

---

# CABBAGE: Comprehensive Agricultural Benchmark Backed by AI-Guided Evaluation

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Artificial intelligence is transforming agriculture, yet the evaluation of Multimodal  
2 Large Language Models (MLLMs) in this domain remains underexplored. We  
3 introduce CABBAGE, the first domain-specific multimodal benchmark designed  
4 to assess MLLMs on real-world agronomic tasks. CABBAGE comprises 74K  
5 questions spanning three key areas: expert agronomic knowledge (25K ques-  
6 tions on plant pathology, soil science, irrigation, etc.), pest and crop identification  
7 (1.2K plant species, 1.2K pest species, 47K images), and agronomic procedural  
8 reasoning (2K task-based scenarios). We employ an LLM-guided scoring frame-  
9 work based on G-Eval to ensure robust and explainable evaluation. We benchmark  
10 state-of-the-art MLLMs—including GPT-4o and Gemini 2.0 Flash—in a zero-shot  
11 setting, revealing their strengths and limitations. While these models demonstrate  
12 strong agricultural scientific knowledge, outperforming open-source alternatives  
13 (e.g., Qwen2-VL-72B by 4% score gap), they struggle with plant pathology and  
14 agronomic visual reasoning. However, their accuracy in visual identification of  
15 taxonomy improves by 20% when observing multiple plant organs. By open-  
16 sourcing CABBAGE<sup>1</sup>, we aim to advance agricultural AI, fostering more reliable,  
17 interpretable models for precision farming, climate resilience, and food security.

## 18 1 Introduction

19 Artificial intelligence is transforming the agricultural sector, enabling innovative solutions for pre-  
20 cision farming, soil health optimization, crop forecasting, pest management, and climate change  
21 adaptation [23, 28, 46, 53]. The emergence of Multimodal Large Language Models (MLLMs) has  
22 further propelled these advancements, offering promising capabilities to bridge the gap between  
23 visual and textual agricultural data [30, 5, 58, 49]. These models could act as AI-powered agricul-  
24 tural assistants, providing expert guidance on addressing critical challenges such as crop disease  
25 management, pest control, and climate adaptation, directly contributing to global food security and  
26 sustainable farming practices. LLMs have succeeded across various domains. Nonetheless, their  
27 application in agriculture remains in its early stages [62, 7]. While recent work has addressed specific  
28 agricultural tasks such as crop disease diagnosis and pest detection [32, 61, 54], a systematic and  
29 standardized evaluation framework designed to unify and benchmark the diverse challenges of the  
30 agricultural domain remains a critical unmet need [16, 27, 59].

31 Benchmarks play a crucial role in assessing AI capabilities, ensuring a consistent and objective  
32 method for comparing models across different domains [17]. Existing MLLM benchmarks primarily  
33 focus on general-purpose vision-language understanding, such as captioning [40], visual question  
34 answering (VQA) [30], and perceptual reasoning [30]. However, agricultural applications demand  
35 more specialized evaluation criteria, encompassing a wide knowledge of plant phytology, regional

---

<sup>1</sup>The benchmark code and datasets can be accessed at <https://github.com/boilnserve/cabbage> (code) and <https://huggingface.co/datasets/boilnserve/cabbage> (datasets).

## CABBAGE — Agronomic Visual Cognition

**Q:** This is the poster boy of exceptions to the rules of plant pathology. This is peach leaf curl caused by *Taphrina deformans*, a fungal pathogen, but what are the symptoms more typical of?

- A. Bacterial infection
- B. Nematode attack
- C. Mite damage
- D. Don't know

**Correct:** C — Mites can cause gross deformities within a leaf such as that seen here. None of the other problems would ever cause such symptoms.



<b>GPT-4o:</b>	✓ C — Peach leaf curl results in thickened, curled, and discolored leaves, which are more typically associated with mite damage.
<b>GPT-4o-mini:</b>	✗ A — Peach leaf curl caused by <i>Taphrina deformans</i> typically presents symptoms similar to those caused by a bacterial infection.
<b>Gemini 2 F:</b>	✗ A — The disease is fungal, but its symptoms are more in line with bacterial infections.
<b>Gemini 2 FL:</b>	✗ A — Although caused by a fungus, the symptoms resemble those of a bacterial infection.
<b>Qwen-2-VL-72B:</b>	✗ A — The symptoms appear more characteristic of fungal infections rather than mite damage.
<b>Qwen-2-VL-7B:</b>	✗ A — The image suggests a viral infection.

Figure 1: **MLLMs reasoning flaws.** Qualitative example illustrating systematic reasoning failures in a task requiring integration of agronomic visual cues, agricultural scientific knowledge, and causal inference. Most models misattribute the symptoms of peach leaf curl—caused by the fungal pathogen *Taphrina deformans*—to incorrect etiologies such as bacterial or viral infections. GPT-4o identifies the correct cause, highlighting superior multimodal reasoning. Model answer rationales are shown.

36 agronomic practices, and complex decision-making under uncertainty. In response to this critical gap,  
37 we introduce CABBAGE (Comprehensive Agricultural Benchmark Backed by AI-Guided Evaluation),  
38 a pioneering and much-needed benchmark designed to evaluate MLLMs on extensive, domain-specific  
39 agricultural tasks. By incorporating open-ended and multiple-choice questions, multimodal reasoning,  
40 and procedural understanding tasks, CABBAGE provides a rigorous framework that ensures AI  
41 models are equipped to handle the complexity and nuances of real-world agronomic intelligence.

42 Our evaluations demonstrate that state-of-the-art MLLMs struggle with tasks requiring simultaneous  
43 integration of advanced agricultural knowledge and causal reasoning—a hallmark of expert decision-  
44 making in agronomy. As shown through Figure 1, diagnostic scenarios requiring cross-domain  
45 knowledge integration expose systematic flaws in the models’ reasoning processes.

46 To systematically characterize these limitations, we designed CABBAGE as a structured, diagnostic  
47 benchmark featuring task-driven categorization to assess models across three key areas (see Figure 2).  
48 These include: (a) *agricultural scientific knowledge*—25K questions evaluating understanding of  
49 core agricultural principles, including plant pathology, soil science, irrigation, and climate adaptation;  
50 (b) *agronomic visual cognition*—47K questions evaluating visual recognition across all growth stages  
51 for 1.2K plant species and 1.2K pest species (insects, fungi, bacteria, and viruses), enabling precise  
52 disease severity assessment and plant health diagnosis; and (c) *agricultural procedural reasoning*—  
53 2K questions assessing the models’ ability to plan, execute, and troubleshoot agronomic workflows  
54 (e.g., tree pruning, soil preparation) while predicting the outcomes of improper practices.

55 Existing agricultural benchmarks offer valuable but limited evaluations of MLLMs, often relying  
56 on coarse commodity groupings or small class sets that overlook species-level diversity essential  
57 for real-world diagnostics. As shown in Table 1, CABBAGE addresses this gap through taxonomic  
58 precision—evaluating plant, pest, and pathogen species—and incorporates procedural reasoning  
59 absent in prior work.

60 Given the complexity of evaluating open-ended responses in highly specialized agricultural domains,  
61 we introduce an LLM-guided evaluation framework within agricultural AI models based on G-Eval  
62 [33]—a framework for natural language generation (NLG) evaluation. Our approach integrates  
63 explicit Chain-of-Thought (CoT) reasoning [51], task-specific scoring rubrics [7], and structured  
64 judgment and scores extraction, enabling consistent and explainable evaluation of model-generated  
65 answers concerning *correctness*, *specificity*, *conciseness* and *procedural accuracy* and *flow* for  
66 procedural tasks. Unlike traditional NLP metrics such as BLEU and ROUGE [38, 12], which fail  
67 to capture the depth of reasoning and context in complex, open-ended tasks, G-Eval offers a more  
68 nuanced and interpretable approach that aligns better with human evaluation [33].

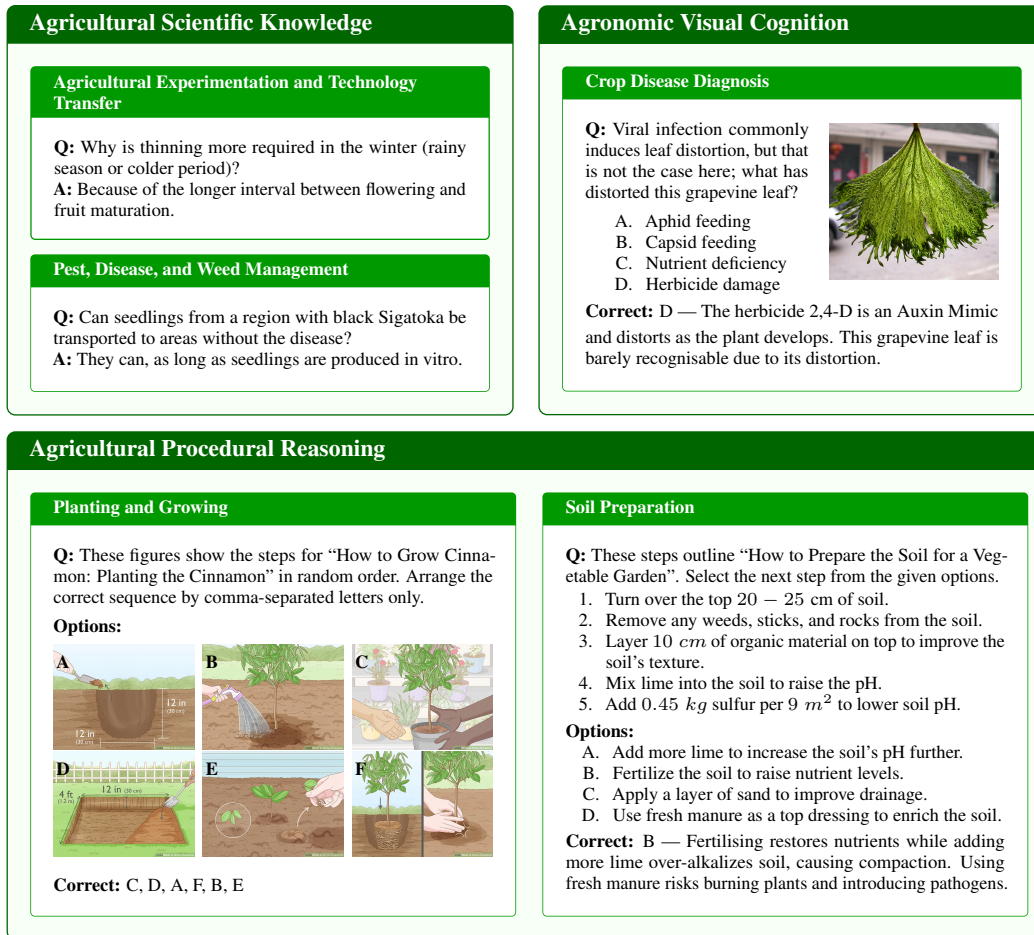


Figure 2: **Overview of the CABBAGE benchmark.** The figure shows some examples of Q&A pairs organized in a structured task categorisation across the three key areas.

69 To benchmark the performance of state-of-the-art MLLMs, we evaluate both commercial and open-  
70 source pre-trained vision-language models, including GPT-4o and GPT-4o-mini [39], Gemini 2.0  
71 Flash and Gemini 2.0 Flash-Lite [48], and Qwen2-VL-7B-Instruct and Qwen2-VL-72B-Instruct  
72 [50]. We simulate real-world deployment by evaluating in a zero-shot learning setting, assessing  
73 the models’ ability to generalize to unseen agronomic tasks without prior fine-tuning. Our findings  
74 provide insights into the strengths and limitations of these models.

75 CABBAGE marks an advancement in AI-driven agronomy, presenting the first comprehensive  
76 multimodal benchmark designed for this field. Through the public release of our dataset and  
77 evaluation framework, we seek to catalyze innovation within agricultural AI, empowering researchers  
78 and practitioners to create robust, domain-specific solutions that address critical challenges in global  
79 food security and sustainable farming practices.

80 In summary, our work makes the following key contributions:

- 81 • We introduce CABBAGE, a large-scale multimodal evaluation framework for AI-powered agricul-  
82 tural assistants, covering expert knowledge, pest and crop identification, and procedural reasoning.
- 83 • We propose an LLM-guided scoring framework for explainable and consistent evaluation of  
84 agricultural AI models’ responses to open-ended questions.
- 85 • We evaluate state-of-the-art MLLMs under a zero-shot setting, analyzing their strengths, limita-  
86 tions, and generalization capabilities in real-world agronomic intelligence.
- 87 • We release an open-source dataset and evaluation framework, providing a standardized benchmark  
88 to accelerate progress in the field.

## 89 2 Related Works

90 **MLLMs benchmarks.** As MLLMs advance [39, 48, 30, 5, 18, 9], specialized benchmarks have  
91 become crucial in evaluating their capabilities. Traditional benchmarks [29, 20, 21, 57, 25] mainly  
92 assess basic image-text integration, focusing on visual question answering and captioning.

93 LLaVA-Bench [30] tests models’ performance in multimodal instruction following and task-solving,  
94 while MME [16] evaluates perceptual and reasoning skills across diverse cognitive domains. SEED-  
95 Bench [27] focuses on generative comprehension, measuring spatial and temporal reasoning in images  
96 and videos. Complementing these efforts, MM-Vet [58] examines integrated multimodal reasoning  
97 by assessing models’ ability to combine core vision–language skills in complex tasks. Furthermore,  
98 additional benchmarks such as LAMM [55], Open-VQA [59], UniBench [3], and TouchStone [6]  
99 provide valuable perspectives on evaluating multimodality, expanding the range of available methods.

100 In parallel, domain-specific benchmarks have been introduced to evaluate MLLMs on specialized,  
101 fine-grained tasks. MMMU [60] covers 30 subjects across six disciplines, testing models’ ability  
102 to process domain-specific multimodal content. MMBench [34] examines expert-level reasoning  
103 under challenging conditions. ScienceQA [42] extends multimodal evaluation to scientific question  
104 answering (QA), requiring models to integrate textual and diagrammatic information. MMT-Bench  
105 [56] introduces a broad set of scenario-based tasks, spanning visual recognition, reasoning, and  
106 planning in real-world multimodal interactions.

107 **Agriculture-specific benchmarks.** Early initiatives, such as the PlantVillage dataset [22], PlantDoc  
108 [44], and CropDeep [65], laid the groundwork for applying computer vision techniques to plant  
109 pathology and precision farming. CLIP-based models like BioCLIP [46], BioTrove [53], and Agri-  
110 CLIP [37] introduce datasets of image-text pairs to enhance biological entity recognition, providing  
111 precise tools for specialized benchmarks in agricultural and ecological domains.

112 Recent benchmarks have been developed to evaluate AI models as agricultural assistants. Due to  
113 the scarcity of high-quality training data for LLMs in agriculture, studies often generate synthetic  
114 data by classifying content, as in Zhao et al. [62], and extracting dialogues from domain-specific  
115 corpora. However, agricultural knowledge is highly context-dependent, limiting these approaches  
116 to test models’ capabilities over specific regions [43, 7, 45] or topics, such as rice- and corn-related  
117 practices in CROP [61], pest management in Yang et al. [54] and irrigation methods in AgXQA [24].

