Supplementary Materials
AgMMU: A Comprehensive Agricultural Multimodal Understanding Benchmark

Aruna Gauba[1,2,5*] **Irene Pi**[1,3,5*] **Yunze Man**[1,4,5†] **Ziqi Pang**[1,4,5†]
Vikram S. Adve[1,4,5] **Yu-Xiong Wang**[1,4,5]

[1]University of Illinois Urbana-Champaign [2]Rice University [3]Carnegie Mellon University
[4]AIFARMS [5]Center for Digital Agriculture at UIUC

*† Equal Contribution † Project Lead

Contents

A Evaluated Large Vision Language Models **1**

B Dataset Curation Details **3**
B.1 Stage 1: Question Categorization 3
B.2 Stage 2: Information Extraction 6
B.3 Stage 3: Question Generation 9
B.4 Final Stage: Human Verification 10

C More Evaluation, Implementation, and Design Choices **10**

D More Dataset Visualization **11**

E Limitations and Future Work **12**

F Societal Impact **13**

A Evaluated Large Vision Language Models

Our evaluation and analysis are conducted mainly on the group of models listed in Table 2 in the main paper. We have chosen models such that they cover most of the popular and best-performing methods used by recent multimodal understanding work. In this part, we discuss all the models we have used in our experiments and explain their evaluation details, the public checkpoints we have chosen, and display the prompts we used to adapt the model to our datasets.

During evaluation, we chose to follow the standard prompt provided by the authors whenever possible for multiple-choice and short-answer questions. When the prompt is not provided for the model, we select a custom prompt that is created through several iterations of prompt engineering to select the one that produces the most effective results. The images are always included as the prefix.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Proprietary Models. We used three proprietary models in our evaluation: GPT-o4-mini [1], Gemini 1.5 Pro [9], and Claude 3 Haiku [10]. Below we note the model API version used for evaluation.

- GPT-o4-mini: May 13-15, 2025.
- Gemini 1.5 Pro: November 1-13, 2024.
- Claude 3 Haiku: November 13-14, 2024.

Cambrian-1 [12]. Cambrian-1 is a recent state-of-the-art model that excels at visual-centric tasks. This model explores combinations of vision encoders, text and image integration techniques, and instruction tuning strategies. We use the official implementation and checkpoint[1] with a LLaMA3-8B-Instruct LLM backbone model in our evaluation.

InternVL2 [11]. InternVL scales up the vision foundation model while aligning it with the backbone LLM, and is trained on web-scale image-text data to achieve strong performance across a variety of vision-centric tasks. We use the official implementation and checkpoint[2] with the InternViT-300M-448px vision backbone and Internlm2.5-7B-chat language backbone in our evaluation.

LLaMA-3.2 [4]. LLaMA-3.2 is the first collection of multimodal large language model from the LLaMA family that was previously text-only. The integration of vision involves utilizing cross-attention layers and a pre-trained vision encoder that feeds directly into the text-processor. The model follows a commonly used training recipe that includes pretraining on noisy image-text pairs and then high-quality knowledge enhanced pairs. Notably, the language-model parameters were frozen during the training of alignment of image and text to retain strong text-only capabilities. We use the official implementation and checkpoint[3] that uses a LLaMA-3.1 text-only language backbone in our evaluation. When evaluating the model, we choose to use a custom prompt since no standard prompt is provided.

LLaVA-NeXT [8]. LLaVA-NeXT expands on LLaVA by using the same instruction tuning method to give the model the ability to process and reason about multi-images, multi-grames, and multi-views. We use the official implementation and checkpoint[4] with LLaMA-3-8B Instruct as the language backbone in our evaluation.

LLaVA-OneVision [5]. LLaVA-OneVision builds on LLaVA-NeXT with the capability to analyze single images, multi-images, and video scenarios. Most impressively, it allows for video understanding through task transfer from images but this is not explored in our evaluation. We use the official implementation and checkpoint[5] that uses a base architecture consisting of SigLIP-SO400M-Patch14-384 and Qwen2-7b in our evaluation.

LLaVA-1.5-7B / LLaVA-1.5-13B [7]. LLaVA introduces the idea of instruction tuning a multimodal model with GPT-4 generated instruction-following data for associated images. This gives it the ability to achieve impressive abilities to act as an instruction-following general agent. We use the official implementation and checkpoints[6][7] with a CLIP ViT-L/14 vision backbone and Vicuna1.5-7B / Vicuna1.5-13B in our evaluation.

Qwen-VL-7B [2]. Qwen-VL is a large vision language model that has the ability to perform various vision-language tasks including image captioning, visual grounding and more, not only limited to question answering. This model is multi-lingual in Chinese and English and was pre-trained using an interleaved image-text technique. We use the official implementation and checkpoint[8] that uses Qwen-7B as the language backbone and CLIP ViT bigG/14 as the vision encoder in our evaluation.

VILA1.5-13B [6]. VILA is trained using an enhanced pre-training method that involves interleaved visual language data. Additionally, during the supervised fine-tuning stage, the data includes

---

[1] https://github.com/cambrian-mllm/cambrian

[2] https://huggingface.co/OpenGVLab/InternVL2-8B

[3] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision

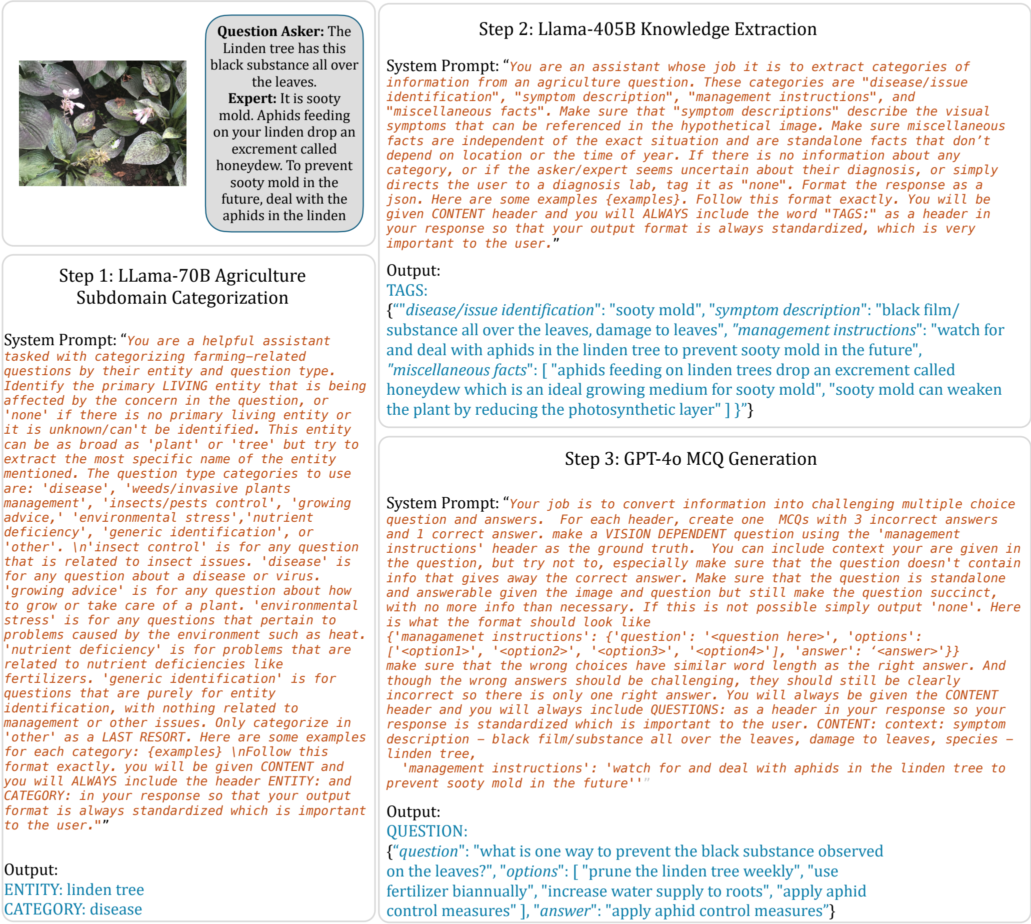
[4] https://huggingface.co/llava-hf/llama3-llava-next-8b-hf

[5] https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-hf

[6] https://huggingface.co/llava-hf/llava-1.5-7b-hf

[7] https://huggingface.co/llava-hf/llava-1.5-13b-hf

[8] https://huggingface.co/Qwen/Qwen-VL



Question Asker: The Linden tree has this black substance all over the leaves.
Expert: It is sooty mold. Aphids feeding on your linden drop an excrement called honeydew. To prevent sooty mold in the future, deal with the aphids in the linden

Step 1: LLama-70B Agriculture Subdomain Categorization

System Prompt: “You are a helpful assistant tasked with categorizing farming-related questions by their entity and question type. Identify the primary LIVING entity that is being affected by the concern in the question, or 'none' if there is no primary living entity or it is unknown/can't be identified. This entity can be as broad as 'plant' or 'tree' but try to extract the most specific name of the entity mentioned. The question type categories to use are: 'disease', 'weeds/invasive plants management', 'insects/pests control', 'growing advice,' 'environmental stress','nutrient deficiency', 'generic identification', or 'other'. \n'insect control' is for any question that is related to insect issues. 'disease' is for any question about a disease or virus. 'growing advice' is for any question about how to grow or take care of a plant. 'environmental stress' is for any questions that pertain to problems caused by the environment such as heat. 'nutrient deficiency' is for problems that are related to nutrient deficiencies like fertilizers. 'generic identification' is for questions that are purely for entity identification, with nothing related to management or other issues. Only categorize in 'other' as a LAST RESORT. Here are some examples for each category: {examples} \nFollow this format exactly. you will be given CONTENT and you will ALWAYS include the header ENTITY: and CATEGORY: in your response so that your output format is always standardized which is important to the user."”

Output:
ENTITY: linden tree
CATEGORY: disease

Step 2: Llama-405B Knowledge Extraction

System Prompt: “You are an assistant whose job it is to extract categories of information from an agriculture question. These categories are "disease/issue identification", "symptom description", "management instructions", and "miscellaneous facts". Make sure that "symptom descriptions" describe the visual symptoms that can be referenced in the hypothetical image. Make sure miscellaneous facts are independent of the exact situation and are standalone facts that don’t depend on location or the time of year. If there is no information about any category, or if the asker/expert seems uncertain about their diagnosis, or simply directs the user to a diagnosis lab, tag it as "none". Format the response as a json. Here are some examples {examples}. Follow this format exactly. You will be given CONTENT header and you will ALWAYS include the word "TAGS:" as a header in your response so that your output format is always standardized, which is very important to the user.”

Output:
TAGS:
{“"*disease/issue identification*": "sooty mold", "*symptom description*": "black film/substance all over the leaves, damage to leaves", *"management instructions*": "watch for and deal with aphids in the linden tree to prevent sooty mold in the future", *"miscellaneous facts*": ["aphids feeding on linden trees drop an excrement called honeydew which is an ideal growing medium for sooty mold", "sooty mold can weaken the plant by reducing the photosynthetic layer"] }”}

Step 3: GPT-4o MCQ Generation

System Prompt: “Your job is to convert information into challenging multiple choice question and answers. For each header, create one MCQs with 3 incorrect answers and 1 correct answer. make a VISION DEPENDENT question using the 'management instructions' header as the ground truth. You can include context your are given in the question, but try not to, especially make sure that the question doesn't contain info that gives away the correct answer. Make sure that the question is standalone and answerable given the image and question but still make the question succinct, with no more info than necessary. If this is not possible simply output 'none'. Here is what the format should look like
{'managamenet instructions': {'question': '<question here>', 'options': ['<option1>', '<option2>', '<option3>', '<option4>'], 'answer': ‘<answer>'}}
make sure that the wrong choices have similar word length as the right answer. And though the wrong answers should be challenging, they should still be clearly incorrect so there is only one right answer. You will always be given the CONTENT header and you will always include QUESTIONS: as a header in your response so your response is standardized which is important to the user. CONTENT: context: symptom description – black film/substance all over the leaves, damage to leaves, species – linden tree,
'management instructions': 'watch for and deal with aphids in the linden tree to prevent sooty mold in the future''”

Output:
QUESTION:
{“*question*": "what is one way to prevent the black substance observed on the leaves?", "*options*": ["prune the linden tree weekly", "use fertilizer biannually", "increase water supply to roots", "apply aphid control measures"], "*answer*": "apply aphid control measures”}

Figure A: Prompts used in different stages of our data curation pipeline.

text-only instruction data to help the model retain strong text-only capabilities. We use the official implementation and checkpoint[9] with a LLaMA3-8B LLM backbone and SigLIP-SO400M-Patch14-384 vision encoder in our evaluation.

