns less and forgets less. *arXiv preprint arXiv: 2405.09673*, 2024. 8
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2
- [6] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2
- [7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023. 2
- [8] Maurilio Di Cicco, Ciro Potena, Giorgio Grisetti, and Alberto Pretto. Automatic model based dataset generation for fast and accurate crop and weeds detection. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5188–5195. IEEE, 2017. 2, 3
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2
- [10] Gaurav Dutta. Farm insect dataset. <https://www.kaggle.com/code/gauravduttakiit/class-dataset-dangerous-farm-insects/data>, 2023. 2, 6
- [11] Xuguang Feng, Chunjiang Zhao, Chunshan Wang, Huarui Wu, Yisheng Miao, and Jingjian Zhang. A vegetable leaf disease identification model based on image-text cross-modal feature fusion. *Frontiers in Plant Science*, 13:918940, 2022. 3
- [12] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*, 2023. 2
- [13] David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015. 2, 3
- [14] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastian Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 2023. 6
- [15] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 4, 6
- [16] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 2
- [17] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. *arXiv preprint arXiv:2311.15826*, 2023. 1, 2
- [18] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2
- [19] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 1, 3, 4
- [20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2
- [21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 3, 4, 5, 6, 7, 8
- [22] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017, 2023. 2
- [23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [24] Marquis. Fruits 100 dataset. "https://www.kaggle.com/datasets/marquis03/fruits-100", 2023. 2
- [25] Ashmitha Mohan. Insects image detection, 2022. Accessed: 2024-09-04. 3
- [26] Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*, 2023. 3
- [27] Sahal Mullappilly, Abdelrahman Shaker, Omkar Thawakar, Hisham Cholakkal, Rao Anwer, Salman Khan, and Fahad Khan. Arabic mini-ClimateGPT : A climate change and sustainability tailored Arabic LLM. In Houda Bouamor, Juan

- Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, Dec. 2023. 1, 2
- [28] Erik P Nyberg, Ann E Nicholson, Kevin B Korb, Michael Wybrow, Ingrid Zukerman, Steven Mascaro, Shreshth Thakur, Abraham Oshni Alvandi, Jeff Riley, Ross Pearson, et al. Bard: A structured technique for group elicitation of bayesian networks to support analytic reasoning. *Risk Analysis*, 42(6):1155–1178, 2022. 1
- [29] Bekhzod Olimov. Fruits classification using pytorch), 2024. Accessed: 2024-09-04. 3
- [30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 1
- [31] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 2
- [32] Muhammad Ardi Putra. Potato leaf disease dataset, 2020. Accessed: 2024-09-09. 3, 7
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 5, 6
- [34] Rashik Rahman. Plant disease recognition dataset, 2021. Accessed: 2024-09-09. 3, 7
- [35] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 3
- [36] Nirmal Sankalana. Sugarcane leaf disease dataset, 2024. Accessed: 2024-09-09. 3, 7
- [37] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. Plantdoc: A dataset for visual plant disease detection. Association for Computing Machinery, 2020. 2, 3, 6
- [38] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2
- [39] Poornima Singh Thakur, Tanuja Sheorey, and Aparajita Ojha. Vgg-icnn: A lightweight cnn model for crop disease identification. *Multimedia Tools and Applications*, 82(1):497–520, 2023. 3
- [40] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv: 2306.07971*, 2023. 1, 2
- [41] David Thulke, Yingbo Gao, Petrus Pelsers, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, et al. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*, 2024. 2
- [42] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2
- [43] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 2
- [44] Xianjun Yang, Junfeng Gao, Wenxin Xue, and Erik Alexandersson. Pillama: An open-source large language model for plant science. *arXiv preprint arXiv:2401.01600*, 2024. 3
- [45] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *ICCV*, 2023. 5, 6, 7
- [46] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv: 2303.00915*, 2023. 1
- [47] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv: 2306.05685*, 2023. 6
- [48] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2
- [49] Minjie Zhu, Yichen Zhu, Xin Liu, Ning Liu, Zhiyuan Xu, Chaomin Shen, Yaxin Peng, Zhicai Ou, Feifei Feng, and Jian Tang. Mipha: A comprehensive overhaul of multimodal assistant with small language models. *arXiv preprint arXiv: 2403.06199*, 2024. 5, 6, 7, 8