118 Lastly, multimodal benchmarks have emerged, merging agricultural knowledge with visual under-  
119 standing. For example, AgEval [4] tests models on plant stress phenotyping tasks. CDDM [49]  
120 provides synthetic dialogues and open-ended question-answering on visual data for crop disease  
121 diagnosis. Similarly, Agri-LLaVA [49] introduces an open-domain exact-match question answer-  
122 ing benchmark and MMMU [60] offers a human-crafted benchmark with multiple choice question  
123 answering (MCQ), both focusing on crops’ pests and disease visual identification.

124 **AI-guided NLG evaluation.** Several techniques that demonstrate human alignment have been  
125 introduced: for example, Min et al. [36], Dubois et al. [12], Zhu et al. [66]. In particular, Liu et al. [33]  
126 proposes an Auto-Chain-of-Thought (Auto-CoT) approach coupled with a weighted scoring function.  
127 The LLM-based evaluation finds broad application. Some studies assess model outputs by comparing  
128 them with expert answers [58, 43] or retrieved contexts [45]. At the same time, other approaches  
129 employ an LLM to generate a reference answer—based on contextual or retrieved information—  
130 and subsequently compare the model’s output to this generated ground truth [10, 30, 6, 49, 32, 7].  
131 Notably, Balaguer et al. [7] distinguishes itself by also evaluating the quality of generated question  
132 and answer (Q&A) pairs and the introduction of a technique called Evaluation with Guideline—where  
133 an LLM produces specific criteria to assess the correctness of models’ outputs tailored to each of the  
134 ground-truth Q&A pairs.

135 LLMs have further been utilized to structure outputs in multiple-choice scenarios [34, 19] and  
136 to extract labels for classification tasks [54]. Additionally, it is common to leverage LLMs to  
137 evaluate intrinsic qualities of generated text, not tied to ground truth, such as relevance and detail  
138 [10, 30, 6, 49], succinctness [7] or, as in Yang et al. [54], coherence, logical consistency, fluency,  
139 relevance, comprehensiveness, and exhaustiveness where these scores are obtained with the G-Eval  
140 framework [33].

Table 1: **Overview of agriculture-specific benchmarks.** Benchmarks differ among synthetic data usage, modality, taxonomic system usage, number of crop and pest & pathogens classes, and task types.

Evaluation Benchmark	Synthetic Data	Input Modality	Categorical Granularity	Categories Number		Tasks Types
				Plant	Pest & Path.	
<i>CV-oriented</i>						
PlantVillage [1]	✗	I	SN	-	-	VI
PantDoc [1]	✗	I	SN	-	-	VI
BioTrove [1]	✗	I	SN	-	-	VI
<i>MLLMs-oriented</i>						
CROP [61]	✓	T	VN	2	-	K
AgEval [4]	✗	T+I	VN	5	46	VI+VR
AgriLLaVA [49]	✓	T+I	VN	21	221	K+VI+VR
CDDM [32]	✓	T+I	VN	16	60	K+VI+VR
Ours	✗	T+I	VN	864	632	K+VI+VR+P
Ours	✗	T+I	SN	1.2K	1.2K	K+VI+VR+P

T = Text, I = Image; VN = Vernacular Name, SN = Scientific Name;  
 K = Scientific Knowledge, VI = Visual Identification, VR = Visual Reasoning,  
 P = Procedural Reasoning.

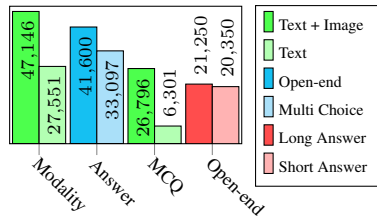


Figure 3: **CABBAGE Q&A types.** Samples count by question modalities, question types and answer types.

### 141 3 The CABBAGE Benchmark

142 In this section, we introduce our proposed CABBAGE dataset by first elaborating its structure  
 143 (Section 3.1), followed by the details of the data collection and annotation process (Section 3.2).

#### 144 3.1 Benchmark Structure

145 CABBAGE comprises *open-ended* and *multiple-choice questions*, covering *visual* and *textual* tasks to  
 146 comprehensively evaluate the depth and breadth of the model’s knowledge. The open-ended questions  
 147 are subdivided into *short-* and *long-answer* formats; see Figure 3. CABBAGE is organized into three  
 148 primary task macro-categories, each targeting a specific aspect of agricultural intelligence.

149 **Agricultural Scientific Knowledge.** Covers fundamental agricultural reasoning and concepts, in-  
 150 cluding plant phytoogy, soil science, irrigation techniques, and integrated pest management (IPM).  
 151 This category draws from Embrapa [13], AgriExam [2], and CCA-CEU [11] ensuring broad and  
 152 diverse knowledge coverage. Figure 7 in Appendix A shows question distribution over top-10 topic  
 153 categories, such as pest, soil, irrigation and post-harvest management and plant physiology and  
 154 pathology as well. See examples in Appendices C.2 and D.

155 **Agronomic Visual Cognition.** This section tests the model’s ability to recognize crops, pests, and  
 156 plant health conditions. Tasks are based on taxons and descriptions from Pl@ntNet [1] and EPPO  
 157 Global Database [14] to test pests and crops visual understanding, and questions from BPPQ [8]  
 158 challenging the models’ visual reasoning to diagnose diseases and recommend interventions.

159 As shown in Figure 4, questions involve identifying species (Taxon), genus (Genus), and vernacular  
 160 names (Common) from a single image. Species recognition is further tested with images depicting  
 161 various plant organ types (Organs). The model must also diagnose plant damage (Damage), identify  
 162 weed species (Weed), and determine pathogen growth stages (Growth). Figure 4 additionally  
 163 illustrates question coverage across natural kingdoms. See examples in Appendix D.

164 **Agricultural Procedural Reasoning.** This macro-category evaluates the model’s reasoning in  
 165 farming workflows such as planting and pruning (Plant Grow.), pest and weed management (Pest &  
 166 Weed), fertilization (Soil Prep.), etc.; see Figure 5. Tasks from wikiHow [52] focus on step-by-step  
 167 procedural knowledge, simulating real-world scenarios that require selecting actions, predicting  
 168 outcomes, and recommending preventive measures; see bottom row in Figure 2.

169 As shown in Figure 5, questions are structured into four main tasks with increasing difficulty levels.  
 170 The simplest task asks the agent to arrange steps correctly for a given procedure (Arrange Steps).  
 171 Next, the agent must identify the missing step in a partially masked procedure (Missing Step) or  
 172 predict the next step in an incomplete procedure (Next Step). The most challenging task involves  
 173 generating a full description of the entire procedure from scratch (All Steps). See Appendix D.

#### 174 3.2 Data Collection and Processing

175 CABBAGE combines publicly available, human-authored Q&As and rule-based generated data from  
 176 reputable, expert- or community-curated sources, ensuring expert knowledge, structured information,  
 177 no personal or sensitive data, and eliminating risks associated with LLM-generated content.

178 **Embrapa.** Embrapa is derived from 42 volumes of the “500 Perguntas 500 Respostas” collection  
 179 [13]. The content was translated into English using Llama 3.3 (70B) [35]. Thematic categories were  
 180 created based on the book and chapter titles with assistance from GPT-4o [39], resulting in 20,217  
 181 open-ended Q&A pairs across 14 categories.

182 **AgriExam.** The AgriExam quiz database was created from an extensive archive of multiple-choice  
 183 questions for agricultural competitive exams across India [2]. After cleaning and preprocessing, 4.5K  
 184 multiple-choice questions were organized into 29 agriculture-specific categories. Non-agricultural  
 185 content was excluded to ensure domain relevance, with thematic categories established using GPT-4o.

186 **CCA-CEU.** Derived from the Continuing Education Unit (CEU) quizzes for the Certified Crop  
 187 Adviser (CCA) exam [11], this dataset evaluates agronomic expertise. It covers core topics such as  
 188 crop protection, pest management, and environmental stewardship.

189 **BPPQ.** The Big Plant Pathology Quiz (BPPQ) dataset, published by the British Society for Plant  
 190 Pathology [8], consists of 370 agronomic visual reasoning questions. These questions focus on plant  
 191 disease symptoms, diagnosis, and management strategies.

192 **PI@ntNet.** PI@ntNet provides a curated subset of human-annotated plant images from the Global  
 193 Biodiversity Information Facility (GBIF) [1]. We algorithmically generated 20K questions for  
 194 identifying plant species based on images of single or multiple plant organs—such as *leaves*, *flowers*,  
 195 *fruits*, and *bark*—along with *habitat* views. This dataset is designed to evaluate model performance  
 196 in species recognition and quantify organ-specific biases in species identification. An example of  
 197 such task is shown in Figure 6.

198 **EPPO.** The EPPO database is a comprehensive source on plant pests, diseases, and associated  
 199 control measures [14]. We algorithmically obtained approximately 26K questions on 14K images,  
 200 spanning roughly 1.2K species of plants and 1.2K species of pests and pathogens. These tasks focus  
 201 on categorising images into species using taxonomic details and damage annotations; see Figure 4.

202 **wikiHow.** The wikiHow [52] dataset was crafted from 1K agriculture-related articles, focusing  
 203 on step-by-step gardening and farming practices guides, obtaining a total of 2.1K questions. We  
 204 adopted GPT-4o to generate option distractors for the multiple-choice tasks, missing (Miss. Step),  
 205 and next-step (Next Step.) prediction tasks; see Figure 5.

## 206 4 Evaluation Methodology

207 In this section, we present our evaluation methodology by first discussing structured output techniques  
 208 to enforce consistent response formatting (Section 4.1), followed by the use of deterministic metrics  
 209 for objective evaluation (Section 4.2). Finally, we describe non-deterministic evaluation methods,  
 210 including LLM-based judges for open-ended tasks (Section 4.3).

### 211 4.1 Structured Output

212 Despite clear multiple-choice options or explicit numerical scoring instructions, many MLLMs  
 213 still generate free-form text responses, particularly those not trained on multiple-choice formats or  
 214 proprietary general-purpose models [34]. To standardize evaluation and enable fair comparisons  
 215 across models with varying instruction-following abilities, we employ structured output techniques  
 216 [31], enforcing predefined JSON response formats.

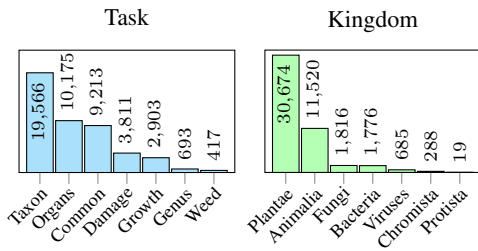


Figure 4: **Plant, pest and pathogens identification.** Sample counts with respect to the question tasks and kingdom types.

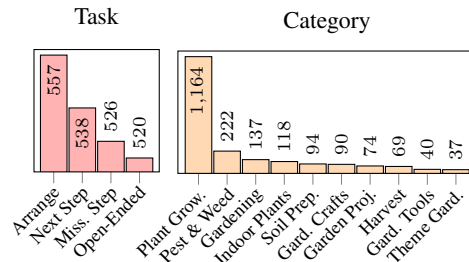


Figure 5: **Agronomic procedural reasoning.** Sample counts with respect to the question tasks and categories.

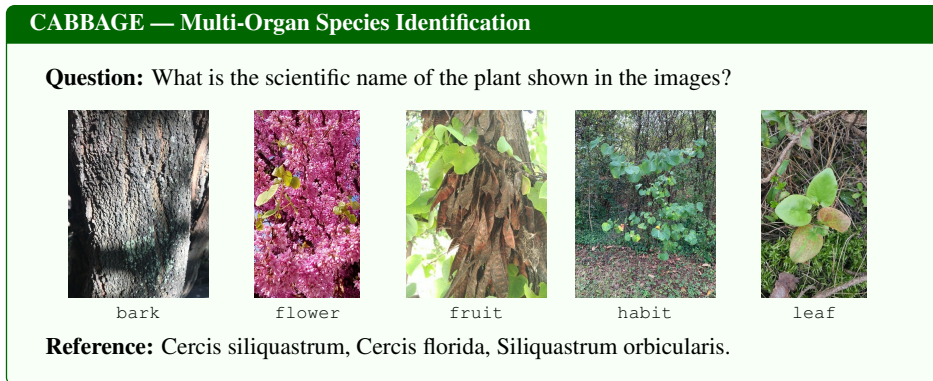


Figure 6: A sample of species identification from multiple organs task. CABBAGE tests MLLMs at identifying the plant species by attaining the global characteristics in five organ images together.

217 The impact of structured outputs on LLM reasoning remains active research [47, 26]. To mitigate  
 218 potential drawbacks, our structured formats incorporate fields that capture reasoning processes,  
 219 preserving explainability while maintaining the benefits of structured responses.

## 220 4.2 Deterministic Evaluation

221 When applicable, we use the deterministic metrics outlined below, as they enable objective and  
 222 consistent evaluation [17]:

- 223 • *Accuracy*: Multiple choice question answering is evaluated assessing how often the model’s output  
 224 matches the correct answer. To ensure fairness in the evaluation process and reduce any bias  
 225 related to the order of options [34], we randomize the sequence of the options.
- 226 • *Exact Match*: For open-ended questions requiring concise textual responses, multiple reference  
 227 answers may be valid. A model’s prediction is considered correct if it exactly matches any of  
 228 the ground truth references after applying a normalization procedure. This includes lowercasing,  
 229 stripping punctuation, standardizing whitespace, and removing articles, thereby ensuring that  
 230 trivial textual discrepancies do not influence the evaluation [41].

## 231 4.3 Non-deterministic Evaluation

232 We employ GPT-4o [39] to evaluate open-ended questions that demand detailed, long-form answers, in  
 233 contexts where conventional metrics fall short; supplementary analyses are presented in Appendix C.

234 **Evaluation prompt.** We provide the evaluator with a question, an expert answer, and a model-  
 235 generated response. When assessing knowledge-based tasks, the evaluator is prompted to assign  
 236 scores (1-5) based on the following criteria:

- 237 • *Correctness*: Measures the extent to which the model’s answer addresses the key points of the  
 238 question compared to the expert.
- 239 • *Specificity*: Evaluates the depth of agriculture-specific knowledge demonstrated in the model’s  
 240 response.
- 241 • *Conciseness*: Assesses the efficiency of the model’s answer in conveying the necessary information  
 242 without unnecessary verbosity.

243 When evaluating answers that describe a procedure, the evaluator will measure *Procedural Concise-*  
 244 *ness*, similar to the metric described above, and additionally:

- 245 • *Procedural Accuracy*: Assesses whether the model’s answer includes all necessary steps to achieve  
 246 the intended outcomes.
- 247 • *Procedural Flow*: Evaluates the clarity and logical sequence of the steps in the response.

248 To ensure consistent and systematic evaluation, we provide the evaluator with a detailed scoring sheet  
 249 for each criterion, similarly to Balaguer et al. [7]. The evaluator first analyses the question, the expert  
 250 answer, and the model’s answer, followed by a step-by-step assessment of the evaluation criteria.  
 251 This approach ensures Chain-of-Thought (CoT) [51], further reinforced by the structured output

(see Section 4.1), where each evaluation step is assigned a dedicated field to capture explicitly the intermediate reasoning. Finally, the evaluator assigns scores to each criterion in the corresponding structured output field.