B Dataset Curation Details

This section outlines the multi-stage curation pipeline of AgMMU and describes the prompts designed for each question type and subdomain.

B.1 Stage 1: Question Categorization

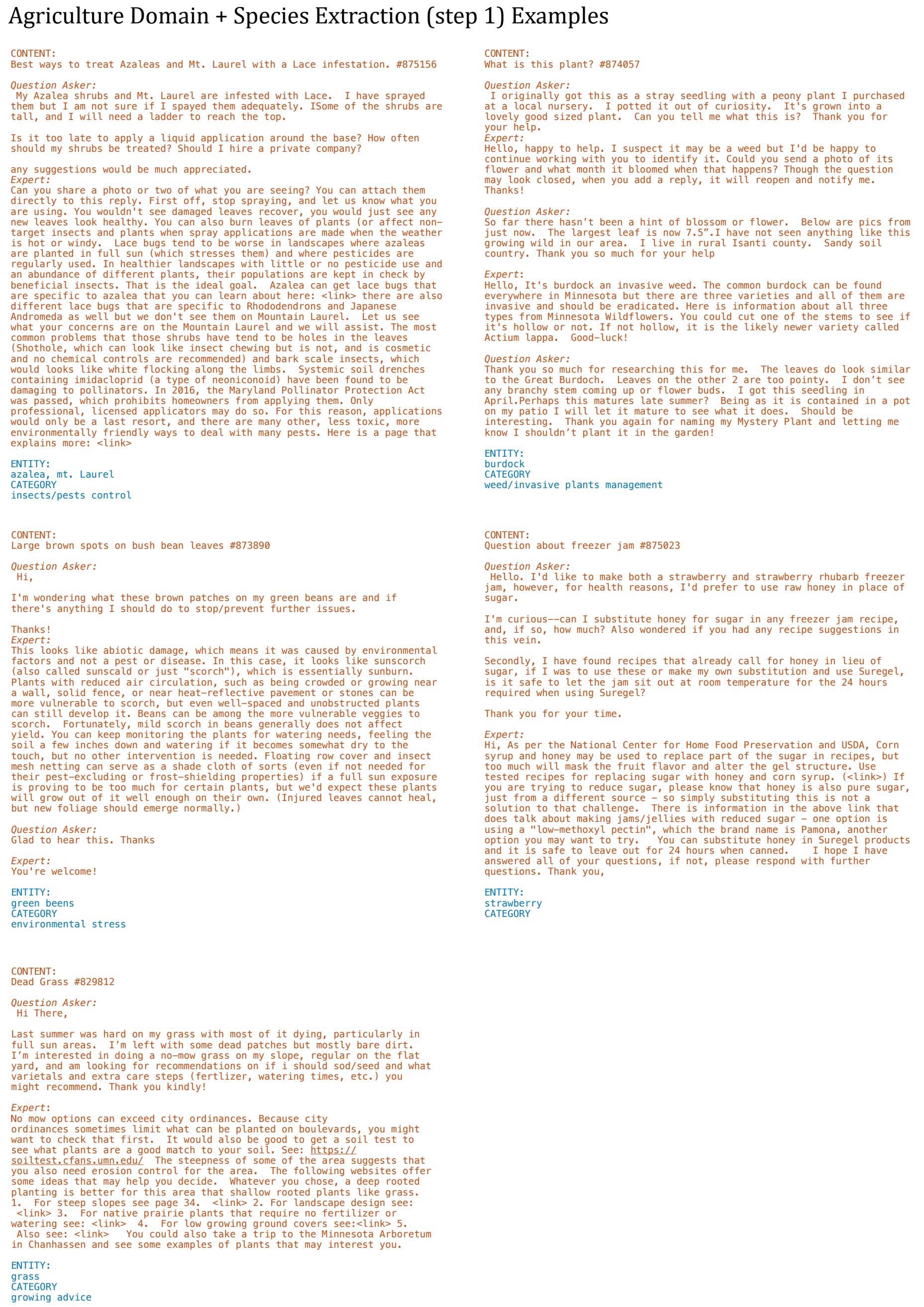
In the first step, we employ the Llama-70B model [4] to categorize questions into predefined agriculture subdomains while identifying the primary living entity affected by the query. Our systematically crafted prompt (Figure A) guides the model to extract the most specific living entity mentioned, such as “apple tree” or “honeybee,” or to assign “none” when the entity is unclear or absent.

The subdomains include *Disease, Weeds/Invasive Plant Management, Insect/Pest Control, Growing Advice, Environmental Stress, Nutrient Deficiency, Generic Identification*, and *Other*. Each subdomain is succinctly defined within the prompt, with illustrative examples provided in Figure B to address ambiguous or edge-case scenarios. The prompt enforces a standardized output format, ensuring consistency with the inclusion of “ENTITY:” and “CATEGORY:” headers.

To enhance robustness, the prompt includes examples of complex or overlapping cases, ensuring accurate classification even for questions that span multiple subdomains or lack explicit details. By

---

[9] https://github.com/NVlabs/VILA



Agriculture Domain + Species Extraction (step 1) Examples

CONTENT:
Best ways to treat Azaleas and Mt. Laurel with a Lace infestation. #875156

Question Asker:
My Azalea shrubs and Mt. Laurel are infested with Lace. I have sprayed them but I am not sure if I spayed them adequately. ISome of the shrubs are tall, and I will need a ladder to reach the top.

Is it too late to apply a liquid application around the base? How often should my shrubs be treated? Should I hire a private company?

any suggestions would be much appreciated.
Expert:
Can you share a photo or two of what you are seeing? You can attach them directly to this reply. First off, stop spraying, and let us know what you are using. You wouldn't see damaged leaves recover, you would just see any new leaves look healthy. You can also burn leaves of plants (or affect non-target insects and plants when spray applications are made when the weather is hot or windy. Lace bugs tend to be worse in landscapes where azaleas are planted in full sun (which stresses them) and where pesticides are regularly used. In healthier landscapes with little or no pesticide use and an abundance of different plants, their populations are kept in check by beneficial insects. That is the ideal goal. Azalea can get lace bugs that are specific to azalea that you can learn about here: <link> there are also different lace bugs that are specific to Rhododendrons and Japanese Andromeda as well but we don't see them on Mountain Laurel. Let us see what your concerns are on the Mountain Laurel and we will assist. The most common problems that those shrubs have tend to be holes in the leaves (Shothole, which can look like insect chewing but is not, and is cosmetic and no chemical controls are recommended) and bark scale insects, which would looks like white flocking along the limbs. Systemic soil drenches containing imidacloprid (a type of neoniconoid) have been found to be damaging to pollinators. In 2016, the Maryland Pollinator Protection Act was passed, which prohibits homeowners from applying them. Only professional, licensed applicators may do so. For this reason, applications would only be a last resort, and there are many other, less toxic, more environmentally friendly ways to deal with many pests. Here is a page that explains more: <link>

ENTITY:
azalea, mt. Laurel
CATEGORY
insects/pests control

CONTENT:
Large brown spots on bush bean leaves #873890

Question Asker:
Hi,

I'm wondering what these brown patches on my green beans are and if there's anything I should do to stop/prevent further issues.

Thanks!
Expert:
This looks like abiotic damage, which means it was caused by environmental factors and not a pest or disease. In this case, it looks like sunscorch (also called sunscald or just "scorch"), which is essentially sunburn. Plants with reduced air circulation, such as being crowded or growing near a wall, solid fence, or near heat-reflective pavement or stones can be more vulnerable to scorch, but even well-spaced and unobstructed plants can still develop it. Beans can be among the more vulnerable veggies to scorch. Fortunately, mild scorch in beans generally does not affect yield. You can keep monitoring the plants for watering needs, feeling the soil a few inches down and watering if it becomes somewhat dry to the touch, but no other intervention is needed. Floating row cover and insect mesh netting can serve as a shade cloth of sorts (even if not needed for their pest-excluding or frost-shielding properties) if a full sun exposure is proving to be too much for certain plants, but we'd expect these plants will grow out of it well enough on their own. (Injured leaves cannot heal, but new foliage should emerge normally.)

Question Asker:
Glad to hear this. Thanks

Expert:
You're welcome!

ENTITY:
green beens
CATEGORY
environmental stress

CONTENT:
Dead Grass #829812

Question Asker:
Hi There,

Last summer was hard on my grass with most of it dying, particularly in full sun areas. I’m left with some dead patches but mostly bare dirt. I’m interested in doing a no-mow grass on my slope, regular on the flat yard, and am looking for recommendations on if i should sod/seed and what varietals and extra care steps (fertlizer, watering times, etc.) you might recommend. Thank you kindly!

Expert:
No mow options can exceed city ordinances. Because city ordinances sometimes limit what can be planted on boulevards, you might want to check that first. It would also be good to get a soil test to see what plants are a good match to your soil. See: https://soiltest.cfans.umn.edu/ The steepness of some of the area suggests that you also need erosion control for the area. The following websites offer some ideas that may help you decide. Whatever you chose, a deep rooted planting is better for this area that shallow rooted plants like grass. 1. For steep slopes see page 34. <link> 2. For landscape design see: <link> 3. For native prairie plants that require no fertilizer or watering see: <link> 4. For low growing ground covers see:<link> 5. Also see: <link> You could also take a trip to the Minnesota Arboretum in Chanhassen and see some examples of plants that may interest you.

ENTITY:
grass
CATEGORY
growing advice

CONTENT:
What is this plant? #874057

Question Asker:
I originally got this as a stray seedling with a peony plant I purchased at a local nursery. I potted it out of curiosity. It's grown into a lovely good sized plant. Can you tell me what this is? Thank you for your help.
Expert:
Hello, happy to help. I suspect it may be a weed but I'd be happy to continue working with you to identify it. Could you send a photo of its flower and what month it bloomed when that happens? Though the question may look closed, when you add a reply, it will reopen and notify me. Thanks!

Question Asker:
So far there hasn’t been a hint of blossom or flower. Below are pics from just now. The largest leaf is now 7.5”.I have not seen anything like this growing wild in our area. I live in rural Isanti county. Sandy soil country. Thank you so much for your help

Expert:
Hello, It's burdock an invasive weed. The common burdock can be found everywhere in Minnesota but there are three varieties and all of them are invasive and should be eradicated. Here is information about all three types from Minnesota Wildflowers. You could cut one of the stems to see if it's hollow or not. If not hollow, it is the likely newer variety called Actium lappa. Good-luck!

Question Asker:
Thank you so much for researching this for me. The leaves do look similar to the Great Burdoch. Leaves on the other 2 are too pointy. I don’t see any branchy stem coming up or flower buds. I got this seedling in April.Perhaps this matures late summer? Being as it is contained in a pot on my patio I will let it mature to see what it does. Should be interesting. Thank you again for naming my Mystery Plant and letting me know I shouldn’t plant it in the garden!

ENTITY:
burdock
CATEGORY
weed/invasive plants management

CONTENT:
Question about freezer jam #875023

Question Asker:
Hello. I'd like to make both a strawberry and strawberry rhubarb freezer jam, however, for health reasons, I'd prefer to use raw honey in place of sugar.

I'm curious--can I substitute honey for sugar in any freezer jam recipe, and, if so, how much? Also wondered if you had any recipe suggestions in this vein.