We use a single prompt injection for all scoring criteria to optimize resource efficiency. For the same reason, we do not generate Evaluation Guidelines for each Q&A pair, as done in Balaguer et al. [7]. The detailed evaluation prompts used can be found in Listings 3 and 4 in Appendix C.1 and an example of the evaluation output procedure is presented in Appendix C.2.

**G-Eval score.** We do not directly use the scores but instead apply the scoring function proposed by Liu et al. [33], which mitigates the issue of low score variance (with 3 being the most frequent in a 1-5 scale) and captures finer distinctions that are not discernible when using integer scores. The score is calculated as  $G\text{-Eval} = \sum_{i=1}^n p(s_i) \times s_i$ . Where  $S = \{s_1, \dots, s_n\}$  is the set of possible scores, and  $p(s_i)$  is the probability assigned by the LLM for each score. These probabilities are derived from the evaluator’s logits corresponding to each score token (located through the structured output).

## 5 Experimental Setup and Results

**Setup and details.** The evaluation includes closed-source and open-source models, covering a range of architectures and parameter sizes. Closed-source models include GPT-4o and GPT-4o-mini [39], Gemini 2.0 Flash and Gemini 2.0 Flash-Lite [48], while open-source models such as Qwen2-VL-7B-Instruct and Qwen2-VL-72B-Instruct [50] were tested using six A6000 GPUs, taking respectively  $\approx 9$ h and  $\approx 2$ h of inference time. The SGLang [64] framework was employed for structured JSON generation to enhance inference efficiency and standardize outputs (see Section 4.1). All metrics—deterministic (Accuracy, Exact Match) and non-deterministic (G-Eval scores)—are computed over  $n = 5$  independent runs with sampling temperature  $t = 0.2$ , and reported as the mean. For conciseness, complete results including standard error of the mean (SEM) are provided in the Appendix C. Empirically, SEM values are consistently small and negligible across all benchmarks, typically on the order of  $\pm 0.01$  for classification metrics (e.g., Accuracy, Exact Match) and  $\pm 0.2$  for open-ended metrics (e.g., G-Eval scores), supporting the reliability of reported means.

**Agricultural Scientific Knowledge.** We evaluate models on agricultural knowledge tasks using multiple-choice and open-ended questions, with results summarized in Table 2. The multiple-choice evaluation is based on the AgriExam and CCA-CEU datasets, where models are assessed using accuracy over four possible answers in a zero-shot setting. For open-ended questions sourced from the Embrapa dataset, models must generate highly detailed, well-structured responses covering fundamental and applied agronomic topics. To ensure consistent and scalable evaluation of these responses, we employ GPT-4o, which scores answers from 1 to 5 based on Correctness, Specificity, and Conciseness, using a curated evaluation prompt (see Section 4.3). Final scores are computed with the G-Eval scoring function to capture subtle distinctions in model performance.

GPT-4o and Gemini 2.0 Flash demonstrate strong agronomic knowledge in multiple-choice tasks, performing at a comparable level across datasets. However, when handling open-ended questions, slight distinctions emerge in their ability to generate well-structured, precise, and informative answers. GPT-4o exhibits a marginal advantage in correctness and specificity, while Gemini 2.0 Flash closely follows, excelling in conciseness.

Table 2: **Zero-shot performance on agricultural scientific knowledge.** For Embrapa we use GPT-4o-guided evaluation metrics to quantify correctness, specificity, and conciseness.

Model	AgriExam	CCA-CEU	Embrapa			
	Acc.	Acc.	Corr.	Spec.	Conc.	Avg.
GPT-4o	<b>0.77</b>	0.89	<b>3.28</b>	3.18	<b>3.4</b>	<b>3.29</b>
GPT-4o-mini	0.61	0.88	2.96	<b>3.24</b>	3.05	3.08
Gemini 2 F	<b>0.77</b>	0.9	3.24	3.14	3.39	3.26
Gemini 2 FL	0.71	<b>0.91</b>	3.06	3.11	3.14	3.1
Qwen-2-VL-72B	0.63	0.82	2.78	3.13	2.81	2.91
Qwen-2-VL-7B	0.45	0.66	2.63	3.02	2.63	2.76

**Agronomic Visual Cognition.** We assess the visual cognition of MLLMs using the EPPO, Pl@ntNet, and BPPQ datasets. The EPPO dataset includes multiple-choice questions with five distractor difficulty levels based on taxonomic proximity and open-ended questions. Performance is quantified using accuracy for the multiple-choice questions and Exact Match for the open-ended questions. The Pl@ntNet dataset features open-ended plant identification tasks with single or multiple organ images evaluated via exact match. Performance on BPPQ is evaluated using accuracy. In all cases, experiments are conducted under zero-shot conditions—results in Table 3. Across these tasks, GPT-4o and Gemini 2.0 Flash demonstrate strong capabilities in visual recognition. We observe

Table 3: **Zero-shot performance on agronomic visual cognition.** Accuracy (Acc.) on EPPO is evaluated at different difficulty levels of multiple choice options (Diff-1, Diff-3, Diff-5) and exact match (EM) for short-answer open-ended questions (Open) and PI@ntNet species recognition from single and multi-organ images.

Model	BPPQ	EPPO				PI@ntNet	
		Diff-1	Diff-3	Diff-5	Open	Single	Multi
	Acc.	Acc.	Acc.	Acc.	EM	EM	EM
GPT-4o	<b>0.63</b>	<b>0.98</b>	<b>0.92</b>	0.71	<b>0.5</b>	<b>0.52</b>	<b>0.74</b>
GPT-4o-mini	0.5	0.89	0.86	0.66	0.35	0.18	0.38
Gemini 2 F	<b>0.63</b>	0.92	0.86	<b>0.72</b>	0.49	<b>0.52</b>	0.68
Gemini 2 FL	<b>0.63</b>	0.95	0.89	0.7	0.38	0.38	0.65
Qwen-2-VL-72B	0.56	0.84	0.81	0.63	0.2	0.26	0.46
Qwen-2-VL-7B	0.44	0.72	0.68	0.63	0.22	0.23	0.44

Table 4: **Zero-shot performance on agriculture procedural reasoning.** Accuracy (Acc.) is evaluated for arranging, identifying missing steps, and predicting the next step. For long-answer open-ended questions, GPT-4o-guided evaluation (G-Eval) metrics assess the procedure’s accuracy, flow, and conciseness.

Model	wikiHow						
	Arr.	Miss.	Next	All Steps			
	Acc.	Acc.	Acc.	P. Acc.	P. Flow	P. Conc.	Avg.
GPT-4o	0.65	<b>0.8</b>	<b>0.68</b>	3.73	<b>4.02</b>	<b>3.17</b>	<b>3.64</b>
GPT-4o-mini	0.58	0.71	0.61	3.54	3.8	2.94	3.42
Gemini 2 F	0.71	0.71	0.66	<b>3.79</b>	3.97	2.99	3.58
Gemini 2 FL	0.68	0.76	0.66	3.75	3.95	3.07	3.59
Qwen-2-VL-72B	<b>0.86</b>	0.6	0.55	2.54	2.84	2.58	2.65
Qwen-2-VL-7B	0.6	0.7	0.65	2.01	2.3	2.29	2.2

304 that question difficulty in EPPO and the number of visible plant organs in PI@ntNet significantly  
 305 impact model performance. The BPPQ dataset, which presents challenges in plant disease diagnosis,  
 306 proves difficult for all models. However, GPT-4o and Gemini 2.0 Flash achieve the highest accuracy,  
 307 suggesting an advantage in integrating visual and agronomic knowledge.

308 **Agricultural Procedural Reasoning.** To evaluate the procedural reasoning abilities of MLLMs,  
 309 we use the wikiHow dataset, which includes four types of tasks: Arrange, Missing Step, Next Step,  
 310 and All Steps. The first three tasks are presented as multiple-choice questions, with performance  
 311 measured by accuracy, while the All Steps task requires the generation of a complete procedure and  
 312 is evaluated using the G-Eval scoring function based on three criteria: Procedural Accuracy, Flow  
 313 and Conciseness. Each criterion is scored on a scale from 1 to 5 (see Section 4.3). We conduct  
 314 experiments under a zero-shot setting. Results are shown in Table 4. In the WikiHow dataset,  
 315 Qwen-2-VL-72B showed high accuracy on the Arrange task, suggesting a strength in understanding  
 316 the correct sequence of steps in a procedure. GPT-4o outperformed other models in the open-ended  
 317 task, as measured by the G-Eval score, indicating a better ability to generate complete and coherent  
 318 procedural descriptions.

## 319 6 Conclusion

320 We introduce CABBAGE, the first comprehensive benchmark to evaluate multimodal large language  
 321 models on domain-specific agricultural tasks. To build this benchmark, we compile an extensive  
 322 dataset comprising about 74K questions across three key areas: scientific knowledge, visual cognition,  
 323 and procedural reasoning in agriculture. The benchmark features a robust LLM-guided evaluation  
 324 framework that leverages explicit Chain-of-Thought reasoning, structured output, and fine-grained  
 325 G-Eval scoring. Experiments demonstrate that state-of-the-art models exhibit promising proficiency  
 326 in handling complex agricultural scenarios. However, they still face significant challenges in fully  
 327 capturing the specialized knowledge demanded by our task design. This reveals opportunities for  
 328 advancing MLLMs by integrating domain-specific knowledge. By addressing these challenges, future  
 329 research can use CABBAGE to improve model reliability and performance in agricultural tasks,  
 330 representing a significant advance in AI-driven agronomy.

331 **Limitations.** While GPT-4o has shown strong alignment with human evaluation in various contexts,  
 332 its reliability in specialized domains such as agriculture requires further investigation. Although the  
 333 benchmark includes diverse regions—Europe, the Mediterranean, Brazil, and India—it lacks full  
 334 global coverage, potentially limiting applicability to areas with different agricultural practices and  
 335 underscoring the need for additional evaluation. Furthermore, the dataset is currently English-only  
 336 and biased toward English-accessible sources, which may restrict the generalizability of results to  
 337 non-English-speaking regions and their agricultural systems.

338 **Broader impact.** CABBAGE promotes AI development for agriculture, a key sector for global  
 339 food security. Risks include incorrect guidance harming crops or communities, and limited coverage  
 340 of diverse regions and languages. To reduce these, we use expert-curated public data and do not  
 341 release trained models.

## References

- 342
- 343 [1] A. Affouard, A. Joly, J. Lombardo, J. Champ, H. Goeau, M. Chouet, H. Gresse, and P. Bonnet. Pl@ntnet  
344 observations. <https://doi.org/10.15468/gtebaa>, 2023. Accessed: 2025-10-13.
- 345 [2] AgriExam. Agriculture previous year question paper. [https://www.agriexam.com/  
346 agriculture-previous-year-question-paper](https://www.agriexam.com/agriculture-previous-year-question-paper), 2024. Accessed: 2024-11-01.
- 347 [3] Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark  
348 Ibrahim. Unibench: Visual reasoning requires rethinking vision-language beyond scaling. *arXiv preprint  
349 arXiv:2408.04810*, 2024.
- 350 [4] Muhammad Arbab Arshad, Talukder Zaki Jubery, Tirtho Roy, Rim Nassiri, Asheesh K Singh, Arti Singh,  
351 Chinmay Hegde, Baskar Ganapathysubramanian, Aditya Balu, Adarsh Krishnamurthy, et al. Ageval: A  
352 benchmark for zero-shot and few-shot plant stress phenotyping with multimodal llms. *arXiv preprint  
353 arXiv:2407.19617*, 2024.
- 354 [5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and  
355 Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint  
356 arXiv:2308.12966*, 2023.
- 357 [6] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang  
358 Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv  
359 preprint arXiv:2308.16890*, 2023.
- 360 [7] Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M Estevão Filho, Todd  
361 Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O Nunes, et al.  
362 Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv e-prints*, pages arXiv–2401,  
363 2024.
- 364 [8] The British Society for Plant Pathology - BSPP. The Big Plant Pathology Quiz. [https://www.bspp.  
365 org.uk/bspp-monthly-symptoms-quiz](https://www.bspp.org.uk/bspp-monthly-symptoms-quiz), 2024. Accessed: 2024-11-02.
- 366 [9] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong  
367 Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv  
368 preprint arXiv:2501.17811*, 2025.
- 369 [10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan  
370 Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4  
371 with 90%\* chatgpt quality. *See https://vicuna.lmsys.org (accessed 14 April 2023)*, 2(3):6, 2023.
- 372 [11] Crop Protection Network. Cca ceu quizzes. [https://cropprotectionnetwork.org/  
373 cca-ceu-quizzes](https://cropprotectionnetwork.org/cca-ceu-quizzes), 2024. Accessed: 2024-11-02.
- 374 [12] Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin,  
375 Percy S Liang, and Tatsunori B Hashimoto. AlpacaFarm: A simulation framework for methods that learn  
376 from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.
- 377 [13] Embrapa. Mais500p500r. <https://mais500p500r.sct.embrapa.br/view/index.php>,  
378 2024. Accessed: 2024-11-01.
- 379 [14] European and Mediterranean Plant Protection Organization - EPPO. European and mediterranean plant  
380 protection organization global database. <https://gd.eppo.int/>, 2024. Accessed: 2024-10-01.
- 381 [15] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir  
382 Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computa-  
383 tional Linguistics*, 9:391–409, 2021.
- 384 [16] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng,  
385 Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A Comprehensive Evaluation Benchmark for  
386 Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*, 2024.
- 387 [17] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun,  
388 Ziwei Liu, Liang Wang, et al. MME-Survey: A comprehensive survey on evaluation of multimodal LLMs.  
389 *arXiv preprint arXiv:2411.15296*, 2024.
- 390 [18] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long,  
391 Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv  
392 preprint arXiv:2501.01957*, 2025.

- 393 [19] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith,  
394 Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In  
395 *European Conference on Computer Vision*, pages 148–166. Springer, 2024.
- 396 [20] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa  
397 matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the*  
398 *IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
- 399 [21] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P  
400 Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the*  
401 *IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.
- 402 [22] David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the  
403 development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015.
- 404 [23] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers*  
405 *and electronics in agriculture*, 147:70–90, 2018.
- 406 [24] Josué Kpodo, Parisa Kordjamshidi, and A Pouyan Nejadhashemi. Agxqa: A benchmark for advanced  
407 agricultural extension question answering. *Computers and Electronics in Agriculture*, 225:109349, 2024.
- 408 [25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen,  
409 Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision  
410 using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- 411 [26] Will Kurt. Say what you mean: A response to ‘let me speak freely’, 2024. Accessed: 2025-02-12.
- 412 [27] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-  
413 Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference*  
414 *on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024.
- 415 [28] Konstantinos G Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine  
416 learning in agriculture: A review. *Sensors*, 18(8):2674, 2018.
- 417 [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár,  
418 and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014:*  
419 *13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages  
420 740–755. Springer, 2014.
- 421 [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural*  
422 *information processing systems*, 36, 2024.
- 423 [31] Michael Xieyang Liu, Frederick Liu, Alexander J Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and  
424 Carrie J Cai. ” we need structured output”: Towards user-centered constraints on large language model  
425 output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9,  
426 2024.
- 427 [32] Xiang Liu, Zhaoxiang Liu, Huan Hu, Zezhou Chen, Kohou Wang, Kai Wang, and Shiguo Lian. A  
428 Multimodal Benchmark Dataset and Model for Crop Disease Diagnosis. In *European Conference on*  
429 *Computer Vision*, pages 157–170. Springer, 2025.
- 430 [33] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg  
431 evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.
- 432 [34] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi  
433 Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In  
434 *European conference on computer vision*, pages 216–233. Springer, 2025.
- 435 [35] Llama Team, AI @ Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- 436 [36] Sewon Min, Kalpesh Krishna, Xinxin Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke  
437 Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision  
438 in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural*  
439 *Language Processing*, pages 12076–12100, 2023.
- 440 [37] Umair Nawaz, Muhammad Awais, Hanan Gani, Muzammal Naseer, Fahad Khan, Salman Khan, and  
441 Rao Muhammad Anwer. AgriCLIP: Adapting CLIP for Agriculture and Livestock via Domain-Specialized  
442 Cross-Model Alignment. *arXiv preprint arXiv:2410.01407*, 2024.

- 443 [38] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new  
444 evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural*  
445 *Language Processing*, pages 2241–2252, 2017.
- 446 [39] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- 447 [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish  
448 Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from  
449 natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR,  
450 2021.
- 451 [41] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for  
452 machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural*  
453 *Language Processing*, pages 2383–2392, 2016.
- 454 [42] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A  
455 novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23  
456 (3):289–301, 2022.
- 457 [43] Bruno Silva, Leonardo Nunes, Roberto Estevão, Vijay Aski, and Ranveer Chandra. Gpt-4 as an agronomist  
458 assistant? answering agriculture exams using large language models. *arXiv preprint arXiv:2310.06225*,  
459 2023.
- 460 [44] Davinder Singh, Naman Jain, Pranjali Jain, Pratik Kayal, Sudhakar Kumawat, and Nipun Batra. PlantDoc:  
461 A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*,  
462 pages 249–253. ACM, 2020.
- 463 [45] Namita Singh, Jacqueline Wang’ombe, Nereah Okanga, Tetyana Zelenska, Jona Repishti, Sanjeev Mishra,  
464 Rajsekhar Manokaran, Vineet Singh, Mohammed Irfan Rafiq, Rikin Gandhi, et al. Farmer.Chat: Scaling  
465 AI-Powered Agricultural Services for Smallholder Farmers. *arXiv preprint arXiv:2409.08916*, 2024.
- 466 [46] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Ed-  
467 ward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. BioCLIP: A vision  
468 foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision*  
469 *and Pattern Recognition*, pages 19412–19424, 2024.
- 470 [47] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me  
471 speak freely? a study on the impact of format restrictions on performance of large language models. *arXiv*  
472 *preprint arXiv:2408.02442*, 2024.
- 473 [48] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan  
474 Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable  
475 multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- 476 [49] Liqiong Wang, Teng Jin, Jinyu Yang, Ales Leonardis, Fangyi Wang, and Feng Zheng. Agri-llava:  
477 Knowledge-infused large multimodal assistant on agricultural pests and diseases. *arXiv preprint*  
478 *arXiv:2412.02158*, 2024.
- 479 [50] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin  
480 Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model’s perception of the world at any  
481 resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- 482 [51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny  
483 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural*  
484 *information processing systems*, 35:24824–24837, 2022.
- 485 [52] wikiHow contributors. wikihow: The how-to manual that you can edit. <https://www.wikihow.com>,  
486 2025. Accessed: 2025-02-11.
- 487 [53] Chih-Hsuan Yang, Benjamin Feuer, Talukder Zaki Jubery, Zi K Deng, Andre Nakkab, Md Zahid Hasan,  
488 Shivani Chiranjeevi, Kelly O Marshall, Nirmal Baishnab, Asheesh K Singh, et al. BioTrove: A Large  
489 Curated Image Dataset Enabling AI for Biodiversity. In *The Thirty-eight Conference on Neural Information*  
490 *Processing Systems Datasets and Benchmarks Track*, 2024.
- 491 [54] Shanglong Yang, Zhipeng Yuan, Shunbao Li, Ruoling Peng, Kang Liu, and Po Yang. GPT-4 as Evaluator:  
492 Evaluating Large Language Models on Pest Management in Agriculture. *arXiv preprint arXiv:2403.11858*,  
493 2024.

- 494 [55] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong  
495 Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset,  
496 framework, and benchmark. *Advances in Neural Information Processing Systems*, 36, 2024.
- 497 [56] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang,  
498 Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large  
499 vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.
- 500 [57] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual  
501 denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the  
502 Association for Computational Linguistics*, 2:67–78, 2014.
- 503 [58] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and  
504 Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint  
505 arXiv:2308.02490*, 2023.
- 506 [59] Zhou Yu, Yuhao Cui, Zhenwei Shao, Pengbing Gao, and Jun Yu. Openvqa. [https://github.com/  
507 MILVLG/openvqa](https://github.com/MILVLG/openvqa), 2019.
- 508 [60] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu  
509 Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding  
510 and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer  
511 Vision and Pattern Recognition*, pages 9556–9567, 2024.
- 512 [61] Hang Zhang, Jiawei Sun, Renqi Chen, Wei Liu, Zhonghang Yuan, Xinzhe Zheng, Zhefan Wang, Zhiyuan  
513 Yang, Hang Yan, Han-Sen Zhong, et al. Empowering and assessing the utility of large language models in  
514 crop science. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and  
515 Benchmarks Track*, 2024.
- 516 [62] Biao Zhao, Weiqiang Jin, Javier Del Ser, and Guang Yang. Chatagri: Exploring potentials of chatgpt on  
517 cross-linguistic agricultural text classification. *Neurocomputing*, 557:126708, 2023.
- 518 [63] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,  
519 Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena.  
520 *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
- 521 [64] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao,  
522 Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient  
523 Execution of Structured Language Model Programs, 2024.
- 524 [65] Yang-Yang Zheng, Jian-Lei Kong, Xue-Bo Jin, Xiao-Yi Wang, Ting-Li Su, and Min Zuo. CropDeep: The  
525 crop vision dataset for deep-learning-based classification and detection in precision agriculture. *Sensors*,  
526 19(5):1058, 2019.
- 527 [66] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are  
528 scalable judges. *arXiv preprint arXiv:2310.17631*, 2023.

## 529 **NeurIPS Paper Checklist**

530 The checklist is designed to encourage best practices for responsible machine learning research,  
531 addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove  
532 the checklist: **The papers not including the checklist will be desk rejected.** The checklist should  
533 follow the references and follow the (optional) supplemental material. The checklist does NOT count  
534 towards the page limit.

535 Please read the checklist guidelines carefully for information on how to answer these questions. For  
536 each question in the checklist:

- 537 • You should answer [Yes], [No], or [NA].
- 538 • [NA] means either that the question is Not Applicable for that particular paper or the  
539 relevant information is Not Available.
- 540 • Please provide a short (1–2 sentence) justification right after your answer (even for NA).

541 **The checklist answers are an integral part of your paper submission.** They are visible to the  
542 reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it  
543 (after eventual revisions) with the final version of your paper, and its final version will be published  
544 with the paper.

545 The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation.  
546 While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]"  
547 provided a proper justification is given (e.g., "error bars are not reported because it would be too  
548 computationally expensive" or "we were unable to find the license for the dataset we used"). In  
549 general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased  
550 in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your  
551 best judgment and write a justification to elaborate. All supporting evidence can appear either in the  
552 main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in  
553 the justification please point to the section(s) where related material for the question can be found.

554 **IMPORTANT, please:**

- 555 • **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist",**
- 556 • **Keep the checklist subsection headings, questions/answers and guidelines below.**
- 557 • **Do not modify the questions and only use the provided macros for your answers.**

### 558 **1. Claims**

559 Question: Do the main claims made in the abstract and introduction accurately reflect the  
560 paper's contributions and scope?

561 Answer: [Yes]

562 Justification: The abstract and the introduction clearly state the contributions made (see lines  
563 2–11, 15–17, 80–89) in the paper and important assumptions (see Section 4) and limitations  
564 (see lines 9–15 and Section 6).

565 Guidelines:

- 566 • The answer NA means that the abstract and introduction do not include the claims  
567 made in the paper.
- 568 • The abstract and/or introduction should clearly state the claims made, including the  
569 contributions made in the paper and important assumptions and limitations. A No or  
570 NA answer to this question will not be perceived well by the reviewers.
- 571 • The claims made should match theoretical and experimental results, and reflect how  
572 much the results can be expected to generalize to other settings.
- 573 • It is fine to include aspirational goals as motivation as long as it is clear that these goals  
574 are not attained by the paper.

### 575 **2. Limitations**

576 Question: Does the paper discuss the limitations of the work performed by the authors?

577 Answer: [Yes]

578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629

Justification: see “Limitations” paragraph in Section 6.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

**3. Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

**4. Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We do provide clear description for reproducibility of the steps taken to make results reproducible or verifiable, see Sections 4 and 5 and Appendices B and C.

Guidelines:

- 630 • The answer NA means that the paper does not include experiments.
- 631 • If the paper includes experiments, a No answer to this question will not be perceived
- 632 well by the reviewers: Making the paper reproducible is important, regardless of
- 633 whether the code and data are provided or not.
- 634 • If the contribution is a dataset and/or model, the authors should describe the steps taken
- 635 to make their results reproducible or verifiable.
- 636 • Depending on the contribution, reproducibility can be accomplished in various ways.
- 637 For example, if the contribution is a novel architecture, describing the architecture fully
- 638 might suffice, or if the contribution is a specific model and empirical evaluation, it may
- 639 be necessary to either make it possible for others to replicate the model with the same
- 640 dataset, or provide access to the model. In general, releasing code and data is often
- 641 one good way to accomplish this, but reproducibility can also be provided via detailed
- 642 instructions for how to replicate the results, access to a hosted model (e.g., in the case
- 643 of a large language model), releasing of a model checkpoint, or other means that are
- 644 appropriate to the research performed.
- 645 • While NeurIPS does not require releasing code, the conference does require all submissions
- 646 to provide some reasonable avenue for reproducibility, which may depend on the
- 647 nature of the contribution. For example
  - 648 (a) If the contribution is primarily a new algorithm, the paper should make it clear how
  - 649 to reproduce that algorithm.
  - 650 (b) If the contribution is primarily a new model architecture, the paper should describe
  - 651 the architecture clearly and fully.
  - 652 (c) If the contribution is a new model (e.g., a large language model), then there should
  - 653 either be a way to access this model for reproducing the results or a way to reproduce
  - 654 the model (e.g., with an open-source dataset or instructions for how to construct
  - 655 the dataset).
  - 656 (d) We recognize that reproducibility may be tricky in some cases, in which case
  - 657 authors are welcome to describe the particular way they provide for reproducibility.
  - 658 In the case of closed-source models, it may be that access to the model is limited in
  - 659 some way (e.g., to registered users), but it should be possible for other researchers
  - 660 to have some path to reproducing or verifying the results.

## 661 5. Open access to data and code

662 Question: Does the paper provide open access to the data and code, with sufficient instruc-  
 663 tions to faithfully reproduce the main experimental results, as described in supplemental  
 664 material?

665 Answer: [Yes]

666 Justification: We do provide full benchmark code and dataset at [https://github.com/](https://github.com/boilnserve/cabbage)  
 667 [boilnserve/cabbage](https://github.com/boilnserve/cabbage) (code) and [https://huggingface.co/datasets/](https://huggingface.co/datasets/boilnserve/cabbage)  
 668 [boilnserve/cabbage](https://huggingface.co/datasets/boilnserve/cabbage) (dataset).

669 Guidelines:

- 670 • The answer NA means that paper does not include experiments requiring code.
- 671 • Please see the NeurIPS code and data submission guidelines ([https://nips.cc/](https://nips.cc/public/guides/CodeSubmissionPolicy)  
 672 [public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 673 • While we encourage the release of code and data, we understand that this might not be
- 674 possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not
- 675 including code, unless this is central to the contribution (e.g., for a new open-source
- 676 benchmark).
- 677 • The instructions should contain the exact command and environment needed to run to
- 678 reproduce the results. See the NeurIPS code and data submission guidelines ([https://](https://nips.cc/public/guides/CodeSubmissionPolicy)  
 679 [nips.cc/public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 680 • The authors should provide instructions on data access and preparation, including how
- 681 to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 682 • The authors should provide scripts to reproduce all experimental results for the new
- 683 proposed method and baselines. If only a subset of experiments are reproducible, they
- 684 should state which ones are omitted from the script and why.

- 685                   • At submission time, to preserve anonymity, the authors should release anonymized  
686                   versions (if applicable).  
687                   • Providing as much information as possible in supplemental material (appended to the  
688                   paper) is recommended, but including URLs to data and code is permitted.

## 689 6. Experimental setting/details

690 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-  
691 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the  
692 results?

693 Answer: [Yes]

694 Justification: We do provide specify all the training and test details, see Sections 4 and 5  
695 and Appendices B and C.

696 Guidelines:

- 697                   • The answer NA means that the paper does not include experiments.  
698                   • The experimental setting should be presented in the core of the paper to a level of detail  
699                   that is necessary to appreciate the results and make sense of them.  
700                   • The full details can be provided either with the code, in appendix, or as supplemental  
701                   material.

## 702 7. Experiment statistical significance

703 Question: Does the paper report error bars suitably and correctly defined or other appropriate  
704 information about the statistical significance of the experiments?

705 Answer: [Yes]

706 Justification: We do report confidence intervals of results for each experiment and for each  
707 metric; see Section 5. We do explain the method to compute the confidence interval and  
708 state explicitly the distribution assumptions made. As shown in the provided code, we use  
709 python libraries `numpy` and `scipy` for computing the intervals.

710 Guidelines:

- 711                   • The answer NA means that the paper does not include experiments.  
712                   • The authors should answer "Yes" if the results are accompanied by error bars, confi-  
713                   dence intervals, or statistical significance tests, at least for the experiments that support  
714                   the main claims of the paper.  
715                   • The factors of variability that the error bars are capturing should be clearly stated (for  
716                   example, train/test split, initialization, random drawing of some parameter, or overall  
717                   run with given experimental conditions).  
718                   • The method for calculating the error bars should be explained (closed form formula,  
719                   call to a library function, bootstrap, etc.)  
720                   • The assumptions made should be given (e.g., Normally distributed errors).  
721                   • It should be clear whether the error bar is the standard deviation or the standard error  
722                   of the mean.  
723                   • It is OK to report 1-sigma error bars, but one should state it. The authors should  
724                   preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis  
725                   of Normality of errors is not verified.  
726                   • For asymmetric distributions, the authors should be careful not to show in tables or  
727                   figures symmetric error bars that would yield results that are out of range (e.g. negative  
728                   error rates).  
729                   • If error bars are reported in tables or plots, The authors should explain in the text how  
730                   they were calculated and reference the corresponding figures or tables in the text.

## 731 8. Experiments compute resources

732 Question: For each experiment, does the paper provide sufficient information on the com-  
733 puter resources (type of compute workers, memory, time of execution) needed to reproduce  
734 the experiments?