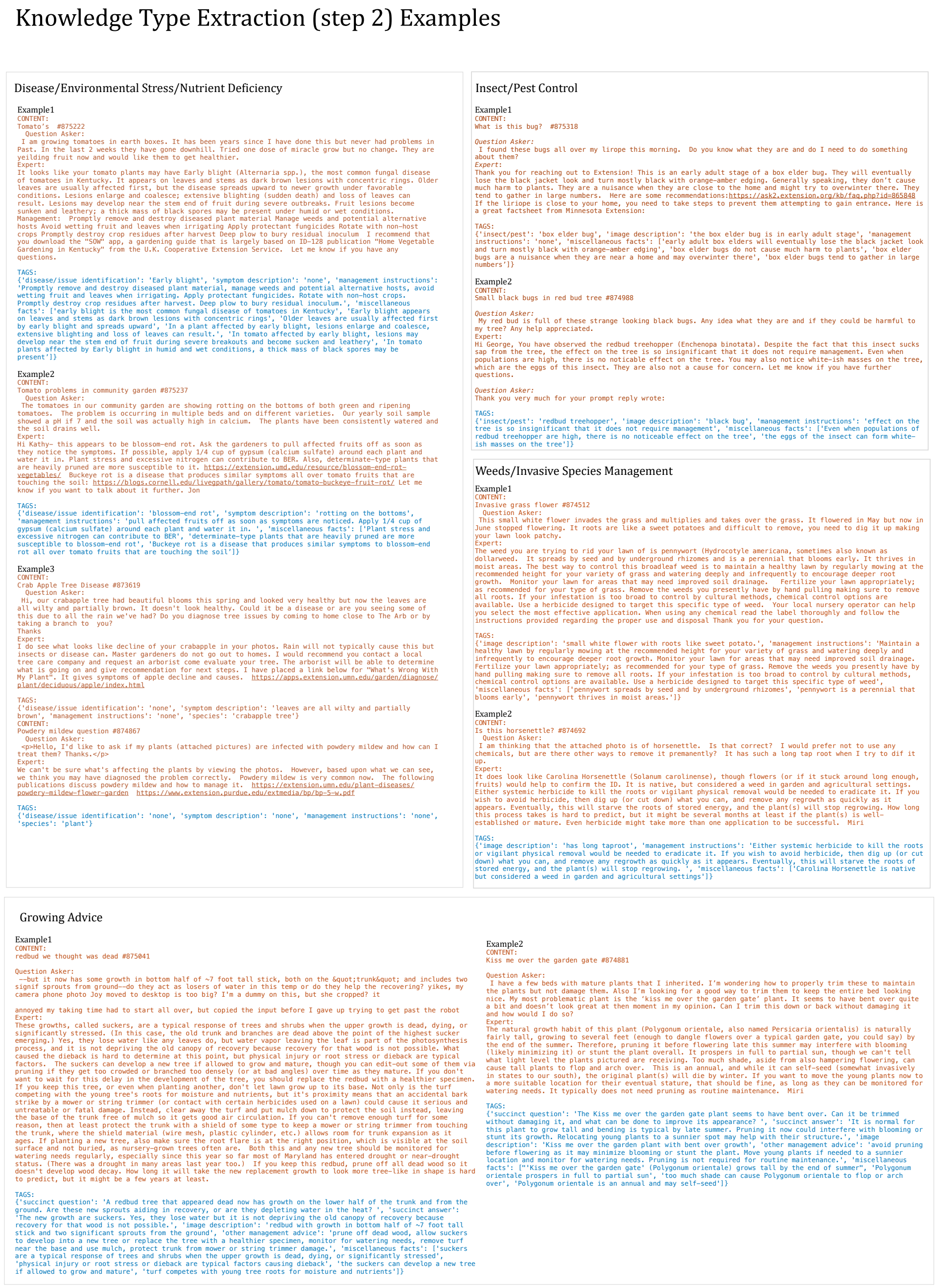
Secondly, I have found recipes that already call for honey in lieu of sugar, if I was to use these or make my own substitution and use Suregel, is it safe to let the jam sit out at room temperature for the 24 hours required when using Suregel?

Thank you for your time.

Expert:
Hi, As per the National Center for Home Food Preservation and USDA, Corn syrup and honey may be used to replace part of the sugar in recipes, but too much will mask the fruit flavor and alter the gel structure. Use tested recipes for replacing sugar with honey and corn syrup. (<link>) If you are trying to reduce sugar, please know that honey is also pure sugar, just from a different source - so simply substituting this is not a solution to that challenge. There is information in the above link that does talk about making jams/jellies with reduced sugar - one option is using a "low-methoxyl pectin", which the brand name is Pamona, another option you may want to try. You can substitute honey in Suregel products and it is safe to leave out for 24 hours when canned. I hope I have answered all of your questions, if not, please respond with further questions. Thank you,

ENTITY:
strawberry
CATEGORY

Figure B: Examples included in prompt during the agriculture domain categorization (step 1).



Knowledge Type Extraction (step 2) Examples

Disease/Environmental Stress/Nutrient Deficiency

Example1

CONTENT:
Tomato’s #875222
Question Asker:
I am growing tomatoes in earth boxes. It has been years since I have done this but never had problems in Past. In the last 2 weeks they have gone downhill. Tried one dose of miracle grow but no change. They are yeilding fruit now and would like them to get healthier.
Expert:
It looks like your tomato plants may have Early blight (Alternaria spp.), the most common fungal disease of tomatoes in Kentucky. It appears on leaves and stems as dark brown lesions with concentric rings. Older leaves are usually affected first, but the disease spreads upward to newer growth under favorable conditions. Lesions enlarge and coalesce; extensive blighting (sudden death) and loss of leaves can result. Lesions may develop near the stem end of fruit during severe outbreaks. Fruit lesions become sunken and leathery; a thick mass of black spores may be present under humid or wet conditions.
Management: Promptly remove and destroy diseased plant material Manage weeds and potential alternative hosts Avoid wetting fruit and leaves when irrigating Apply protectant fungicides Rotate with non-host crops Promptly destroy crop residues after harvest Deep plow to bury residual inoculum I recommend that you download the "SOW" app, a gardening guide that is largely based on ID-128 publication "Home Vegetable Gardening in Kentucky" from the U.K. Cooperative Extension Service. Let me know if you have any questions.

TAGS:
{'disease/issue identification': 'Early blight', 'symptom description': 'none', 'management instructions': 'Promptly remove and destroy diseased plant material, manage weeds and potential alternative hosts, avoid wetting fruit and leaves when irrigating. Apply protectant fungicides. Rotate with non-host crops. Promptly destroy crop residues after harvest. Deep plow to bury residual inoculum.', 'miscellaneous facts': ['early blight is the most common fungal disease of tomatoes in Kentucky', 'Early blight appears on leaves and stems as dark brown lesions with concentric rings', 'Older leaves are usually affected first by early blight and spreads upward', 'In a plant affected by early blight, lesions enlarge and coalesce, extensive blighting and loss of leaves can result.', 'In tomato affected by early blight, lesions may develop near the stem end of fruit during severe breakouts and become sucken and leathery', 'In tomato plants affected by Early blight in humid and wet conditions, a thick mass of black spores may be present’]}

Example2

CONTENT:
Tomato problems in community garden #875237
Question Asker:
The tomatoes in our community garden are showing rotting on the bottoms of both green and ripening tomatoes. The problem is occurring in multiple beds and on different varieties. Our yearly soil sample showed a pH if 7 and the soil was actually high in calcium. The plants have been consistently watered and the soil drains well.
Expert:
Hi Kathy- this appears to be blossom-end rot. Ask the gardeners to pull affected fruits off as soon as they notice the symptoms. If possible, apply 1/4 cup of gypsum (calcium sulfate) around each plant and water it in. Plant stress and excessive nitrogen can contribute to BER. Also, determinate-type plants that are heavily pruned are more susceptible to it. https://extension.umd.edu/resource/blossom-end-rot-vegetables/ Buckeye rot is a disease that produces similar symptoms all over tomato fruits that are touching the soil: https://blogs.cornell.edu/livegpath/gallery/tomato/tomato-buckeye-fruit-rot/ Let me know if you want to talk about it further. Jon

TAGS:
{'disease/issue identification': 'blossom-end rot', 'symptom description': 'rotting on the bottoms', 'management instructions': 'pull affected fruits off as soon as symptoms are noticed. Apply 1/4 cup of gypsum (calcium sulfate) around each plant and water it in. ', 'miscellaneous facts': ['Plant stress and excessive nitrogen can contribute to BER', 'determinate-type plants that are heavily pruned are more susceptible to blossom-end rot', 'Buckeye rot is a disease that produces similar symptoms to blossom-end rot all over tomato fruits that are touching the soil’]}

Example3

CONTENT:
Crab Apple Tree Disease #873619
Question Asker:
Hi, our crabapple tree had beautiful blooms this spring and looked very healthy but now the leaves are all wilty and partially brown. It doesn't look healthy. Could it be a disease or are you seeing some of this due to all the rain we've had? Do you diagnose tree issues by coming to home close to The Arb or by taking a branch to you?
Thanks
Expert:
I do see what looks like decline of your crabapple in your photos. Rain will not typically cause this but insects or disease can. Master gardeners do not go out to homes. I would recommend you contact a local tree care company and request an arborist come evaluate your tree. The arborist will be able to determine what is going on and give recommendation for next steps. I have placed a link below for "What's Wrong With My Plant". It gives symptoms of apple decline and causes. https://apps.extension.umn.edu/garden/diagnose/plant/deciduous/apple/index.html

TAGS:
{'disease/issue identification': 'none', 'symptom description': 'leaves are all wilty and partially brown', 'management instructions': 'none', 'species': 'crabapple tree'}
CONTENT:
Powdery mildew question #874867
Question Asker:
<p>Hello, I'd like to ask if my plants (attached pictures) are infected with powdery mildew and how can I treat them? Thanks.</p>
Expert:
We can't be sure what's affecting the plants by viewing the photos. However, based upon what we can see, we think you may have diagnosed the problem correctly. Powdery mildew is very common now. The following publications discuss powdery mildew and how to manage it. https://extension.umn.edu/plant-diseases/powdery-mildew-flower-garden https://www.extension.purdue.edu/extmedia/bp/bp-5-w.pdf

TAGS:
{'disease/issue identification': 'none', 'symptom description': 'none', 'management instructions': 'none', 'species': 'plant'}

Insect/Pest Control

Example1

CONTENT:
What is this bug? #875318

Question Asker:
I found these bugs all over my lirope this morning. Do you know what they are and do I need to do something about them?
Expert:
Thank you for reaching out to Extension! This is an early adult stage of a box elder bug. They will eventually lose the black jacket look and turn mostly black with orange-amber edging. Generally speaking, they don’t cause much harm to plants. They are a nuisance when they are close to the home and might try to overwinter there. They tend to gather in large numbers. Here are some recommendations:https://ask2.extension.org/kb/faq.php?id=865848 If the liriope is close to your home, you need to take steps to prevent them attempting to gain entrance. Here is a great factsheet from Minnesota Extension:

TAGS:
{'insect/pest': 'box elder bug', 'image description': 'the box elder bug is in early adult stage', 'management instructions': 'none', 'miscellaneous facts': ['early adult box elders will eventually lose the black jacket look and turn mostly black with orange-amber edging', 'box elder bugs do not cause much harm to plants', 'box elder bugs are a nuisance when they are near a home and may overwinter there', 'box elder bugs tend to gather in large numbers’]}

Example2

CONTENT:
Small black bugs in red bud tree #874988

Question Asker:
My red bud is full of these strange looking black bugs. Any idea what they are and if they could be harmful to my tree? Any help appreciated.
Expert:
Hi George, You have observed the redbud treehopper (Enchenopa binotata). Despite the fact that this insect sucks sap from the tree, the effect on the tree is so insignificant that it does not require management. Even when populations are high, there is no noticable effect on the tree. You may also notice white-ish masses on the tree, which are the eggs of this insect. They are also not a cause for concern. Let me know if you have further questions.

Question Asker:
Thank you very much for your prompt reply wrote:

TAGS:
{'insect/pest': 'redbud treehopper', 'image description': 'black bug', 'management instructions': 'effect on the tree is so insignificant that it does not require management', 'miscellaneous facts': ['Even when populations of redbud treehopper are high, there is no noticeable effect on the tree', 'the eggs of the insect can form white-ish masses on the tree']}

Weeds/Invasive Species Management

Example1

CONTENT:
Invasive grass flower #874512
Question Asker:
This small white flower invades the grass and multiplies and takes over the grass. It flowered in May but now in June stopped flowering. It roots are like a sweet potatoes and difficult to remove, you need to dig it up making your lawn look patchy.
Expert:
The weed you are trying to rid your lawn of is pennywort (Hydrocotyle americana, sometimes also known as dollarweed. It spreads by seed and by underground rhizomes and is a perennial that blooms early. It thrives in moist areas. The best way to control this broadleaf weed is to maintain a healthy lawn by regularly mowing at the recommended height for your variety of grass and watering deeply and infrequently to encourage deeper root growth. Monitor your lawn for areas that may need improved soil drainage. Fertilize your lawn appropriately; as recommended for your type of grass. Remove the weeds you presently have by hand pulling making sure to remove all roots. If your infestation is too broad to control by cultural methods, chemical control options are available. Use a herbicide designed to target this specific type of weed. Your local nursery operator can help you select the most effective application. When using any chemical read the label thoroughly and follow the instructions provided regarding the proper use and disposal Thank you for your question.

TAGS:
{'image description': 'small white flower with roots like sweet potato.', 'management instructions': 'Maintain a healthy lawn by regularly mowing at the recommended height for your variety of grass and watering deeply and infrequently to encourage deeper root growth. Monitor your lawn for areas that may need improved soil drainage. Fertilize your lawn appropriately; as recommended for your type of grass. Remove the weeds you presently have by hand pulling making sure to remove all roots. If your infestation is too broad to control by cultural methods, chemical control options are available. Use a herbicide designed to target this specific type of weed', 'miscellaneous facts': ['pennywort spreads by seed and by underground rhizomes', 'pennywort is a perennial that blooms early', 'pennywort thrives in moist areas.']}

Example2

CONTENT:
Is this horsenettle? #874692
Question Asker:
I am thinking that the attached photo is of horsenettle. Is that correct? I would prefer not to use any chemicals, but are there other ways to remove it premanently? It has such a long tap root when I try to dif it up.
Expert:
It does look like Carolina Horsenettle (Solanum carolinense), though flowers (or if it stuck around long enough, fruits) would help to confirm the ID. It is native, but considered a weed in garden and agricultural settings. Either systemic herbicide to kill the roots or vigilant physical removal would be needed to eradicate it. If you wish to avoid herbicide, then dig up (or cut down) what you can, and remove any regrowth as quickly as it appears. Eventually, this will starve the roots of stored energy, and the plant(s) will stop regrowing. How long this process takes is hard to predict, but it might be several months at least if the plant(s) is well-established or mature. Even herbicide might take more than one application to be successful. Miri

TAGS:
{'image description': 'has long taproot', 'management instructions': 'Either systemic herbicide to kill the roots or vigilant physical removal would be needed to eradicate it. If you wish to avoid herbicide, then dig up (or cut down) what you can, and remove any regrowth as quickly as it appears. Eventually, this will starve the roots of stored energy, and the plant(s) will stop regrowing. ', 'miscellaneous facts': ['Carolina Horsenettle is native but considered a weed in garden and agricultural settings']}

Growing Advice

Example1

CONTENT:
redbud we thought was dead #875041

Question Asker:
--but it now has some growth in bottom half of ~7 foot tall stick, both on the &quot;trunk&quot; and includes two signif sprouts from ground--do they act as losers of water in this temp or do they help the recovering? yikes, my camera phone photo Joy moved to desktop is too big? I'm a dummy on this, but she cropped? it

annoyed my taking time had to start all over, but copied the input before I gave up trying to get past the robot
Expert:
These growths, called suckers, are a typical response of trees and shrubs when the upper growth is dead, dying, or significantly stressed. (In this case, the old trunk and branches are dead above the point of the highest sucker emerging.) Yes, they lose water like any leaves do, but water vapor leaving the leaf is part of the photosynthesis process, and it is not depriving the old canopy of recovery because recovery for that wood is not possible. What caused the dieback is hard to determine at this point, but physical injury or root stress or dieback are typical factors. The suckers can develop a new tree if allowed to grow and mature, though you can edit-out some of them via pruning if they get too crowded or branched too densely (or at bad angles) over time as they mature. If you don't want to wait for this delay in the development of the tree, you should replace the redbud with a healthier specimen. If you keep this tree, or even when planting another, don't let lawn grow up to its base. Not only is the turf competing with the young tree's roots for moisture and nutrients, but it's proximity means that an accidental bark strike by a mower or string trimmer (or contact with certain herbicides used on a lawn) could cause it serious and untreatable or fatal damage. Instead, clear away the turf and put mulch down to protect the soil instead, leaving the base of the trunk free of mulch so it gets good air circulation. If you can't remove enough turf for some reason, then at least protect the trunk with a shield of some type to keep a mower or string trimmer from touching the trunk, where the shield material (wire mesh, plastic cylinder, etc.) allows room for trunk expansion as it ages. If planting a new tree, also make sure the root flare is at the right position, which is visible at the soil surface and not buried, as nursery-grown trees often are. Both this and any new tree should be monitored for watering needs regularly, especially since this year so far most of Maryland has entered drought or near-drought status. (There was a drought in many areas last year too.) If you keep this redbud, prune off all dead wood so it doesn't develop wood decay. How long it will take the new replacement growth to look more tree-like in shape is hard to predict, but it might be a few years at least.

TAGS:
{'succinct question': 'A redbud tree that appeared dead now has growth on the lower half of the trunk and from the ground. Are these new sprouts aiding in recovery, or are they depleting water in the heat? ', 'succinct answer': 'The new growth are suckers. Yes, they lose water but it is not depriving the old canopy of recovery because recovery for that wood is not possible.', 'image description': 'redbud with growth in bottom half of ~7 foot tall stick and two significant sprouts from the ground', 'other management advice': 'prune off dead wood, allow suckers to develop into a new tree or replace the tree with a healthier specimen, monitor for watering needs, remove turf near the base and use mulch, protect trunk from mower or string trimmer damage.', 'miscellaneous facts': ['suckers are a typical response of trees and shrubs when the upper growth is dead, dying, or significantly stressed', 'physical injury or root stress or dieback are typical factors causing dieback', 'the suckers can develop a new tree if allowed to grow and mature', 'turf competes with young tree roots for moisture and nutrients']}

Example2

CONTENT:
Kiss me over the garden gate #874881

Question Asker:
I have a few beds with mature plants that I inherited. I’m wondering how to properly trim these to maintain the plants but not damage them. Also I’m looking for a good way to trim them to keep the entire bed looking nice. My most problematic plant is the ‘kiss me over the garden gate’ plant. It seems to have bent over quite a bit and doesn’t look great at then moment in my opinion. Can I trim this down or back without damaging it and how would I do so?
Expert:
The natural growth habit of this plant (Polygonum orientale, also named Persicaria orientalis) is naturally fairly tall, growing to several feet (enough to dangle flowers over a typical garden gate, you could say) by the end of the summer. Therefore, pruning it before flowering late this summer may interfere with blooming (likely minimizing it) or stunt the plant overall. It prospers in full to partial sun, though we can't tell what light level the plants pictured are receiving. Too much shade, aside from also hampering flowering, can cause tall plants to flop and arch over. This is an annual, and while it can self-seed (somewhat invasively in states to our south), the original plant(s) will die by winter. If you want to move the young plants now to a more suitable location for their eventual stature, that should be fine, as long as they can be monitored for watering needs. It typically does not need pruning as routine maintenance. Miri

TAGS:
{'succinct question': 'The Kiss me over the garden gate plant seems to have bent over. Can it be trimmed without damaging it, and what can be done to improve its appearance? ', 'succinct answer': 'It is normal for this plant to grow tall and bending is typical by late summer. Pruning it now could interfere with blooming or stunt its growth. Relocating young plants to a sunnier spot may help with their structure.', 'image description': 'Kiss me over the garden plant with bent over growth', 'other management advice': 'avoid pruning before flowering as it may minimize blooming or stunt the plant. Move young plants if needed to a sunnier location and monitor for watering needs. Pruning is not required for routine maintenance.', 'miscellaneous facts': ["'Kiss me over the garden gate' (Polygonum orientale) grows tall by the end of summer", 'Polygonum orientale prospers in full to partial sun', 'too much shade can cause Polygonum orientale to flop or arch over', 'Polygonum orientale is an annual and may self-seed']}

Figure C: Examples included in prompt during the knowledge extraction (step 2) based on agriculture subdomain type.

Growing Advice

`You are an assistant whose job it is to extract categories of information from an agriculture question. These categories are "succinct question", "succinct answer", "image description", "other management advice", and "miscellaneous facts". Make sure that "image descriptions" describe the visual symptoms that can be referenced in the hypothetical image. Make sure miscellaneous facts are independent of the exact situation and are standalone facts that don't depend on location or the time of year. If there is no information about any category, or if the asker/expert seems very uncertain about the information, tag it as "none". Format the response as a json. Here are some examples {examples}. Follow this format exactly. You will be given CONTENT header and you will ALWAYS include the word "TAGS:" as a header in your response so that your output format is always standardized, which is very important to the user.`

Insects/Pest Control

`You are an assistant whose job it is to extract categories of information from an agriculture question. These categories are "insect/pest", "image description", "management instructions", and "miscellaneous facts". Make sure that "image descriptions" describe the visual qualities that can be referenced in the hypothetical image. Make sure miscellaneous facts are independent of the exact situation and are standalone facts that don't depend on location or the time of year. If there is no information about any category, or if the asker/expert seems very uncertain about the information, tag it as "none". Format the response as a json. Here are some examples {examples}. Follow this format exactly. You will be given CONTENT header and you will ALWAYS include the word "TAGS:" as a header in your response so that your output format is always standardized, which is very important to the user.'`

Weeds/Invasive Species Management

`You are an assistant whose job it is to extract categories of information from an agriculture question. These categories are "image description", "management instructions", and "miscellaneous facts". Make sure that "symptom descriptions" describe the visual symptoms that can be referenced in the hypothetical image. Make sure miscellaneous facts are independent of the exact situation and are standalone facts that don't depend on location or the time of year. If there is no information about any category, or if the asker/expert seems very uncertain about the information, tag it as "none". Format the response as a json. Here are some examples {examples}. Follow this format exactly. You will be given CONTENT header and you will ALWAYS include the word "TAGS:" as a header in your response so that your output format is always standardized, which is very important to the user.' else: return None`

Disease/Environmental Stress/Nutrient Deficiency

`You are an assistant whose job it is to extract categories of information from an agriculture question. These categories are "disease/issue identification", "symptom description", "management instructions", and "miscellaneous facts". Make sure that "symptom descriptions" describe the visual symptoms that can be referenced in the hypothetical image. Make sure miscellaneous facts are independent of the exact situation and are standalone facts that don't depend on location or the time of year. If there is no information about any category, or if the asker/expert seems uncertain about their diagnosis, or simply directs the user to a diagnosis lab, tag it as "none". Format the response as a json. Here are some examples {examples}. Follow this format exactly. You will be given CONTENT header and you will ALWAYS include the word "TAGS:" as a header in your response so that your output format is always standardized, which is very important to the user.`

Figure D: Prompts used for each subdomain during knowledge extraction (step 2).

embedding these clarifications, the design supports reliable categorization across diverse agricultural contexts.