735 Answer: [Yes]

736 Justification: See Section 5.

737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: We did all necessary to conform to the NeurIPS Code of Ethics and preserve anonymity.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See paragraph "Broader impacts" in Section 6.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843

Answer: [NA]

Justification: We do not release models or datasets that pose a high risk of misuse. The data used in our benchmark are collected exclusively from reliable, publicly available sources with established reputations for domain-specific accuracy and educational intent. These include institutional repositories (e.g., governmental agencies, international organizations), expert-moderated platforms such as PI@ntNet (curated by botanists and research institutions), and high-quality community-driven resources like wikiHow, where content is generated, reviewed, and refined by a large base of knowledgeable contributors following a transparent editorial process.

All included datasets are carefully curated to reflect authoritative knowledge, consensus practices, or validated annotations, and no personal or sensitive information is ever used. Furthermore, we do not release any models trained on these datasets. Thus, the benchmark poses no significant ethical, privacy, or security risks associated with data or model misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We do explicitly mention original owners of the assets as well as source URLs. We do provide complete and clear reference to the original owners of the assets as well as versions and last access date to the source for online data. For each scraped data and the relative entry in the benchmark dataset we do include all original copyright information, name of licenses and term of use of data both in the 'metadata' payload of the entry and, if any, in the image metadata (e.g. using Python Pillow package API: `PIL.Image.info`).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894

Answer: [Yes]

Justification: We do provide documented code and dataset for the benchmark.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

#### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

895 Justification: We do describe how LLM have been used for categorising QA pairs and  
896 for generating distractors for procedural reasoning task (Section 3.2), also to score model  
897 responses at evaluation time via G-Eval [33] (Section 4).

898 Guidelines:

- 899 • The answer NA means that the core method development in this research does not  
900 involve LLMs as any important, original, or non-standard components.
- 901 • Please refer to our LLM policy ([https://neurips.cc/Conferences/2025/](https://neurips.cc/Conferences/2025/LLM)  
902 LLM) for what should or should not be described.

## 903 A Benchmark Dataset Details

### 904 A.1 Agricultural Scientific Knowledge Categories

905 Figure 7 shows the distribution of question categories by count across the three knowledge datasets:  
906 Embrapa, AgriExam, and CCA CEU. Each bar chart highlights the top 10 categories for each dataset,  
907 illustrating the focus areas such as pest management, soil science, and plant diseases, providing an  
overview of the thematic coverage in agricultural expert knowledge.

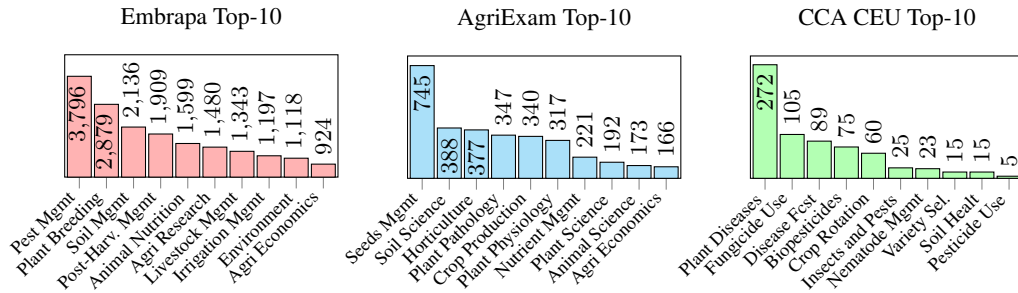


Figure 7: **Agricultural Scientific Knowledge topics.** Question categories by count for each source.

908

### 909 A.2 Data Provenance and Reliability

910 All datasets in our benchmark are drawn from sources with well-established expertise, rigorous  
911 curation procedures, and broad recognition in their respective domains. Each provider operates  
912 under transparent governance—through governmental bodies, international consortia, professional  
913 societies, or expert communities—and makes its materials freely accessible for educational and  
914 research purposes.

915 **Embrapa [13].** The Brazilian Agricultural Research Corporation is Brazil’s leading public agronomic  
916 research institute, with decades of peer-reviewed outputs and a national mandate for rural education.  
917 Its “500 Perguntas 500 Respostas” volumes are developed by in-house specialists and distributed at no  
918 cost to support agricultural learning.

919 **AgriExam [2].** AgriExam is overseen by a consortium of academic instructors and certified  
920 agronomists who author and validate each quiz item. Its open access content is continuously re-  
921 viewed to align with current agronomic curricula and professional standards, ensuring high relevance  
922 for both students and practitioners.

923 **CCA-CEU [11].** The Certified Crop Adviser (CCA) program is administered by the American  
924 Society of Agronomy in partnership with the American Society for Horticultural Science. Its  
925 continuing-education quizzes and materials undergo formal approval by expert working groups,  
926 guaranteeing up-to-date, peer-endorsed guidance in crop science. The quizzes are made available  
927 free of charge for professional development.

928 **BPPQ [8].** The Big Plant Pathology Quiz is hosted by the British Society for Plant Pathology,  
929 with content authored and vetted by leading plant-disease researchers. All quizzes and explanatory  
930 materials are freely accessible for academic and training use under the Society’s educational terms.

931 **PI@ntNet [1].** PI@ntNet operates under a consortium of CIRAD, Inria, INRAE and IRD, with  
932 botanical oversight from major research institutions. Its community-verified observations adhere  
933 to strict quality controls and expert validation. Its occurrence datasets are published via the Global  
934 Biodiversity Information Facility (GBIF).

935 **EPPO [14].** The European and Mediterranean Plant Protection Organization is an intergovernmental  
936 body that develops and maintains international phytosanitary standards through expert panels. Its  
937 EPPO Codes are released under an EPPO Open Data Licence, granting free, perpetual reuse (including  
938 derivatives and commercial use) provided the EPPO source is attributed.

939 **wikiHow [52].** wikiHow’s how-to articles are composed and refined through a layered editorial  
940 process: initial drafts by subject enthusiasts, fact-checking by volunteer editors, and ongoing peer

941 review by a large global community, ensuring clarity and practical accuracy. Text content is published  
942 under CC BY-NC-SA 3.0 for non-commercial reuse.

## 943 B Experimental Details

### 944 B.1 Structured Output Examples

945 Listing 1 presents the structured output for multiple-choice, short answer, and open-ended tasks  
946 and for evaluation. These outputs are used in the model’s inference to ensure that it responds in the  
expected format, maintaining consistency and alignment with predefined criteria.

```
class MultipleChoiceReasoningProcess(BaseModel):
    reasoning: str
    answer: Literal['A', 'B', 'C', 'D', 'E']

class ShortAnswerReasoningProcess(BaseModel):
    reasoning: str
    short_answer: str

class OpenEndedReasoningProcess(BaseModel):
    reasoning: str
    complete_answer: str

class Correctness(Enum):
    INCORRECT = '1'
    PARTIALLY = '2'
    MOSTLY = '3'
    LARGELY = '4'
    COMPLETE = '5'

class Specificity(Enum):
    GENERAL = '1'
    BASIC = '2'
    COMPARABLE = '3'
    IN_DEPTH = '4'
    SUPERIOR = '5'

class Conciseness(Enum):
    VERBOSE = '1'
    DETAILED = '2'
    BALANCED = '3'
    CONCISE = '4'
    EFFICIENT = '5'

class EvaluationStep(BaseModel):
    step_analysis: str

class Scores(BaseModel):
    correctness_score: Correctness
    specificity_score: Specificity
    conciseness_score: Conciseness

class EvaluationProcess(BaseModel):
    evaluation_steps: list[EvaluationStep]
    scores: Scores
    final_review: str
```

Listing 1: **Structured output Python classes.** Each question modality requires a different structure, and each metric requires its own format.

947

## 948 C Non-deterministic Evaluation

### 949 C.1 Evaluation Prompts

950 Listings 3 and 4 present the complete evaluation prompts provided to the LLM evaluator for assessing  
951 knowledge-based and procedural reasoning tasks, respectively. These prompts guide the evaluator  
952 in scoring model responses based on predefined criteria, ensuring consistency and reliability in our  
953 assessment.

```

Task Description:
  You will be given a question, an expert answer (ground truth), and a model-generated
  answer. Your task is to evaluate the model's answer using the defined evaluation criteria
  and the relative scoring sheet. Follow the outlined steps systematically, providing a
  detailed analysis for each step before assigning scores for the criteria.
---
Evaluation Criteria:
  1. Correctness [1-5]: Assesses how thoroughly the model's answer addresses the key points
  and requirements of the question compared to the expert answer.
  2. Specificity [1-5]: Evaluates the depth of agriculture-specific knowledge in the model's
  response relative to the expert answer.
  3. Conciseness [1-5]: Rates the efficiency of the model's response in using words while
  covering the required information compared to the expert answer. Where 1 indicates a
  verbose answer and 5 a highly efficient one. Brevity without meaningful or relevant
  information should receive a low score.
---
Scoring Sheet:
  - Correctness:
    1: Incorrect or irrelevant answer.
    2: Partially correct but lacks some important details.
    3: Mostly correct with some missing details.
    4: Largely correct with minor omissions.
    5: Fully addresses the question, matching expert-level detail.

  - Specificity:
    1: General knowledge, no agriculture specificity.
    2: Basic agriculture knowledge, minimal depth.
    3: Moderate agriculture-specific knowledge with partial depth.
    4: Highly relevant agricultural knowledge, near expert level.
    5: Matches or exceeds expert-level agriculture knowledge.

  - Conciseness:
    1: Verbose, inefficient response.
    2: Over-detailed with unnecessary information.
    3: Balanced and clear.
    4: Concise and focused on essentials.
    5: Efficient and precise.
---
Evaluation Steps:
  1. Understand the Question: Identify its key requirements and scope.
  2. Understand the Expert Answer: Determine its key points, agricultural specificity, and
  conciseness.
  3. Understand the Model's Answer: Determine its key points, agricultural specificity, and
  conciseness.
  4. Correctness Analysis: Assess whether the model's answer fully addresses the key points
  from the expert answer.
  5. Specificity Analysis: Determine if the model's answer demonstrates agricultural
  specificity in alignment with the expert answer.
  6. Conciseness Analysis: Determine how efficiently the model's answer conveys information
  compared to the expert answer, ensuring that the response is not only concise but also
  meaningful.
  7. Assign Scores: Using the Evaluation Criteria and the relative Scoring Sheet, assign
  scores for Correctness, Specificity, and Conciseness based on your analyses.
---
Question:
  {QUESTION}

Expert Answer:
  {EXPERT_ANSWER}

Model's Answer:
  {MODEL_ANSWER}
---
Evaluation Process:
  For each Evaluation Step, provide your answer and then assign scores based on the
  Evaluation Criteria. The final scores should be numerical values based on the defined
  Scoring Sheet.

```

**Listing 2: Evaluation prompt for knowledge-based tasks.** Evaluation prompt comprising evaluation criteria, scoring sheet and explicit evaluation steps guideline for judging model-generated answers for knowledge-based tasks.

Task Description:  
 You will be given a question, an expert answer (ground truth), and a model-generated answer. In this task, both responses present procedures in a step-by-step format. Your task is to evaluate the model's answer using the defined evaluation criteria and the corresponding scoring sheet. Follow the outlined steps systematically, providing a detailed analysis for each step before assigning scores for the criteria.

---

Evaluation Criteria:

1. Procedural Accuracy [1-5]: Assesses whether the model's answer covers all the essential steps and intermediate actions required to achieve the intended outcomes as specified by the expert answer.
2. Procedural Flow [1-5]: Evaluates whether the procedure is presented in a clear, sequential order with well-divided steps that are practically feasible and closely aligned with the structure of the expert answer.
3. Conciseness [1-5]: Rates the efficiency of the model's answer in terms of both the number of steps and the length of each step relative to the expert answer. Maximum scores can be awarded even if the number of steps differs, provided that the conveyed information is equivalent and communicated efficiently.

---

Scoring Sheet:

- Procedural Accuracy:
  - 1: Largely incorrect; several essential steps are missing.
  - 2: Partially correct; multiple key steps are missing or inaccurate.
  - 3: Mostly correct but lacks some critical components.
  - 4: Nearly all steps are present with minor omissions.
  - 5: Fully accurate; all necessary steps and actions are included.
- Procedural Flow:
  - 1: Extremely disorganized and illogical.
  - 2: Poorly organized with gaps in logical flow.
  - 3: Generally logical but with uneven step divisions.
  - 4: Well organized with minor deviations from the expert answer.
  - 5: Very clear and logically sequenced.
- Conciseness:
  - 1: Verbose and inefficient.
  - 2: Overly detailed with unnecessary information.
  - 3: Balanced but with some extra detail.
  - 4: Concise and focused on essentials.
  - 5: Exceptionally concise and precise.

---

Evaluation Steps:

1. Identify the procedure's objectives, constraints, and contextual factors as present in the question.
2. Understand the Expert Answer: Determine the expert answer's essential steps, logical flow, and critical implicit actions.
3. Understand the Model's Answer: Determine the model answer's essential steps, logical flow, and any straightforward implicit actions.
3. Understand the Model's Answer: Extract and identify the essential procedural steps and intermediate actions as presented by the model, noting the logical sequence and division of steps, including any straightforward implicit instructions that are clear and simple enough to follow without causing ambiguity or confusion.
4. Procedural Accuracy Analysis: Compare the essential steps and intermediate actions identified from the expert answer with those present in the model's answer. Evaluate how completely and accurately the model's answer covers the critical components required to achieve the intended outcomes.
5. Procedural Flow Analysis: Evaluate whether the model's answer presents the procedure in a clear and logically sequential order with well-defined steps. Verify the procedure's feasibility by ensuring its logical flow and structure align with the expert answer.
6. Conciseness Analysis: Evaluate the model's answer for efficiency by comparing step count and length to the expert answer, ensuring it conveys necessary information without excess or omission.
7. Assign Scores: Based on your analysis for each criterion, assign numerical scores (1-5) according to the provided scoring sheets.

---

Question:  
 {QUESTION}

Expert Answer:  
 {EXPERT\_ANSWER}

Model's Answer:  
 {MODEL\_ANSWER}

---

Evaluation Process:  
 For each Evaluation Step, provide your detailed analysis and then assign scores based on the Evaluation Criteria. The final scores should be numerical values between 1 and 5 for Procedural Accuracy, Procedural Flow, and Conciseness.

**Listing 3: Evaluation prompt for procedural reasoning tasks.** Evaluation prompt comprising evaluation criteria, scoring sheet and explicit evaluation steps guideline for judging model-generated answers for procedural reasoning tasks.

954 **C.2 Evaluation Example**

955 We present an example of an open-ended agricultural knowledge question from Embrapa, along with  
 956 model responses and detailed G-Eval evaluations. Specifically, Table 5 shows the input question,  
 957 expert answer, and evaluator analysis (first two evaluation steps). The models’ answers with their  
 958 detailed evaluations and scores are presented in subsequent tables: Qwen2-VL-72B-Instruct in Table 7,  
 959 Qwen2-VL-7B-Instruct in Table 6, GPT-4o in Table 8, GPT-4o-mini in Table 9, Gemini-2.0-Flash in  
 Table 10 and Gemini-2.0-Flash-Lite in Table 11.

Table 5: **Qualitative result of G-Eval evaluation using GPT-4o as evaluator model.** The evaluator input prompt consists of a preamble, a question and an expert answer. The first two analysis steps of the G-Eval evaluation, which consist of the comprehension of the question and the expert answer, are executed offline only once. The next analysis steps and the final evaluation scores are computed for each tested model separately in Tables 6 to 11.