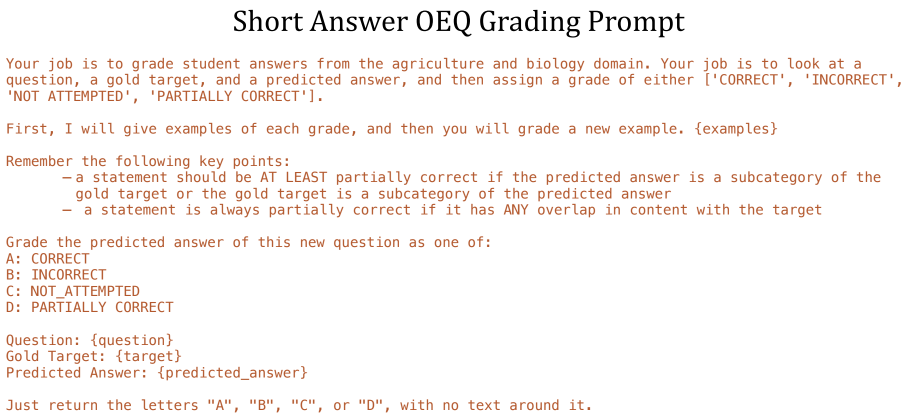
B.2 Stage 2: Information Extraction

In the second step, we design prompts to extract granular categories of information from agricultural questions. These categories are tailored to the specific subdomain identified in Step 1, ensuring that the extracted information is both relevant and actionable.

Weeds/Invasive Plants Management. For the "weeds/invasive plants management" subdomain, the extraction focuses on: (1) *Image Description*, visual characteristics of the weed or invasive plant, (2) *Management Instructions*, actionable strategies for control, and (3) *Miscellaneous Facts*, contextual expert insights. The name of the weed itself is already extracted in Step 1. This ensures that the emphasis remains on descriptions, actionable measures, and expert knowledge.

Insects/Pests Control. For this subdomain, the categories include: (1) *Insect/Pest*, identifying the pest in focus, (2) *Image Description*, visual traits of the pest or evidence of damage, (3) *Management Instructions*, guidance for mitigation, and (4) *Miscellaneous Facts*, contextual expert insights. The primary plant affected, if exists, is identified in Step 1, thus this step concentrates on pest-specific details, such as visual features or damage patterns, and the corresponding management strategies.

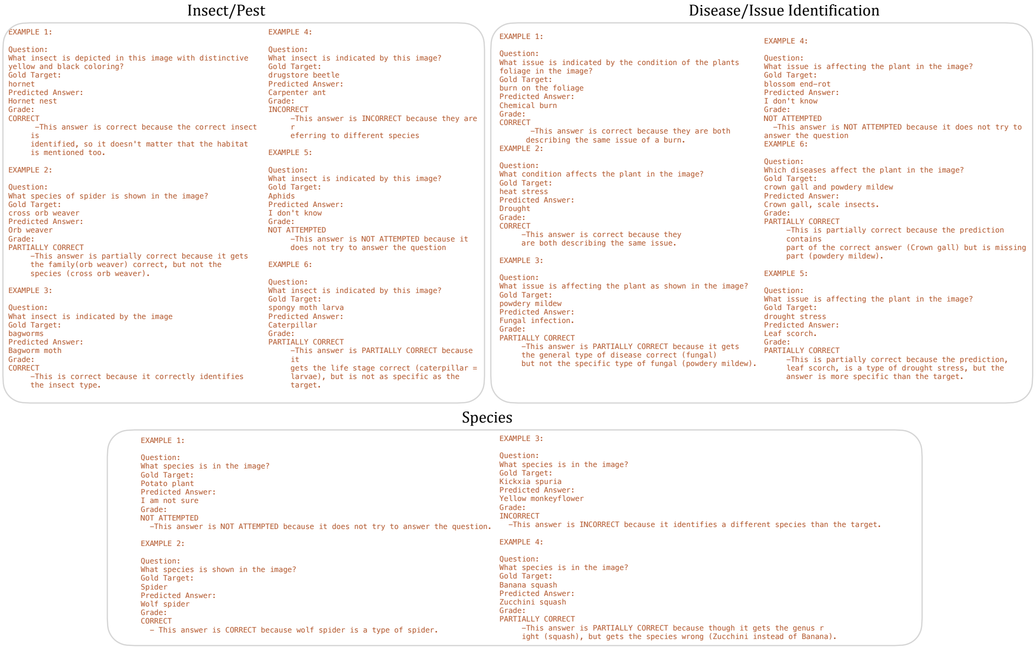
Nutrient Deficiency, Disease, Environmental Stress. For these subdomains, we group them due to shared characteristics. The extracted categories are: (1) *Disease/Issue Identification*, specifying the underlying cause, (2) *Symptom Description*, observable signs such as discoloration or stunted growth, (3) *Management Instructions*, remediation or prevention strategies, and (4) *Miscellaneous Facts*, contextual expert insights. These subdomains are defined by their symptomatic presentation, the underlying conditions, and the need for targeted management interventions.



Short Answer OEQ Grading Prompt

```
Your job is to grade student answers from the agriculture and biology domain. Your job is to look at a
question, a gold target, and a predicted answer, and then assign a grade of either ['CORRECT', 'INCORRECT',
'NOT ATTEMPTED', 'PARTIALLY CORRECT'].

First, I will give examples of each grade, and then you will grade a new example. {examples}

Remember the following key points:
       −a statement should be AT LEAST partially correct if the predicted answer is a subcategory of the
        gold target or the gold target is a subcategory of the predicted answer
       −  a statement is always partially correct if it has ANY overlap in content with the target

Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT
C: NOT_ATTEMPTED
D: PARTIALLY CORRECT

Question: {question}
Gold Target: {target}
Predicted Answer: {predicted_answer}

Just return the letters "A", "B", "C", or "D", with no text around it.
```

Figure E: Grading prompt for our LLM-as-judge on short-answer OEQ.



Insect/Pest

```
EXAMPLE 1:

Question:
What insect is depicted in this image with distinctive
yellow and black coloring?
Gold Target:
hornet
Predicted Answer:
Hornet nest
Grade:
CORRECT
      −This answer is correct because the correct insect
     is
     identified, so it doesn't matter that the habitat
     is mentioned too.

EXAMPLE 2:

Question:
What species of spider is shown in the image?
Gold Target:
cross orb weaver
Predicted Answer:
Orb weaver
Grade:
PARTIALLY CORRECT
     −This answer is partially correct because it gets
     the family(orb weaver) correct, but not the
     species (cross orb weaver).

EXAMPLE 3:

Question:
What insect is indicated by the image
Gold Target:
bagworms
Predicted Answer:
Bagworm moth
Grade:
CORRECT
     −This is correct because it correctly identifies
     the insect type.
```

```
EXAMPLE 4:

Question:
What insect is indicated by this image?
Gold Target:
drugstore beetle
Predicted Answer:
Carpenter ant
Grade:
INCORRECT
     −This answer is INCORRECT because they are
     r
     eferring to different species

EXAMPLE 5:

Question:
What insect is indicated by this image?
Gold Target:
Aphids
Predicted Answer:
I don't know
Grade:
NOT ATTEMPTED
     −This answer is NOT ATTEMPTED because it
     does not try to answer the question

EXAMPLE 6:

Question:
What insect is indicated by this image?
Gold Target:
spongy moth larva
Predicted Answer:
Caterpillar
Grade:
PARTIALLY CORRECT
     −This answer is PARTIALLY CORRECT because
     it
     gets the life stage correct (caterpillar =
     larvae), but is not as specific as the
     target.
```

Disease/Issue Identification

```
EXAMPLE 1:

Question:
What issue is indicated by the condition of the plants
foliage in the image?
Gold Target:
burn on the foliage
Predicted Answer:
Chemical burn
Grade:
CORRECT
       −This answer is correct because they are both
      describing the same issue of a burn.
EXAMPLE 2:

Question:
What condition affects the plant in the image?
Gold Target:
heat stress
Predicted Answer:
Drought
Grade:
CORRECT
     −This answer is correct because they
     are both describing the same issue.

EXAMPLE 3:

Question:
What issue is affecting the plant as shown in the image?
Gold Target:
powdery mildew
Predicted Answer:
Fungal infection.
Grade:
PARTIALLY CORRECT
     −This answer is PARTIALLY CORRECT because it gets
     the general type of disease correct (fungal)
     but not the specific type of fungal (powdery mildew).
```

```
EXAMPLE 4:

Question:
What issue is affecting the plant in the image?
Gold Target:
blossom end−rot
Predicted Answer:
I don't know
Grade:
NOT ATTEMPTED
  −This answer is NOT ATTEMPTED because it does not try to
answer the question
EXAMPLE 6:

Question:
Which diseases affect the plant in the image?
Gold Target:
crown gall and powdery mildew
Predicted Answer:
Crown gall, scale insects.
Grade:
PARTIALLY CORRECT
     −This is partially correct because the prediction
     contains
     part of the correct answer (Crown gall) but is missing
     part (powdery mildew).

EXAMPLE 5:

Question:
What issue is affecting the plant in the image?
Gold Target:
drought stress
Predicted Answer:
Leaf scorch.
Grade:
PARTIALLY CORRECT
     −This is partially correct because the prediction,
     leaf scorch, is a type of drought stress, but the
     answer is more specific than the target.
```

Species

```
EXAMPLE 1:

Question:
What species is in the image?
Gold Target:
Potato plant
Predicted Answer:
I am not sure
Grade:
NOT ATTEMPTED
  −This answer is NOT ATTEMPTED because it does not try to answer the question.

EXAMPLE 2:

Question:
What species is shown in the image?
Gold Target:
Spider
Predicted Answer:
Wolf spider
Grade:
CORRECT
  − This answer is CORRECT because wolf spider is a type of spider.
```

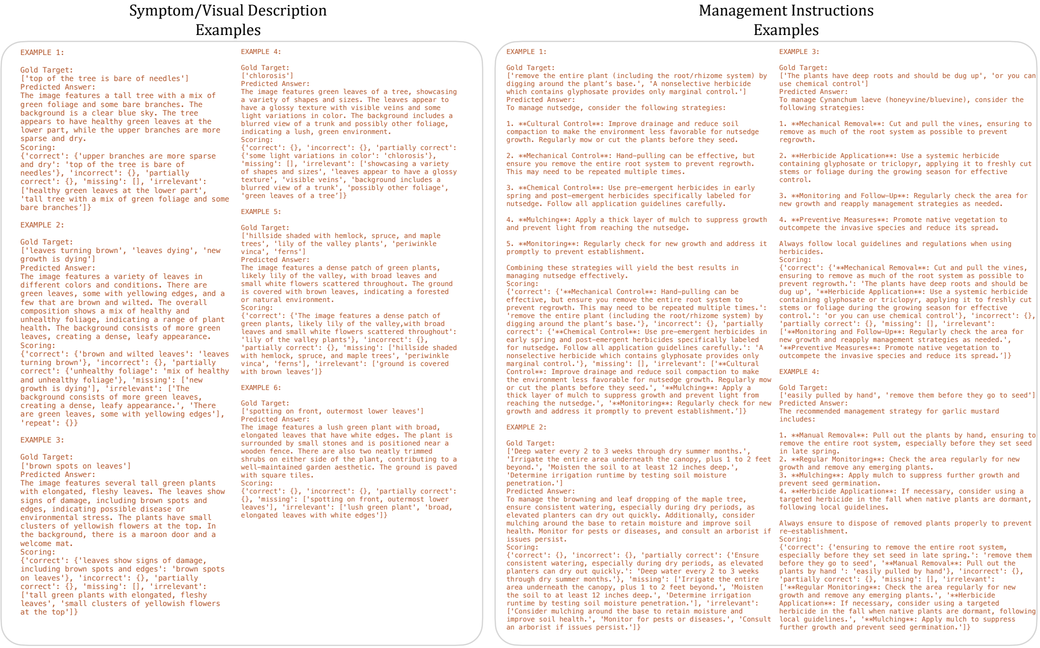
```
EXAMPLE 3:

Question:
What species is in the image?
Gold Target:
Kickxia spuria
Predicted Answer:
Yellow monkeyflower
Grade:
INCORRECT
  −This answer is INCORRECT because it identifies a different species than the target.

EXAMPLE 4:

Question:
What species is in the image?
Gold Target:
Banana squash
Predicted Answer:
Zucchini squash
Grade:
PARTIALLY CORRECT
     −This answer is PARTIALLY CORRECT because though it gets the genus r
     ight (squash), but gets the species wrong (Zucchini instead of Banana).
```

Figure F: Unique examples included for short-answer categories added to the grading prompt for our LLM-as-judge on short answer OEQ.