Sample ID – val_Embrapa_20	
<i>Preamble</i>	You will be given an open-ended question about Peanut, specifically focused on Diseases. Respond fully and clearly to the question, avoiding unrelated information.
<i>Question</i>	Is there a relationship between the occurrence of thrips and the incidence of scab in peanut plants?
<i>Expert answer</i>	The higher incidence of scab may be associated with the presence of thrips in the crop, probably due to the fact that the wounds caused by these insects favor the penetration of the pathogen that causes the disease.
Evaluator answer	
<i>G-Eval steps</i>	
<i>1. Question analysis</i>	The question asks about the relationship between thrips and scab in peanut plants. It seeks to understand if there is a connection between the presence of thrips and the incidence of scab disease in peanuts.
<i>2. Expert answer analysis</i>	The expert answer suggests that there is a relationship between thrips and scab. It explains that thrips may cause wounds on the plants, which can facilitate the entry of the pathogen responsible for scab, indicating an indirect relationship.

Table 6: **Qualitative result of Qwen2-VL-7B-Instruct answer on CABBAGE sample val\_Embrapa\_20 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 5, followed by a step-by-step G-Eval analysis and scoring. The analysis includes correctness, specificity, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

Qwen2-VL-7B-Instruct answer – val_Embrapa_20	
No, there is no relationship between the occurrence of thrips and the incidence of scab in peanut plants.	
Evaluator answer	
<i>G-Eval steps</i>	
<i>3. Model answer analysis</i>	The model’s answer states that there is no relationship between thrips and scab in peanut plants, directly contradicting the expert answer.
<i>4. Correctness analysis</i>	The model’s answer does not address the key point made by the expert, which is that thrips can facilitate the entry of the scab pathogen through the wounds they create. Therefore, the model’s answer is incorrect in relation to the expert answer.
<i>5. Specificity analysis</i>	The model’s answer lacks any agriculture-specific knowledge. It provides a general statement without any explanation or depth related to the interaction between thrips and scab in peanuts, unlike the expert answer, which provides a specific mechanism of interaction.
<i>6. Conciseness analysis</i>	The model’s answer is concise, but it is overly simplistic and lacks the necessary information to be meaningful. The expert answer provides a concise explanation that includes the reason for the relationship, which is missing in the model’s response.
<i>7. Final review</i>	The model’s answer is incorrect as it contradicts the expert’s explanation of the relationship between thrips and scab in peanut plants. It lacks specificity, as it does not provide any agriculture-specific details or mechanisms that are present in the expert answer. While the model’s answer is concise, it is overly simplistic and does not convey the necessary information, resulting in a low conciseness score. Overall, the model’s response fails to meet the criteria for correctness, specificity, and conciseness.
<i>G-Eval score</i>	{correctness: 1.11, specificity: 1.24, conciseness: 2.68}

960

Table 7: **Qualitative result of Qwen2-VL-72B-Instruct answer on CABBAGE sample val.Embrapa.20 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 5, followed by a step-by-step G-Eval analysis and scoring. The analysis includes correctness, specificity, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

Qwen2-VL-72B-Instruct answer – val.Embrapa.20	
Yes, there is a relationship between the occurrence of thrips and the incidence of scab in peanut plants. Thrips can cause damage to the plant's leaves and stems, creating wounds that allow pathogens such as the fungus that causes scab to enter the plant. Additionally, thrips can spread fungal spores from plant to plant, increasing the risk of scab infection. Therefore, the presence of thrips can contribute to the incidence of scab in peanut plants.	
G-Eval steps	Evaluator answer
3. <i>Model answer analysis</i>	The model's answer confirms that there is a relationship between thrips and scab in peanut plants. It explains that thrips cause damage that allows pathogens to enter and also mentions that thrips can spread fungal spores, increasing scab risk. This answer provides additional detail about thrips spreading spores.
4. <i>Correctness analysis</i>	The model's answer correctly identifies the relationship between thrips and scab, similar to the expert answer. It includes the key point about thrips causing damage that allows pathogen entry and adds the detail about thrips spreading spores, which is not in the expert answer but is relevant. Therefore, the model's answer is largely correct with minor additional information not covered by the expert answer.
5. <i>Specificity analysis</i>	The model's answer demonstrates a good level of agricultural specificity by discussing how thrips damage can lead to pathogen entry and mentioning the spread of fungal spores. This shows a moderate to high level of agricultural knowledge, similar to the expert answer, with some additional detail.
6. <i>Conciseness analysis</i>	The model's answer is slightly more detailed than the expert answer, providing additional information about the spread of fungal spores. While this adds value, it makes the answer slightly less concise than the expert's. However, the model's answer remains clear and focused on the topic.
7. <i>Final review</i>	The model's answer is largely correct, providing a thorough explanation of the relationship between thrips and scab, including additional detail about the spread of fungal spores. It demonstrates a good level of agricultural specificity, similar to the expert answer. However, the additional detail makes the answer slightly less concise compared to the expert's more straightforward response. Overall, the model's answer is informative and relevant, earning high marks for correctness and specificity, with a moderate score for conciseness due to the extra detail.
G-Eval score	{correctness: 4.32, specificity: 3.99, conciseness: 3.38}

Table 8: **Qualitative result of GPT-4o answer on CABBAGE sample val.Embrapa.20 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 5, followed by a step-by-step G-Eval analysis and scoring. The analysis includes correctness, specificity, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

GPT-4o answer – val.Embrapa.20	
There is no direct relationship between the occurrence of thrips and the incidence of scab in peanut plants, as they are caused by different pathogens. Thrips are insects that can transmit viruses like the Tomato Spotted Wilt Virus (TSWV), while scab is a fungal disease caused by <i>Sphaceloma arachidis</i> . However, thrips can weaken plants by feeding on them and spreading viruses, potentially making them more susceptible to other diseases, including scab. Additionally, environmental conditions favorable for thrips might also promote scab, leading to an indirect association.	
G-Eval steps	Evaluator answer
3. <i>Model answer analysis</i>	The model's answer states that there is no direct relationship between thrips and scab because they are caused by different pathogens. However, it acknowledges that thrips can weaken plants, potentially making them more susceptible to scab, and mentions that environmental conditions favorable to thrips might also promote scab, suggesting an indirect association.
4. <i>Correctness analysis</i>	The model's answer partially addresses the expert's points. It correctly identifies that thrips and scab are caused by different pathogens, but it also acknowledges an indirect relationship through plant weakening and environmental conditions, similar to the expert's explanation of wounds facilitating pathogen entry. However, it does not explicitly mention the role of wounds caused by thrips, which is a key point in the expert answer.
5. <i>Specificity analysis</i>	The model's answer demonstrates moderate agricultural specificity. It accurately identifies the pathogens involved and explains the indirect relationship through plant weakening and environmental conditions, showing a good understanding of the agricultural context, but it lacks the specific detail about wounds facilitating pathogen entry.
6. <i>Conciseness analysis</i>	The model's answer is somewhat verbose. It provides additional information about the pathogens and potential indirect relationships, which, while informative, is not as concise as the expert answer. The expert answer is more focused and directly addresses the question with fewer words.
7. <i>Final review</i>	The model's answer is largely correct, acknowledging an indirect relationship between thrips and scab, similar to the expert's explanation. However, it misses the specific mention of wounds facilitating pathogen entry, which is a key point in the expert answer. The model demonstrates a good level of agricultural specificity by identifying the pathogens and explaining potential indirect relationships, although it could be more precise in highlighting the role of wounds. The response is somewhat verbose compared to the expert answer, as it includes additional details that, while informative, are not directly necessary to answer the question. Overall, the model's answer is informative and mostly aligns with the expert's explanation, but it could improve in conciseness and by explicitly mentioning the role of wounds.
G-Eval score	{correctness: 3.32, specificity: 3.49, conciseness: 2.91}

Table 9: **Qualitative result of GPT-4o-mini answer on CABBAGE sample val\_Embrapa\_20 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 5, followed by a step-by-step G-Eval analysis and scoring. The analysis includes correctness, specificity, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

GPT-4o-mini answer – val_Embrapa_20	
Yes, there is a relationship between the occurrence of thrips and the incidence of scab in peanut plants. Thrips are known to cause direct damage to the plants by feeding on the leaves, which can lead to increased stress on the plants. This stress can make peanut plants more susceptible to diseases, including scab, which is caused by the fungus <i>Fusarium</i> spp. Although thrips do not directly cause scab, their feeding can weaken the plants, reducing their ability to resist fungal infections. Therefore, the presence of thrips can indirectly increase the incidence of scab in peanut plants.	
G-Eval steps	Evaluator answer
3. <i>Model answer analysis</i>	The model's answer acknowledges a relationship between thrips and scab, explaining that thrips cause damage by feeding on leaves, which stresses the plants and makes them more susceptible to diseases like scab. It specifies that thrips do not directly cause scab but can weaken plants, increasing their susceptibility to fungal infections.
4. <i>Correctness analysis</i>	The model's answer correctly identifies a relationship between thrips and scab, similar to the expert answer. However, it introduces the idea that thrips cause stress, which indirectly increases susceptibility to scab, rather than focusing on wounds facilitating pathogen entry. The mention of <i>*Fusarium* spp.</i> is incorrect, as scab is not caused by this fungus. Overall, the model's answer is mostly correct but includes a factual error and a different emphasis than the expert answer.
5. <i>Specificity analysis</i>	The model's answer demonstrates moderate agricultural specificity by discussing the impact of thrips on plant stress and susceptibility to diseases, but it inaccurately attributes scab to <i>Fusarium</i> spp., which detracts from its specificity. The expert answer is more precise in linking thrips wounds to pathogen entry.
6. <i>Conciseness analysis</i>	The model's answer is concise, providing a clear explanation of the relationship between thrips and scab. It includes some additional context about plant stress, which is somewhat relevant but not directly aligned with the expert's focus on wounds. The answer could be slightly more focused on the key point about wounds facilitating pathogen entry.
7. <i>Final review</i>	The model's answer acknowledges the relationship between thrips and scab, similar to the expert's explanation, but introduces an incorrect pathogen and a different emphasis on plant stress. It demonstrates moderate agricultural specificity but inaccurately identifies the pathogen. The answer is concise and mostly focused, though it includes some additional context that is not directly aligned with the expert's focus. Overall, the model's answer is mostly correct but includes a factual error and could be more aligned with the expert's explanation.
G-Eval score	{correctness: 3.00, specificity: 2.99, conciseness: 3.95}

Table 10: **Qualitative result of Gemini-2.0-Flash answer on CABBAGE sample val\_Embrapa\_20 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 5, followed by a step-by-step G-Eval analysis and scoring. The analysis includes correctness, specificity, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

Gemini-2.0-Flash answer – val_Embrapa_20	
While thrips can vector some plant diseases and their feeding damage can create entry points for pathogens, there is no direct relationship established between thrips occurrence and the incidence of scab in peanut plants. Scab is caused by the fungus <i>Sphaeloma arachidis</i> , and its development is favored by moist conditions, unlike thrips which prefer warm, dry weather. Therefore, thrips are not considered a direct factor in the development of peanut scab.	
G-Eval steps	Evaluator answer
3. <i>Model answer analysis</i>	The model's answer states that there is no direct relationship between thrips and scab in peanuts. It explains that while thrips can create entry points for pathogens, scab is caused by a fungus and thrives in moist conditions, unlike thrips which prefer dry conditions. Therefore, thrips are not considered a direct factor in scab development.
4. <i>Correctness analysis</i>	The model's answer correctly identifies that there is no direct relationship between thrips and scab, which aligns with the expert's implication of an indirect relationship. However, it contradicts the expert's suggestion of a possible indirect relationship through wound facilitation. The model provides additional context about the environmental conditions favoring scab and thrips, which is not mentioned in the expert answer.
5. <i>Specificity analysis</i>	The model's answer demonstrates moderate agricultural specificity. It identifies the causal agent of scab and the environmental conditions favoring both scab and thrips. However, it lacks the expert's nuance about the potential indirect relationship via wounds caused by thrips.
6. <i>Conciseness analysis</i>	The model's answer is concise and focuses on the key points: the lack of a direct relationship and the environmental conditions for scab and thrips. It provides relevant information without unnecessary detail, making it efficient.
7. <i>Final review</i>	The model's answer provides a mostly correct response by identifying the lack of a direct relationship between thrips and scab, aligning with the expert's implication of an indirect relationship. However, it contradicts the expert's suggestion about the potential for thrips to facilitate pathogen entry through wounds. The model demonstrates moderate specificity by discussing the causal agent and environmental conditions but lacks the expert's nuance about thrips' indirect role. The response is concise and efficiently communicates the key points. Overall, the model's answer is mostly correct but could improve in aligning with the expert's nuanced understanding of the relationship.
G-Eval score	{correctness: 2.92, specificity: 3.21, conciseness: 4.17}

Table 11: **Qualitative result of Gemini-2.0-Flash-Lite answer on CABBAGE sample val Embrapa\_20 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 5, followed by a step-by-step G-Eval analysis and scoring. The analysis includes correctness, specificity, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

Gemini-2.0-Flash-Lite answer – val.Embrapa_20	
Yes, there is a relationship between thrips and the incidence of scab in peanut plants. Thrips can act as vectors for the fungal pathogen that causes scab, spreading the disease from plant to plant.	
G-Eval steps	Evaluator answer
3. <i>Model answer analysis</i>	The model’s answer states that there is a relationship between thrips and scab, suggesting that thrips act as vectors for the pathogen, spreading the disease. This implies a direct role of thrips in transmitting the pathogen rather than just facilitating its entry.
4. <i>Correctness analysis</i>	The model’s answer partially addresses the expert’s key point. While it correctly identifies a relationship, it incorrectly suggests that thrips act as vectors for the pathogen, which is not mentioned by the expert. The expert describes thrips as facilitating pathogen entry due to the wounds they cause, not as vectors. Therefore, the model’s answer is partially correct but contains inaccuracies about the nature of the relationship.
5. <i>Specificity analysis</i>	The model’s answer shows some agriculture-specific knowledge by identifying a relationship between thrips and scab, but it inaccurately describes thrips as vectors, which is not supported by the expert answer. This indicates a moderate level of specificity with some incorrect details.
6. <i>Conciseness analysis</i>	The model’s answer is concise, directly stating the relationship between thrips and scab. However, it lacks the nuanced detail about how thrips contribute to scab incidence through wounding, as mentioned by the expert, which affects the meaningfulness of its conciseness.
7. <i>Final review</i>	The model’s answer correctly identifies a relationship between thrips and scab in peanut plants but inaccurately describes thrips as vectors for the pathogen, which is not supported by the expert answer. This results in a partially correct response with moderate agricultural specificity. The answer is concise but lacks the depth and accuracy of the expert’s explanation about the role of thrips in facilitating pathogen entry through wounding. Overall, the model’s response is partially correct and moderately specific, but it could be improved by accurately reflecting the expert’s explanation.
G-Eval score	{correctness: 2.42, specificity: 2.75, conciseness: 3.27}

961 **C.3 Supplementary Analyses**

962 We conduct additional experiments to assess the reliability of non-deterministic evaluation protocols—  
 963 specifically G-Eval—in the context of scientific question answering in agriculture. These analyses are  
 964 performed on a random sampled subset of 100 question–answer pairs from the Embrapa dataset [13].