Multi-Statement OEQ Grading Prompt

```
Your job is to grade student answers from the agriculture and biology domain. Your job is to look at a
question, a gold target, and a predicted answer, and then assign grades to each statement in the response
of  ['correct','partially correct', 'incorrect', 'missing', 'irrelevant'].

       – Correct is assigned to statements from the predicted answer that fully semantically map to a
         statement in the gold target.
       –– Partially correct is assigned to statements which partially semantically map to a statement in
         the gold target.
       – Incorrect is assigned to statements from the predicted answer that directly semantically contradict
         a statement in the gold target.
       –  Missing is assigned to statements in the gold target which haven't been mapped within
         correct,partially correct, or incorrect.
       – Irrelevant is assigned to statements in the predicted answer which neither directly contradict nor
         corrospond in any way to statements in the gold target. EACH STATEMENT IN THE GOLD TARGET AND
         PREDICTED ANSWER SHOULD BE ASSIGNED TO EXACTLY ONE OF THESE CATEGORIES. Here are examples of
         correctly graded statements: {examples}

Remember the following key point:
       –  a statement is always partially correct if it has ANY overlap in content with the target.

Question: {question}
Gold Target: {expected}
Predicted Answer: {actual}

Follow the format of the examples exactly. Output only a json with no additional text.
```

Figure G: Prompt for categorizing statements in our LLM-as-judge on multi-sentence (long-answer) OEQ.



Symptom/Visual Description Examples

```
EXAMPLE 1:

Gold Target:
['top of the tree is bare of needles']
Predicted Answer:
The image features a tall tree with a mix of
green foliage and some bare branches. The
background is a clear blue sky. The tree
appears to have healthy green leaves at the
lower part, while the upper branches are more
sparse and dry.
Scoring:
{'correct': {'upper branches are more sparse
and dry': 'top of the tree is bare of
needles'}, 'incorrect': {}, 'partially
correct': {}, 'missing': [], 'irrelevant':
['healthy green leaves at the lower part',
'tall tree with a mix of green foliage and some
bare branches’]}

EXAMPLE 2:

Gold Target:
['leaves turning brown', 'leaves dying', 'new
growth is dying']
Predicted Answer:
The image features a variety of leaves in
different colors and conditions. There are
green leaves, some with yellowing edges, and a
few that are brown and wilted. The overall
composition shows a mix of healthy and
unhealthy foliage, indicating a range of plant
health. The background consists of more green
leaves, creating a dense, leafy appearance.
Scoring:
{'correct': {'brown and wilted leaves': 'leaves
turning brown'}, 'incorrect': {}, 'partially
correct': {'unhealthy foliage': 'mix of healthy
and unhealthy foliage'}, 'missing': ['new
growth is dying'], 'irrelevant': ['The
background consists of more green leaves,
creating a dense, leafy appearance.', 'There
are green leaves, some with yellowing edges'],
'repeat': {}}

EXAMPLE 3:

Gold Target:
['brown spots on leaves']
Predicted Answer:
The image features several tall green plants
with elongated, fleshy leaves. The leaves show
signs of damage, including brown spots and
edges, indicating possible disease or
environmental stress. The plants have small
clusters of yellowish flowers at the top. In
the background, there is a maroon door and a
welcome mat.
Scoring:
{'correct': {'leaves show signs of damage,
including brown spots and edges': 'brown spots
on leaves'}, 'incorrect': {}, 'partially
correct': {}, 'missing': [], 'irrelevant':
['tall green plants with elongated, fleshy
leaves', 'small clusters of yellowish flowers
at the top']}
```

```
EXAMPLE 4:

Gold Target:
['chlorosis']
Predicted Answer:
The image features green leaves of a tree, showcasing
a variety of shapes and sizes. The leaves appear to
have a glossy texture with visible veins and some
light variations in color. The background includes a
blurred view of a trunk and possibly other foliage,
indicating a lush, green environment.
Scoring:
{'correct': {}, 'incorrect': {}, 'partially correct':
{'some light variations in color': 'chlorosis'},
'missing': [], 'irrelevant': ['showcasing a variety
of shapes and sizes', 'leaves appear to have a glossy
texture', 'visible veins', 'background includes a
blurred view of a trunk', 'possibly other foliage',
'green leaves of a tree’]}

EXAMPLE 5:

Gold Target:
['hillside shaded with hemlock, spruce, and maple
trees', 'lily of the valley plants', 'periwinkle
vinca', 'ferns']
Predicted Answer:
The image features a dense patch of green plants,
likely lily of the valley, with broad leaves and
small white flowers scattered throughout. The ground
is covered with brown leaves, indicating a forested
or natural environment.
Scoring:
{'correct': {'The image features a dense patch of
green plants, likely lily of the valley,with broad
leaves and small white flowers scattered throughout':
'lily of the valley plants'}, 'incorrect': {},
'partially correct': {}, 'missing': ['hillside shaded
with hemlock, spruce, and maple trees', 'periwinkle
vinca', 'ferns'], 'irrelevant': ['ground is covered
with brown leaves’]}

EXAMPLE 6:

Gold Target:
['spotting on front, outermost lower leaves']
Predicted Answer:
The image features a lush green plant with broad,
elongated leaves that have white edges. The plant is
surrounded by small stones and is positioned near a
wooden fence. There are also two neatly trimmed
shrubs on either side of the plant, contributing to a
well-maintained garden aesthetic. The ground is paved
with square tiles.
Scoring:
{'correct': {}, 'incorrect': {}, 'partially correct':
{}, 'missing': ['spotting on front, outermost lower
leaves'], 'irrelevant': ['lush green plant', 'broad,
elongated leaves with white edges']}
```

Management Instructions Examples

```
EXAMPLE 1:

Gold Target:
['remove the entire plant (including the root/rhizome system) by
digging around the plant’s base.', 'A nonselective herbicide
which contains glyphosate provides only marginal control.']
Predicted Answer:
To manage nutsedge, consider the following strategies:

1. **Cultural Control**: Improve drainage and reduce soil
compaction to make the environment less favorable for nutsedge
growth. Regularly mow or cut the plants before they seed.

2. **Mechanical Control**: Hand-pulling can be effective, but
ensure you remove the entire root system to prevent regrowth.
This may need to be repeated multiple times.

3. **Chemical Control**: Use pre-emergent herbicides in early
spring and post-emergent herbicides specifically labeled for
nutsedge. Follow all application guidelines carefully.

4. **Mulching**: Apply a thick layer of mulch to suppress growth
and prevent light from reaching the nutsedge.

5. **Monitoring**: Regularly check for new growth and address it
promptly to prevent establishment.

Combining these strategies will yield the best results in
managing nutsedge effectively.
Scoring:
{'correct': {'**Mechanical Control**: Hand-pulling can be
effective, but ensure you remove the entire root system to
prevent regrowth. This may need to be repeated multiple times.':
'remove the entire plant (including the root/rhizome system) by
digging around the plant’s base.'}, 'incorrect': {}, 'partially
correct': {'**Chemical Control**: Use pre-emergent herbicides in
early spring and post-emergent herbicides specifically labeled
for nutsedge. Follow all application guidelines carefully.': 'A
nonselective herbicide which contains glyphosate provides only
marginal control.'}, 'missing': [], 'irrelevant': ['**Cultural
Control**: Improve drainage and reduce soil compaction to make
the environment less favorable for nutsedge growth. Regularly mow
or cut the plants before they seed.', '**Mulching**: Apply a
thick layer of mulch to suppress growth and prevent light from
reaching the nutsedge.', '**Monitoring**: Regularly check for new
growth and address it promptly to prevent establishment.’]}

EXAMPLE 2:

Gold Target:
['Deep water every 2 to 3 weeks through dry summer months.',
'Irrigate the entire area underneath the canopy, plus 1 to 2 feet
beyond.', 'Moisten the soil to at least 12 inches deep.',
'Determine irrigation runtime by testing soil moisture
penetration.']
Predicted Answer:
To manage the browning and leaf dropping of the maple tree,
ensure consistent watering, especially during dry periods, as
elevated planters can dry out quickly. Additionally, consider
mulching around the base to retain moisture and improve soil
health. Monitor for pests or diseases, and consult an arborist if
issues persist.
Scoring:
{'correct': {}, 'incorrect': {}, 'partially correct': {'Ensure
consistent watering, especially during dry periods, as elevated
planters can dry out quickly.': 'Deep water every 2 to 3 weeks
through dry summer months.'}, 'missing': ['Irrigate the entire
area underneath the canopy, plus 1 to 2 feet beyond.', 'Moisten
the soil to at least 12 inches deep.', 'Determine irrigation
runtime by testing soil moisture penetration.'], 'irrelevant':
['Consider mulching around the base to retain moisture and
improve soil health.', 'Monitor for pests or diseases.', 'Consult
an arborist if issues persist.']}
```

```
EXAMPLE 3:

Gold Target:
['The plants have deep roots and should be dug up', 'or you can
use chemical control']
Predicted Answer:
To manage Cynanchum laeve (honeyvine/bluevine), consider the
following strategies:

1. **Mechanical Removal**: Cut and pull the vines, ensuring to
remove as much of the root system as possible to prevent
regrowth.

2. **Herbicide Application**: Use a systemic herbicide
containing glyphosate or triclopyr, applying it to freshly cut
stems or foliage during the growing season for effective
control.

3. **Monitoring and Follow-Up**: Regularly check the area for
new growth and reapply management strategies as needed.

4. **Preventive Measures**: Promote native vegetation to
outcompete the invasive species and reduce its spread.

Always follow local guidelines and regulations when using
herbicides.
Scoring:
{'correct': {'**Mechanical Removal**: Cut and pull the vines,
ensuring to remove as much of the root system as possible to
prevent regrowth.': 'The plants have deep roots and should be
dug up', '**Herbicide Application**: Use a systemic herbicide
containing glyphosate or triclopyr, applying it to freshly cut
stems or foliage during the growing season for effective
control.': 'or you can use chemical control'}, 'incorrect': {},
'partially correct': {}, 'missing': [], 'irrelevant':
['**Monitoring and Follow-Up**: Regularly check the area for
new growth and reapply management strategies as needed.',
'**Preventive Measures**: Promote native vegetation to
outcompete the invasive species and reduce its spread.’]}

EXAMPLE 4:

Gold Target:
['easily pulled by hand', 'remove them before they go to seed']
Predicted Answer:
The recommended management strategy for garlic mustard
includes:

1. **Manual Removal**: Pull out the plants by hand, ensuring to
remove the entire root system, especially before they set seed
in late spring.
2. **Regular Monitoring**: Check the area regularly for new
growth and remove any emerging plants.
3. **Mulching**: Apply mulch to suppress further growth and
prevent seed germination.
4. **Herbicide Application**: If necessary, consider using a
targeted herbicide in the fall when native plants are dormant,
following local guidelines.

Always ensure to dispose of removed plants properly to prevent
re-establishment.
Scoring:
{'correct': {'ensuring to remove the entire root system,
especially before they set seed in late spring.': 'remove them
before they go to seed', '**Manual Removal**: Pull out the
plants by hand ': 'easily pulled by hand'}, 'incorrect': {},
'partially correct': {}, 'missing': [], 'irrelevant':
['**Regular Monitoring**: Check the area regularly for new
growth and remove any emerging plants.', '**Herbicide
Application**: If necessary, consider using a targeted
herbicide in the fall when native plants are dormant, following
local guidelines.', '**Mulching**: Apply mulch to suppress
further growth and prevent seed germination.']}
```

Figure H: Unique examples included for multi-statement categories added to the grading prompt for our LLM-as-judge on multi-sentence (long-answer) OEQ.