965 **On the Evaluation Efficacy with Single- vs. Multi-Metric Prompting.** We begin by ob-  
 966 serving that automatic evaluators often produce scores that are correlated across distinct evalu-  
 967 ation criteria. To quantify these dependencies, we compute Kendall’s  $\tau$  and Spearman’s  $\rho$  cor-  
 968 relations between the *Correctness*, *Specificity*, and *Conciseness* scores generated by the evalu-  
 969 ator. Interestingly, such inter-metric correlations are not unique to automated evaluation: hu-  
 970 man judgments on complex generation tasks frequently exhibit similar dependencies. For ex-  
 971 ample, in the SummEval benchmark [15], human ratings for metrics such as *relevance*, *co-*  
 972 *herence*, and *consistency* display moderate to strong rank correlations (see Table 12), indi-  
 973 cating that even expert annotators do not fully disentangle different evaluation dimensions.

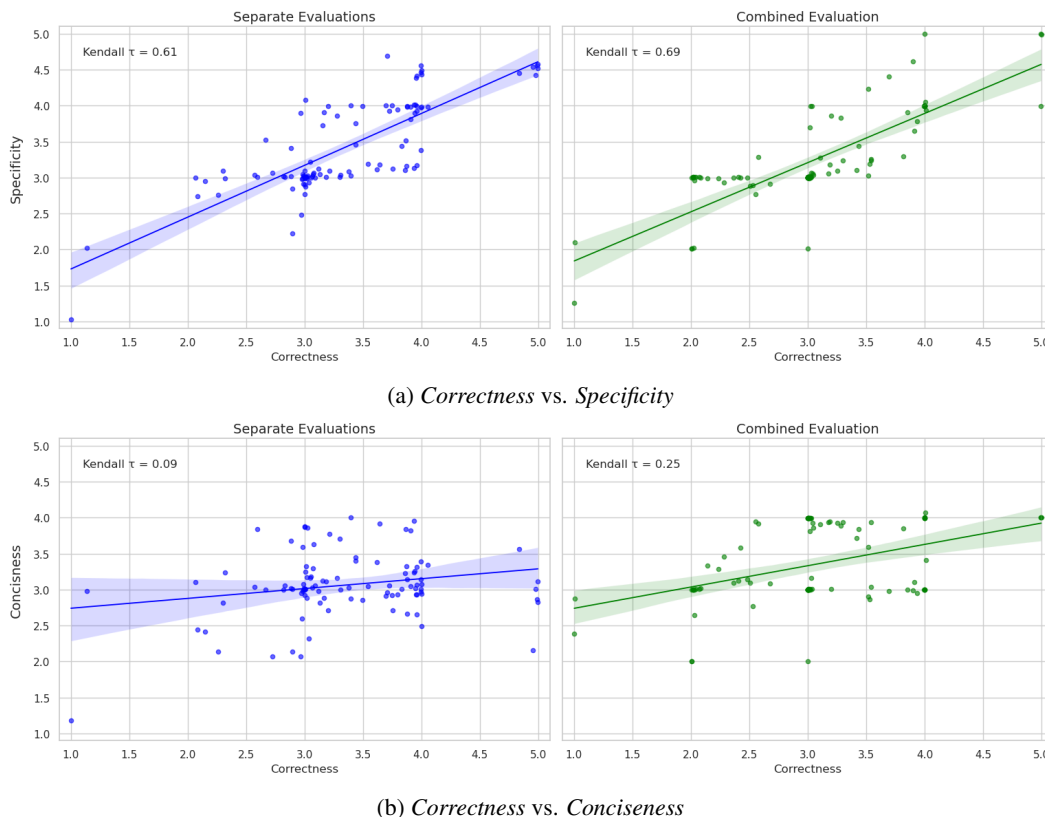
974  
 975 Motivated by these observations, we hy-  
 976 pothesize that the design of the evalua-  
 977 tion prompt may influence the degree of  
 978 inter-metric correlation. In particular, we  
 979 investigate whether prompting the eval-  
 980 uator model to assess multiple metrics  
 981 simultaneously—as opposed to issuing sep-  
 982 arate prompts per metric—affects the struc-  
 983 ture of metric dependencies. This approach  
 984 generalizes the original G-Eval protocol,  
 985 which evaluates a single metric per prompt,  
 986 by enabling multi-metric evaluation in a single forward pass, potentially reducing cost.

Table 12: Pairwise metric correlations from human judgments of model answers in the SummEval dataset [15]. Each cell reports the correlation coefficient with the associated  $p$ -value in parentheses. Spearman’s  $\rho$  above the diagonal and Kendall’s  $\tau$  below.

	Relevance	Coherence	Fluency	Consistency
Relevance	–	0.654 (3e-96)	0.277 (1e-29)	0.322 (5e-40)
Coherence	0.527 (1e-78)	–	0.330 (4e-42)	0.302 (4e-35)
Fluency	0.229 (5e-29)	0.270 (9e-41)	–	0.415 (1e-67)
Consistency	0.270 (1e-38)	0.249 (7e-34)	0.381 (3e-63)	–

Above the diagonal: Spearman’s  $\rho$ . Below the diagonal: Kendall’s  $\tau$ .

987 To test this hypothesis, we compare the correlation of the evaluator scores under two prompting  
 988 formats: (1) *single-metric prompting*, in which each metric is elicited independently, and (2) *multi-*  
 989 *metric prompting*, in which all three metrics are requested simultaneously (see Appendix C.2). As



**Figure 8: Impact of Prompt Format on Inter-Metric Score Correlation.** Comparison of evaluator score correlations under two prompting strategies: separate metric prompts (left) and simultaneous multi-metric prompts (right). Scatter plots report Kendall’s  $\tau$  rank correlation between *Correctness* and (a) *Specificity*, (b) *Conciseness*, over a subset of Embrapa QA pairs. Results show that joint evaluation via multi-metric prompting slightly increases inter-metric dependency, particularly between *Correctness* and *Specificity* ( $\tau = 0.69$  vs.  $\tau = 0.61$ ). However, the increase in correlation is less pronounced between *Correctness* and *Conciseness*, suggesting that metric interactions vary depending on semantic proximity. These findings indicate that single-metric prompting may better preserve metric independence, albeit at a higher evaluation cost.

990 the evaluator model, we employ GPT-4o with temperature = 0, and we use GPT-4o-mini as the  
 991 answering model, with temperature = 0.2, as it exhibits the highest score variance across the selected  
 992 Embrapa samples. Each sample is evaluated over  $n = 5$  trials, and scores are averaged across runs.

993 Our experiments reveal that single-metric prompting slightly mitigates correlations—*Correctness*  
 994 and *Specificity*, and particularly between these two and *Conciseness*—suggesting reduced evaluator  
 995 bias; see Figure 8. However, the improvements are modest, and the additional computational cost of  
 996 separate prompts is not clearly justified when weighed against the efficiency gains of multi-metric  
 997 prompting. Indeed, from Table 13 we observe that the token utilisation (obtained from OpenAI  
 998 Dashboard [39]) of the *multi-metric prompting* is much more cost-effective.

Table 13: Correlation between evaluation metrics under single- and multi-metric prompting formats. Kendall’s  $\tau$  and Spearman’s  $\rho$  are reported along with their  $p$ -values in parentheses. Token counts indicate the total number of input and output tokens consumed per evaluation format.

Metric Pair	Single-Metric Prompt					Multi-Metric Prompt				
	Correlation		Total Tokens			Correlation		Total Tokens		
	$\tau$	$\rho$	Input	Output	Cache	$\tau$	$\rho$	Input	Output	Cache
Corr vs. Spec	0.61 (2e-19)	0.79 (2e-23)				0.69 (4e-24)	0.86 (1e-30)			
Corr vs. Conc	0.09 (2e-1)	0.13 (2e-1)	197K	460K	113K	0.25 (2e-4)	0.36 (2e-4)	92K	261K	64K
Spec vs. Conc	-0.02 (7e-1)	-0.03 (8e-1)				0.19 (4e-3)	0.28 (4e-3)			

999 **On the Bias of Evaluators Toward Answer Length in Judging Correctness.**  
 1000 We begin by observing that our evaluator  
 1001 exhibits a bias in judging the correctness of  
 1002 answers, favoring more verbose responses.  
 1003 To investigate whether this phenomenon  
 1004 also occurs in human evaluation, we con-  
 1005 ducted a comparative analysis and found  
 1006 that human judgment similarly tends to fa-  
 1007 vor longer answers—even in settings where brevity is typically preferred, such as text summarization  
 1008 (see Table 14).  
 1009

1010 To quantify this bias, we constructed a synthetic dataset of augmented answers by rephrasing,  
 1011 summarizing, and expanding the original content without altering its semantic meaning. Start-  
 1012 ing from the original answers produced by GPT-4o-mini, we used GPT-4.1 with a temperature  
 1013 setting of temperature = 0 to generate the augmented versions, following the prompting scheme  
 1014 shown in Listing 4. We ask to limit the word utilization setting the two thresholds (`min_words`,  
 1015 `max_words`) to (90%, 110%), (40%, 60%) and (190%, 210%) of the original answer length, for  
 1016 *rephrase*, *summarize*, and *expand*, respectively. On average, the rephrased answers were 92% the  
 1017 length of the originals, the summarized versions were 50%, and the expanded responses reached 236%.  
 1018

1019 Subsequently, we prompted GPT-4o to independently evalu-  
 1020 ate the *Correctness* of the original and each set of aug-  
 1021 mented answers. The results show that, on average, cor-  
 1022 rectness scores for the augmented answers remain closely  
 1023 aligned with those of the original responses, suggesting  
 1024 that the evaluator is generally robust to these forms of  
 1025 semantic-preserving transformation. As illustrated in Fig-  
 1026 ure 9, the median difference in correctness scores is near  
 1027 zero (indicated by the red horizontal lines). However, we  
 1028 also observe that approximately 50% of the mass shifts  
 1029 vertically in the case of expanded answers (highlighted  
 1030 by the green boxes), indicating a systematic increase in  
 1031 their evaluation scores. Conversely, summarized answers  
 1032 tend to receive lower scores, implying a penalization for  
 1033 brevity.

1034 These results suggest that while the evaluator maintains  
 1035 robustness under semantic-preserving perturbations, it nev-  
 1036 ertheless exhibits a measurable preference for longer answers. This length bias, even when subtle,  
 1037 has potential implications for both model evaluation protocols and the design of prompting strategies,  
 1038 particularly in tasks where brevity is essential.

1039 **On the Self-Judging Bias: G-Eval Arena.** To probe potential evaluator biases inherent to the G-  
 1040 Eval framework, we investigate whether models exhibit preferential scoring behaviour when assessing  
 1041 responses generated by themselves. Building on the approach introduced in *Judging LLM-as-a-Judge*  
 1042 *with MT-Bench and Chatbot Arena* by Zheng et al. [63], we design an evaluation arena where each  
 1043 model—GPT-4o, GPT-4o-mini, Gemini 2 Flash, and Gemini 2 Flash Lite—acts as a judge over  
 1044 anonymized responses produced by all models, including itself. This setup allows us to assess  
 1045 the extent of self-judging bias and its variability across models of different sizes and architectures.  
 1046 The evaluation is conducted under the same experimental setting as the previous experiments using  
 1047 *multi-metric prompting* for G-Eval and the results are shown in Table 15.

1048 To further explore the extent of self-judging tendencies under the G-Eval Arena setup, we compute the  
 1049 *subjectivity bias* for each model. This metric captures the deviation between a model’s self-assigned  
 1050 score and the average score it receives from all other  $N - 1$  judges. Formally, for a model  $M$ , the  
 1051 subjectivity bias is defined as:

$$\text{Bias}(M) = \text{Judge}(M \rightarrow M) - \frac{1}{N-1} \sum_{J \neq M} \text{Judge}(J \rightarrow M)$$

Table 14: Correlation between summary length and human evaluation metrics in the SummEval dataset [15]. Each cell reports the correlation coefficient with its associated  $p$ -value in parentheses.

Metric	Relevance	Coherence	Fluency	Consistency
Spearman’s $\rho$	0.24 (2e-22)	0.07 (2e-03)	0.01 (6e-01)	0.11 (9e-06)
Kendall’s $\tau$	0.17 (1e-22)	0.05 (1e-03)	0.01 (6e-01)	0.08 (1e-05)

Correlation between summary length and human judgment scores.

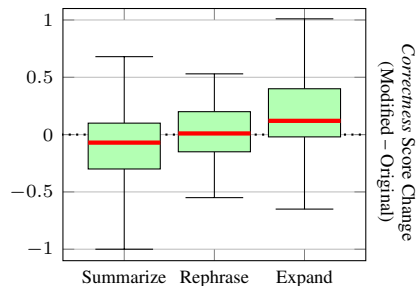


Figure 9: **Evaluator Bias Towards Answer Length.** Boxplots show the distribution of GPT-4o evaluated answer *Correctness* score differences for augmented (summarized, rephrased, expanded) and original answers.

Table 15: **G-Eval arena scores.** Self- and cross-model judgments on a subset of 100 Embrapa QA pairs. Maximum per row is highlighted in blue. Self-judgments are highlighted in bold.

Correctness					Specificity					Conciseness				
Judge →	4o	4o-mini	G2F	G2F-L	Judge →	4o	4o-mini	G2F	G2F-L	Judge →	4o	4o-mini	G2F	G2F-L
4o	<b>3.26</b>	2.96	3.22	3.06	4o	<b>3.38</b>	3.05	3.37	3.14	4o	<b>3.16</b>	3.24	3.12	3.11
4o-mini	3.24	<b>3.12</b>	3.32	3.08	4o-mini	3.39	<b>3.27</b>	<b>3.47</b>	3.18	4o-mini	3.36	<b>3.46</b>	3.37	3.31
G2F	3.45	3.33	<b>3.55</b>	3.36	G2F	3.21	3.03	<b>3.22</b>	3.07	G2F	2.79	3.03	<b>2.82</b>	3.12
G2F-L	3.56	3.41	3.75	<b>3.53</b>	G2F-L	3.38	3.10	3.54	<b>3.22</b>	G2F-L	2.99	3.10	2.98	<b>3.25</b>

1052 where  $N$  is the total number of models participating in the evaluation. A positive value indicates a  
 1053 self-favouring bias, while a negative value suggests that the model underrates itself compared to peer  
 1054 evaluations.

1055 Table 16 reports the computed biases across the  
 1056 three evaluation metrics: *Correctness*, *Speci-*  
 1057 *ficity*, and *Conciseness*. We observe that Gem-  
 1058 ini 2 Flash Lite (G2F-L) exhibits the largest  
 1059 positive bias in *Correctness* (+0.363), indicat-  
 1060 ing a pronounced tendency to overrate its own  
 1061 answers. Conversely, Gemini 2 Flash (G2F)  
 1062 tends to underrate its own outputs, particularly in  
 1063 *Specificity* (-0.240) and *Conciseness* (-0.337).  
 1064 Interestingly, GPT-4o-mini (4o-mini) shows a  
 1065 strong self-preference in *Conciseness* (+0.337),  
 while GPT-4o (4o) remains comparatively balanced across all metrics, with only minor deviations.