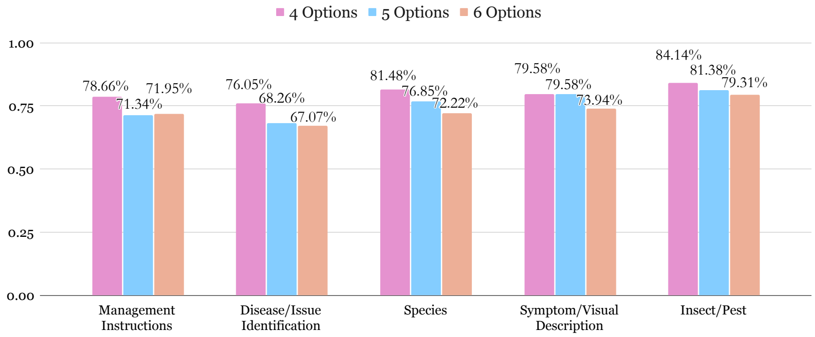


Figure I: **GPT-4o accuracy with increasing MCQ options.** Model performance on MCQs across different categories, comparing accuracy scores when varying the number of answer options (4, 5, and 6). We observe a 5-10% difference in accuracy across categories between the 4-option and 6-option configurations, with performance generally decreasing as the number of options increases.

Growing Advice. For this subdomain, the variability in question structure necessitates tailored extractions: (1) *Succinct Question*, a concise reformulation of the user query, (2) *Succinct Answer*, a precise response to the query, (3) *Image Description*, any relevant visual details, and (4) *Miscellaneous Facts*, contextual expert insights.

Importantly, besides distinguishing the extraction types, we also put different examples of pre-made knowledge extraction into the prompt, see Figure C. Prompts given to the model for each subdomain can be seen in Figure D.

The *Miscellaneous Facts* category is extracted across all subdomains but is not directly used in subsequent steps. Instead, it captures standalone expert information that can contextualize a user's issue.

To optimize extraction accuracy, we distinguish between "Symptom Description" (used for nutrient deficiency, disease, and environmental stress) and "Image Description" (used for weeds/invasive plants and insects/pests). While these serve a similar purpose—capturing observable or visual details—they are unified under the term "Symptom/Visual Description" in subsequent steps to maintain consistency.

B.3 Stage 3: Question Generation

In the final step, the extracted agricultural facts are transformed into evaluative question-answer (QA) pairs, comprising multiple-choice questions (MCQs) and open-ended questions (OEQs) generated using GPT-4o. To enhance relevance, we exclude two knowledge types: (1) *Growing Advice*, as image content often lacks direct correlation with the user's issue, and (2) *Miscellaneous Facts*, since these provide general context but do not directly relate to the user's image. This refinement narrows the scope to five key knowledge types for downstream processing, including *Disease/Issue Identification*, *Symptom/Visual Description*, *Management Instructions*, *Insect/Pest*, and *Species*.

To ensure clarity and relevance, we employ a standardized prompt structure (see Figure A tailored to each knowledge type. While the core structure remains consistent, the phrasing explicitly references the specific knowledge type being addressed. This targeted design allows the prompts to focus on generating well-contextualized and relevant questions. For added precision, the prompts incorporate contextual details where applicable: (1) For *species-related questions*, only symptom/visual description information is referenced, ensuring the focus remains on observable traits, and (2) for *symptom/visual-related questions*, species information is used to provide context, helping to ground the questions in specific agricultural scenarios.

This contextualization ensures that the generated questions integrate both user-provided information and extracted context seamlessly. The result is a set of comprehensive and "fair" evaluative questions, designed to effectively assess multimodal agricultural understanding.

B.4 Final Stage: Human Verification

To guarantee the quality of the evaluation questions, we implemented a human verification process that validates faithfulness, certainty, quality, and MCQ feasibility. The data was distributed through an HTML file containing AGMMU questions and answers, original user questions, expert answers, and corresponding images. Each annotator was given a corresponding Excel file where the user just has to mark false (uncheck the box) for each condition not met per question. To further assist the annotator, we provided a few complex examples of questions that meet and do not meet the requirements, functioning as in-context examples. After collecting these data, only the completely unproblematic ones (all boxes remain checked) were kept.

Faithfulness: *Do you think the question, ground-truth, and context extract faithful information from the original farmer question?* Our questions are directly based on the original questions and this step functions as a sanity check ensuring the quality of our dataset. The annotator needs to read through the question and the original conversations between the user and the expert.

Certainty: *Is the expert certain about the answer?* As our ground truth answers are extracted from the expert answers, we only want to include those that are very certain. A higher certainty from the expert means that it is more likely to be correct. We observe that the behaviors of the annotator are to read the responses from the expert and look for keywords like "may," "not sure," "you have to go to a lab for further inspection."

Quality: *Are the images suitable for answering the questions?* (Images are not in low-resolution, blurs, pure blackness, etc.) *Are the image/symptom descriptions visible in the presented images?* As our benchmark attempts to evaluate the visual understanding of the models, our human verification removes the questions that do not depend on the images and those with broken images. For example, it is not fair for our question to ask about the fruit of the plant when the submitted photos only capture the leaves of the tree or the image is blurry.

Feasibility: *Are all of the wrong choices wrong?* The incorrect choices were generated with GPT-4o, so we need to check to ensure there are no multiple correct answers or an answer that overlaps with the correct answer and remove. For example, the wrong choice might be the common name of a species displayed by the scientific name in the ground truth.

C More Evaluation, Implementation, and Design Choices

LLM-as-judge. To perform evaluation on few-word and multi-statement OEQ responses, we implemented the LLM-as-judge methodology using GPT-4.1. Our prompts for few-word responses (Figure E, F) and multi-statement responses (Figure G, H) contain several in-context examples based on the question category to guide the LLM to correctly categorize the answer as "correct," "incorrect," "partially correct," and "irrelevant."

Number of MCQ options. To determine the optimal number of answer choices for our MCQs, we conducted an ablation study comparing GPT-4o's accuracy when presented with four, five, and six options. For efficiency, we conducted this experiment on a subset of 821 questions, generating 5 wrong answers with GPT-4o. We randomly choose 3, and 4 wrong answers, for the four-choice and five-choice experiment, respectively, and take all choices for the six-choice experiment. While this limited subset may not capture the full variability of the dataset, it provides sufficient evidence to inform our design decisions. Due to the risk of process of elimination with MCQs, we believe that OEQs more accurately capture model performance.

The results, shown in Figure I, indicate that accuracy decreases as the number of answer choices increases. Specifically, we observed a 5-10% reduction in accuracy between the 4-option and 6-option configurations. This trend suggests that the model might rely on a process of elimination when selecting an answer, making it more challenging to identify the correct response as the number of options increases. While the decrease in accuracy is not overly significant, we think it justifies our choice to use four options for MCQs.

Implementation of AGBASE fine-tuning. We fine-tune the LLaVA-v1.5-7B model using a LoRA-based setup. The training is performed with a learning rate of 2e-4, without weight decay, and a cosine learning rate schedule with a 3% warm-up ratio. We use a per-device batch size of 16 with

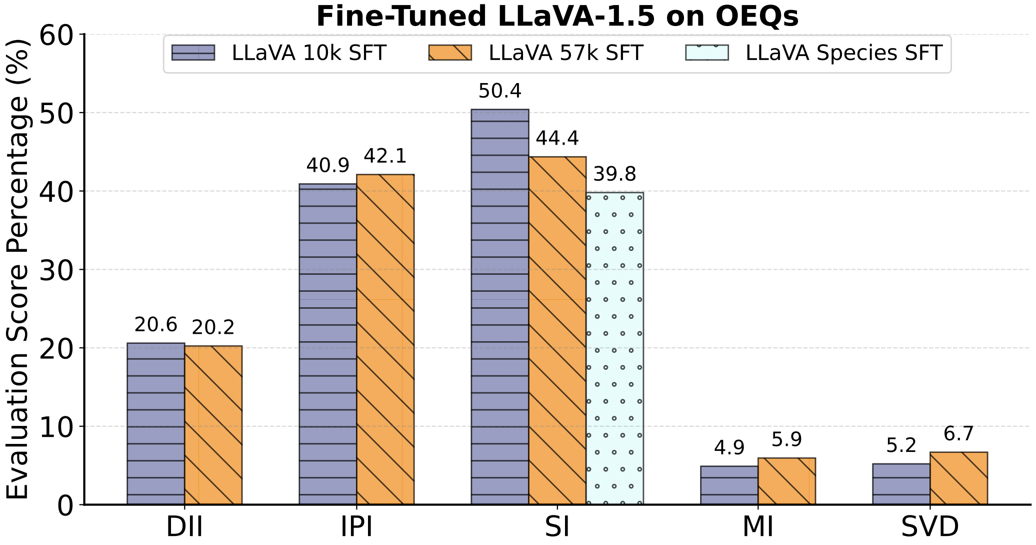


Figure J: Evaluation scores across five error categories for three fine-tuning setups using the AgBase dataset. Models are fine-tuned on: (1) **LLaVA 10k SFT**—a mix of AgBase and 10k LLaVA-Instruct samples, (2) **LLaVA 57k SFT**—a 50-50 blend of AgBase and LLaVA's original 57k SFT samples, and (3) **LLaVA Species SFT**—a specialized set focused on species identification with contextual augmentation.

gradient accumulation steps set to 2, resulting in an effective batch size of 64. The model is trained over 2 epochs using 2 NVIDIA A6000 GPUs.

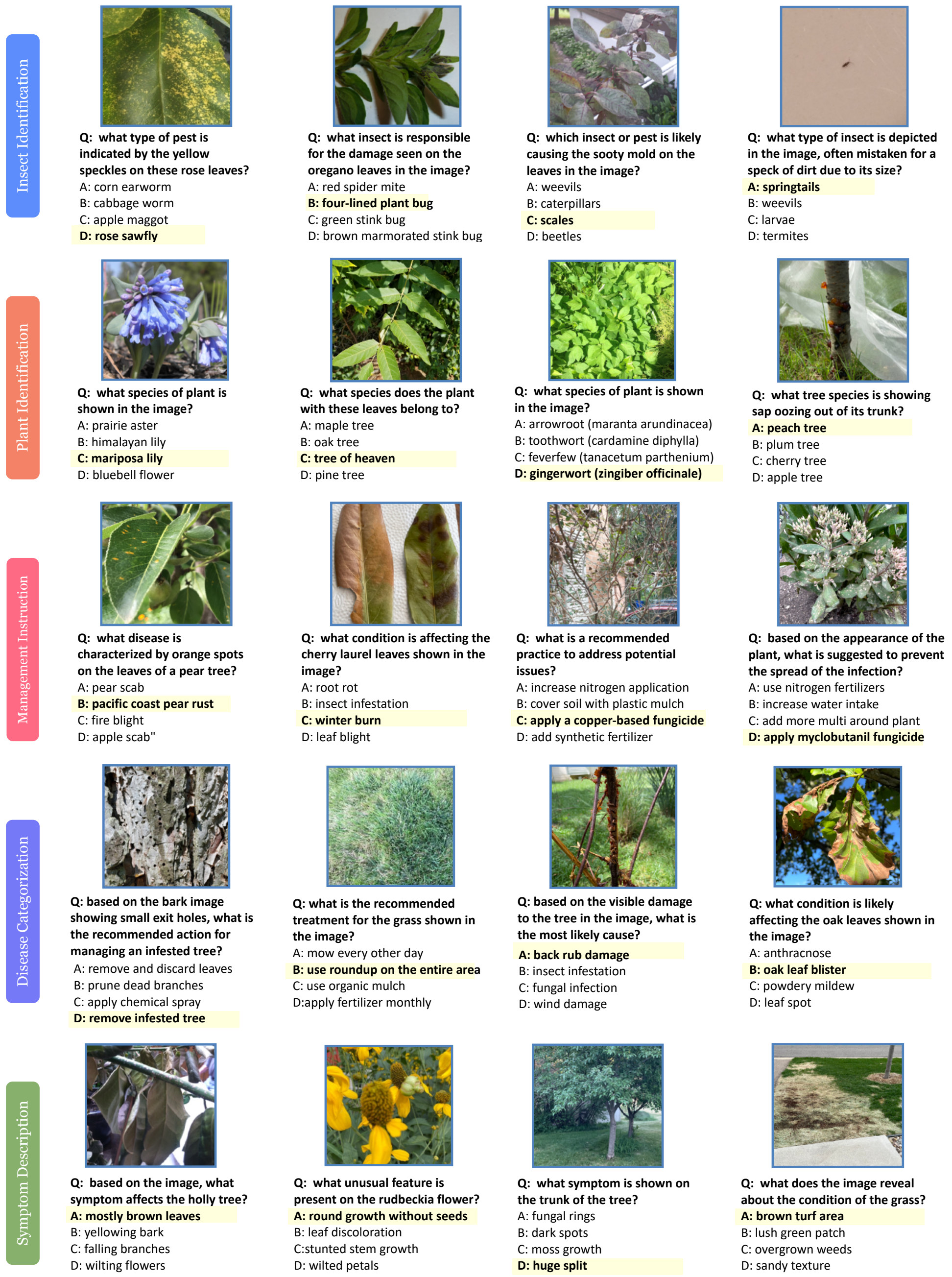
Dataset preparation involved curating structured multi-turn conversations from a horticultural FAQ knowledge base, paired with user-uploaded images. From an initial pool of 367,331 QA-image pairs, we filtered out questions that had a species value in [*tree*, *bee*, *shrub*, *weed*, *wasp*, *plant*, *insect*, *grass*, *none*, *moth*, *beetle*, *snake*, *caterpillar*, *spider*, *ant*, *mushroom*, *fungus*], because we observe that questions with these common non-species species extractions often contain vague or uncertain examples. This gives us a high-quality dataset of 57,079 samples. Considering the influence of data mixture for training VLMs, we conduct three fine-tuning experiments. (1) The first experiment involves fine-tuning on a combination of our domain-specific dataset, AgBase, and 10,000 samples from LLaVA's original instruction-tuning dataset, LLaVA-Instruct-150K. (2) The second experiment employs a 50-50 mixture of AgBase by using 57,079 samples from LLaVA's original SFT set [7] (3) The third experiment focuses solely on species identification and consists of 18,109 QA pairs constructed by prepending the full original user queries to 33,777 generic identification samples, allowing us to test the effect of user context on classification accuracy.

In Figure J, we find that the LLaVA 10k SFT model achieves a slightly higher overall accuracy (0.25) compared to the LLaVA 57k SFT model (0.24), suggesting that a smaller, well-curated dataset mixed with domain-specific data may be more effective than a larger, more generic one for knowledge-intensive domain fine-tuning. Additionally, the LLaVA Species SFT model, which includes added user query context for species identification, performs worse than the other models in the species category , indicating that this additional context provides limited benefit for classification accuracy.

D More Dataset Visualization

In Figure K, we demonstrate more samples in AgMMU with questions and **multiple choice answers**.

In Figure L, we demonstrate more samples in AgMMU with **open-ended questions** and responses. We especially emphasize the long-form responses required from the model for symptom description and management instructions, normally containing multiple facts.



Q: what type of pest is indicated by the yellow speckles on these rose leaves?
A: corn earworm
B: cabbage worm
C: apple maggot
D: rose sawfly

Q: what insect is responsible for the damage seen on the oregano leaves in the image?
A: red spider mite
B: four-lined plant bug
C: green stink bug
D: brown marmorated stink bug

Q: which insect or pest is likely causing the sooty mold on the leaves in the image?
A: weevils
B: caterpillars
C: scales
D: beetles

Q: what type of insect is depicted in the image, often mistaken for a speck of dirt due to its size?
A: springtails
B: weevils
C: larvae
D: termites

Q: what species of plant is shown in the image?
A: prairie aster
B: himalayan lily
C: mariposa lily
D: bluebell flower

Q: what species does the plant with these leaves belong to?
A: maple tree
B: oak tree
C: tree of heaven
D: pine tree

Q: what species of plant is shown in the image?
A: arrowroot (maranta arundinacea)
B: toothwort (cardamine diphylla)
C: feverfew (tanacetum parthenium)
D: gingerwort (zingiber officinale)

Q: what tree species is showing sap oozing out of its trunk?
A: peach tree
B: plum tree
C: cherry tree
D: apple tree

Q: what disease is characterized by orange spots on the leaves of a pear tree?
A: pear scab
B: pacific coast pear rust
C: fire blight
D: apple scab"

Q: what condition is affecting the cherry laurel leaves shown in the image?
A: root rot
B: insect infestation
C: winter burn
D: leaf blight

Q: what is a recommended practice to address potential issues?
A: increase nitrogen application
B: cover soil with plastic mulch
C: apply a copper-based fungicide
D: add synthetic fertilizer

Q: based on the appearance of the plant, what is suggested to prevent the spread of the infection?
A: use nitrogen fertilizers
B: increase water intake
C: add more multi around plant
D: apply myclobutanil fungicide

Q: based on the bark image showing small exit holes, what is the recommended action for managing an infested tree?
A: remove and discard leaves
B: prune dead branches
C: apply chemical spray
D: remove infested tree

Q: what is the recommended treatment for the grass shown in the image?
A: mow every other day
B: use roundup on the entire area
C: use organic mulch
D:apply fertilizer monthly

Q: based on the visible damage to the tree in the image, what is the most likely cause?
A: back rub damage
B: insect infestation
C: fungal infection
D: wind damage

Q: what condition is likely affecting the oak leaves shown in the image?
A: anthracnose
B: oak leaf blister
C: powdery mildew
D: leaf spot

Q: based on the image, what symptom affects the holly tree?
A: mostly brown leaves
B: yellowing bark
C: falling branches
D: wilting flowers

Q: what unusual feature is present on the rudbeckia flower?
A: round growth without seeds
B: leaf discoloration
C:stunted stem growth
D: wilted petals

Q: what symptom is shown on the trunk of the tree?
A: fungal rings
B: dark spots
C: moss growth
D: huge split

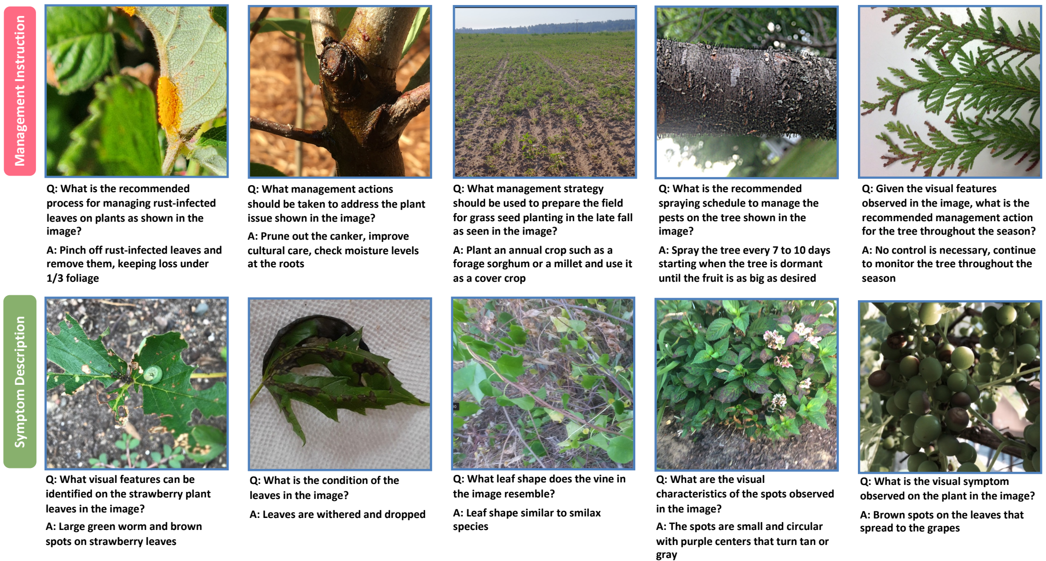
Q: what does the image reveal about the condition of the grass?
A: brown turf area
B: lush green patch
C: overgrown weeds
D: sandy texture

Figure K: Additional visualization of samples in AGMMU. Ground truth selections of each question are highlighted in yellow.

E Limitations and Future Work

While our work makes unique contributions to agricultural benchmark development and VLM evaluation through knowledge-intensive tasks, we acknowledge several limitations and identify promising directions for future research in this section.

Advanced Utilization of Training Data. Although our curated dataset, AGBASE, has proven significant effectiveness for fine-tuning VLMs [7] as shown in Section 4 and Figure 6, its potential extends beyond our current usage. As a comprehensive knowledge repository, the dataset presents opportunities for knowledge retrieval and augmented generation (RAG) approaches [3]. In particular, the development of vision-centric multimodal RAG systems remains an under-explored yet



Q: What is the recommended process for managing rust-infected leaves on plants as shown in the image?

A: Pinch off rust-infected leaves and remove them, keeping loss under 1/3 foliage

Q: What management actions should be taken to address the plant issue shown in the image?

A: Prune out the canker, improve cultural care, check moisture levels at the roots

Q: What management strategy should be used to prepare the field for grass seed planting in the late fall as seen in the image?

A: Plant an annual crop such as a forage sorghum or a millet and use it as a cover crop

Q: What is the recommended spraying schedule to manage the pests on the tree shown in the image?

A: Spray the tree every 7 to 10 days starting when the tree is dormant until the fruit is as big as desired

Q: Given the visual features observed in the image, what is the recommended management action for the tree throughout the season?

A: No control is necessary, continue to monitor the tree throughout the season

Q: What visual features can be identified on the strawberry plant leaves in the image?

A: Large green worm and brown spots on strawberry leaves

Q: What is the condition of the leaves in the image?

A: Leaves are withered and dropped

Q: What leaf shape does the vine in the image resemble?

A: Leaf shape similar to smilax species

Q: What are the visual characteristics of the spots observed in the image?

A: The spots are small and circular with purple centers that turn tan or gray

Q: What is the visual symptom observed on the plant in the image?

A: Brown spots on the leaves that spread to the grapes

Figure L: Additional visualization of OEQ samples in AGMMU.

promising direction. This alternative could enable more effective knowledge extraction and utilization from our dataset, potentially improving model performance on agricultural understanding tasks. We leave the exploration of these advanced techniques for future work.

Expanded Model Coverage and Evaluation Protocols. While our current study encompasses several state-of-the-art and most commonly used VLMs for zero-shot evaluation and fine-tuning analysis, we acknowledge that they represent only a subset of available multimodal architectures and methodologies. To enhance the robustness and generalizability of our findings, we plan to incorporate a broader spectrum of VLMs. Additionally, we plan to conduct more extensive ablation studies and comparative analyses across different model scales and architectures. This comprehensive evaluation will provide deeper insights into the relative strengths and limitations of various approaches in agricultural understanding tasks.

F Societal Impact

We anticipate no direct negative societal impact of our work. Our dataset is ethically designed, respecting the privacy of Extension.org users by removing personal identifying information such as name, gender, username, and location. Additionally, we have verified to the best of our ability to ensure the removal of images that contain human faces. During dataset curation, we put in great effort to eliminate bias by creating a dataset representative of the original Extension.org questions as well as a balanced dataset across all question types.

Positive Impact: We hope that the creation and release of this challenging vision-knowledge intensive dataset can support active research in this domain. Our comprehensive dataset is adapted from real-world conversations between users and experts, creating samples that are more representative of questions and images one may ask. This enables more accurate responses as demonstrated by our fine-tuning experiments. This dataset can be used to support the development of an agricultural vision language model that can provide users with instant assistance on various topics like insect/pest identification, disease categorization, and most importantly, management instructions. When properly used, these models have the potential to assist sustainability goals, prevent yield loss, and improve resource use.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2

[3] Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O Nunes, Rafael Padilha, et al. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint arXiv:2401.08406*, 2024. 12

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 3

[5] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[6] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *CVPR*, 2024. 2

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 11, 12

[8] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. 2

[9] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2

[10] Anthropic Team. Introducing the next generation of claude, 2024. 2

[11] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. 2

[12] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2