Table 16: **Subjectivity Bias per Model.** Bias is computed as the difference between self-judgment and the average judgment from other models, across three metrics. The last column reports the aggregate bias magnitude as the absolute sum across all metrics.

Model	Correctness	Specificity	Conciseness	Aggregate Bias ( )
4o	-0.157	+0.053	+0.113	<b>0.323</b>
4o-mini	-0.113	+0.210	+0.337	0.660
G2F	+0.120	-0.240	-0.337	0.697
G2F-L	+0.363	+0.090	+0.070	0.523

```
# ----- REPHRASE PROMPT -----
You are an expert in agronomy.
Your task is to paraphrase the answer below:
1. Retain all key agronomic terms and concepts (e.g., crop physiology, soil management, pest control).
2. You may reorder phrases or sentences, while staying true to the original meaning and factual content.
3. Your paraphrase should closely match the original answer's length, falling between {min_words} and {max_words} words.

Respond with only the rephrased answer--no commentary or labels.
Answer: {answer}

# ----- SUMMARIZE PROMPT -----
You are an expert in agronomy.
Your task is to summarize the answer below:
1. Preserve all key agronomic terms and concepts (e.g., crop physiology, soil management, pest control).
2. You may reorganize or rephrase for clarity, while staying true to the original meaning and factual content.
3. Your summary should be approximately half the length of the original, falling between {min_words} and {max_words} words.

Respond with only the summary--no commentary or labels.
Answer: {answer}

# ----- EXPAND PROMPT -----
You are an expert in agronomy.
Your task is to elaborate on the answer below:
1. Retain all core agricultural facts, terms, and concepts (e.g., crop physiology, soil management, pest control).
2. You may expand for clarity and style (e.g., add transitions or background phrasing), while staying true to the original meaning and factual content.
3. Do not introduce new factual claims, mechanisms, or relationships.
4. Your elaboration should be approximately twice the length of the original, falling between {min_words} and {max_words} words.

Respond with only the elaborated answer--no commentary or labels.
Answer: {answer}
```

Listing 4: **Prompts for guided augmentation of model-generated answers.** Prompts for paraphrasing, summarization, and elaboration-based augmentation of model-generated answers in agronomic tasks.

1067 **D Other Results**

1068 In this section, we present additional results. In Table 17 we present complete quantitative results  
 1069 for the tested models over the CABBAGE benchmark. All metrics—deterministic (Accuracy, Ex-  
 1070 act Match) and non-deterministic (G-Eval scores)—are computed over  $n = 5$  independent runs  
 1071 (temperature  $t = 0.2$ ) and reported as mean  $\pm 95\%$  confidence interval (i.e. mean  $\pm 1.96 \times$  SEM),  
 1072 assuming asymptotically normal per-run distributions. Empirically, SEM values are consistently  
 1073 small and negligible across all benchmarks, typically on the order of  $\pm 0.01$  for classification metrics  
 1074 (e.g., Accuracy, Exact Match) and  $\pm 0.2$  for open-ended metrics (e.g., G-Eval scores), supporting  
 1075 the reliability of reported means. We limit our report to *Correctness* and *Procedural Accuracy* for  
 1076 G-Eval scores only and difficulty level  $D5$  for EPPO. Additionally, we present qualitative results  
 1077 for the three key areas of the CABBAGE benchmark, specifically agricultural scientific knowledge  
 1078 in Appendix D.1, agronomic visual identification and cognition in Appendices D.2 and D.3 and  
 1079 agronomic procedural reasoning in Appendix D.4.

Table 17: **Unified zero-shot performance across CABBAGE.** Evaluation spans scientific knowledge (AgriExam, CCA-CEU, Embrapa), visual cognition (BPPQ, EPPO, PI@ntNet), and procedural reasoning (wikiHow). Metrics include multiple-choice *Accuracy* (Acc.), open-ended *Exact Match* (EM), and GPT-4o-guided evaluation using G-Eval framework for *Correctness* (Corr.), and *Procedural Accuracy* (P. Acc.).

Model	Agricultural Scientific Knowledge				Agronomic Visual Cognition				Agronomic Procedural Reasoning				
	AgriExam		CCA-CEU	Embrapa	BPPQ	EPPO		PI@ntNet		wikiHow			
	Acc.	Acc.	Corr.	Acc.	D5 Acc.	Open EM	Single EM	Multi EM	Arr. Acc.	Miss. Acc.	Next Acc.	All Steps P. Acc.	
GPT-4o	<b>0.77 ± 0.01</b>	0.89 ± 0.01	<b>3.28 ± 0.18</b>	<b>0.63 ± 0.02</b>	0.71 ± 0.02	<b>0.50 ± 0.02</b>	<b>0.52 ± 0.02</b>	<b>0.74 ± 0.01</b>	0.65 ± 0.02	<b>0.80 ± 0.02</b>	<b>0.68 ± 0.02</b>	3.75 ± 0.22	
GPT-4o-mini	0.61 ± 0.02	0.88 ± 0.01	2.96 ± 0.36	0.50 ± 0.02	0.66 ± 0.02	0.35 ± 0.02	0.18 ± 0.01	0.38 ± 0.02	0.58 ± 0.02	0.71 ± 0.02	0.61 ± 0.02	3.54 ± 0.33	
Gemini 2 F	<b>0.77 ± 0.01</b>	0.90 ± 0.01	3.24 ± 0.11	<b>0.63 ± 0.02</b>	<b>0.72 ± 0.01</b>	0.49 ± 0.02	<b>0.52 ± 0.02</b>	0.68 ± 0.02	0.71 ± 0.01	0.71 ± 0.01	0.66 ± 0.01	<b>3.79 ± 0.09</b>	
Gemini 2 FL	0.71 ± 0.02	<b>0.91 ± 0.01</b>	3.06 ± 0.27	<b>0.63 ± 0.02</b>	0.70 ± 0.02	0.38 ± 0.02	0.38 ± 0.02	0.65 ± 0.02	0.68 ± 0.02	0.76 ± 0.01	0.66 ± 0.01	3.75 ± 0.18	
Qwen-2-VL-72B	0.63 ± 0.02	0.82 ± 0.02	2.78 ± 0.31	0.56 ± 0.02	0.63 ± 0.02	0.20 ± 0.01	0.26 ± 0.01	0.46 ± 0.02	<b>0.86 ± 0.01</b>	0.60 ± 0.01	0.55 ± 0.02	2.54 ± 0.33	
Qwen-2-VL-7B	0.45 ± 0.02	0.66 ± 0.02	2.63 ± 0.42	0.44 ± 0.02	0.63 ± 0.02	0.22 ± 0.01	0.23 ± 0.01	0.44 ± 0.02	0.60 ± 0.02	0.70 ± 0.02	0.65 ± 0.02	2.01 ± 0.49	

1080 **D.1 Agricultural Scientific Knowledge**

1081 Qualitative results about agricultural scientific knowledge can be found in Appendix C.2.

1082 **D.2 Agronomic Visual Identification**

1083 This section presents a set of qualitative examples from the CABBAGE benchmark, focusing on  
 1084 the visual identification of plant species by multiple multimodal large language models (MLLMs).  
 1085 Each example demonstrates how different models approach the identification task using images  
 1086 of various plant organs (such as bark, flower, fruit, habit, and leaf). The models’ responses are  
 1087 analyzed alongside their explanations, highlighting the differences in their ability to identify plants  
 1088 based on visual information. The goal is to evaluate the models’ integration of advanced agricultural  
 1089 knowledge and their reasoning capabilities in accurately identifying plant species from visual cues.  
 1090 Figures 10 and 11 illustrate how model accuracy varies based on the available visual information.  
 1091 Here we show how GPT-4o improve the response when looking at the five organs simultaneously.

Figure 10: **Qualitative Comparison on Species Visual Identification Task by Multiple Organs.** GPT-4o successfully identify the plant species by attaining the global characteristics of the plant in the five organ images together, as motivated by the model. The ★ indicates the model’s reasoning. Models’ responses to single organs are shown in Figure 11.

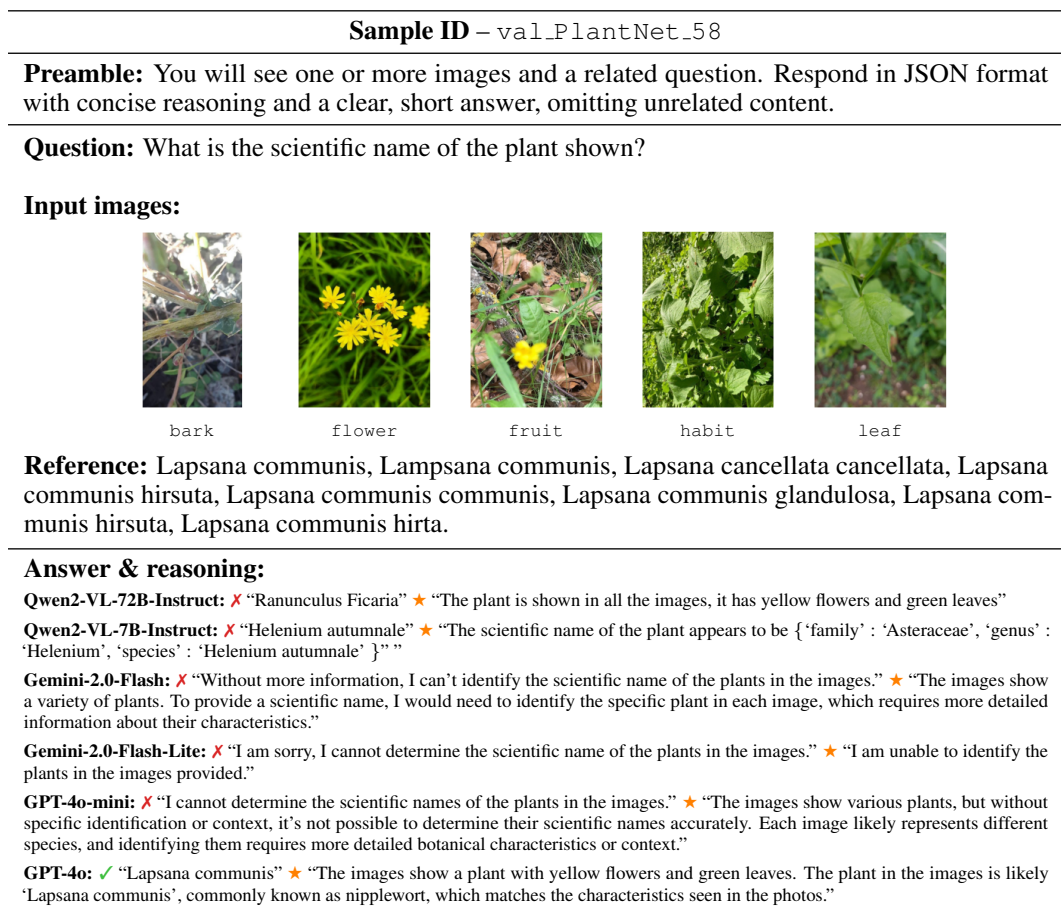




Figure 11: **Qualitative Comparison on Species Visual Identification Task by Single Organ.** The input preamble, question template and reference remain unchanged (see Figure 10). All the models fail to identify the plant species from the image of the plant’s organs individually. The ★ indicates the model’s explanation.

Input image	Answer & reasoning
 <p data-bbox="331 562 381 583">bark</p>	<p><b>Qwen2-VL-72B-Instruct:</b> ✗ “N/A” ★ “There is no plant in the image.”</p> <p><b>Qwen2-VL-7B-Instruct:</b> ✗ “Urtica dioica” ★ “The plant in the image is a type of plant that has been identified as a weed. The scientific name of the plant is Urtica dioica”</p> <p><b>Gemini-2.0-Flash:</b> ✗ “Arachis pintoi” ★ “Based on the leaves, stems, and buds in the image, the plant is likely Arachis pintoi”</p> <p><b>Gemini-2.0-Flash-Lite:</b> ✗ “I am sorry, I cannot identify the plant in the image.” ★ “I am unable to identify the plant in the image without more information”</p> <p><b>GPT-4o-mini:</b> ✗ “I don’t know the scientific name of the plant shown.” ★ “The image shows a plant, but without additional context or specific features, it’s difficult to accurately identify the scientific name. Identification often requires knowledge of the plant’s characteristics, habitat, and other details”</p> <p><b>GPT-4o:</b> ✗ “Arundinaria alpina” ★ “The plant in the image appears to be a type of bamboo, specifically Arundinaria alpina, which is commonly found in certain regions. The characteristics such as the woody stem and the growth pattern suggest this identification”</p>
 <p data-bbox="331 882 381 903">flower</p>	<p><b>Qwen2-VL-72B-Instruct:</b> ✗ “Crepis triasii” ★ “The image shows a plant with yellow flowers, which is identified as Crepis triasii.”</p> <p><b>Qwen2-VL-7B-Instruct:</b> ✗ “The plant is a yellow wildflower surrounded by grass.” ★ “The image shows a yellow wildflower that is surrounded by grass. However, the image is not clear enough to identify the exact species of the plant. The scientific name of the plant is not provided in the image.”</p> <p><b>Gemini-2.0-Flash:</b> ✗ “Youngia japonica” ★ “The image shows a plant with yellow flowers and dandelion-like leaves, which is characteristic of Youngia japonica, also known as Japanese hawksbeard.”</p> <p><b>Gemini-2.0-Flash-Lite:</b> ✗ “Hypochaeris radicata” ★ “The image shows a plant with yellow flowers and a rosette of leaves, which is consistent with the characteristics of Hypochaeris radicata, also known as catsear.”</p> <p><b>GPT-4o-mini:</b> ✗ “I don’t know the scientific name of the plant shown in the image.” ★ “The image shows yellow flowers, which could belong to various species. Without specific identification, I cannot provide a scientific name.”</p> <p><b>GPT-4o:</b> ✗ “Possibly Hieracium or Crepis (exact species unknown).” ★ “The plant in the image features small, bright yellow flowers with multiple petals, resembling those of the genus Hieracium or Crepis, which are commonly known as hawkweeds or hawksbeards. These plants are part of the Asteraceae family and are characterized by their dandelion-like flowers. Without more specific details, it’s challenging to determine the exact species, but Hieracium or Crepis are likely candidates.”</p>
 <p data-bbox="331 1228 381 1249">fruit</p>	<p><b>Qwen2-VL-72B-Instruct:</b> ✗ “Ranunculus” ★ “The plant is a buttercup, which has the scientific name Ranunculus.”</p> <p><b>Qwen2-VL-7B-Instruct:</b> ✗ “Ranunculus acris” ★ “The plant in the image is yellow and has a green stem. It appears to be a type of buttercup, which is commonly found in forests and meadows. Buttercups are flowering plants in the family Ranunculaceae, which includes many species with yellow flowers. The most common buttercup species in North America is Ranunculus acris, commonly known as meadow buttercup. Therefore, the scientific name of the plant shown in the image is Ranunculus acris.”</p> <p><b>Gemini-2.0-Flash:</b> ✗ “Ranunculus” ★ “Based on the yellow flower and surrounding foliage, it is likely a type of buttercup, which has the scientific name Ranunculus.”</p> <p><b>Gemini-2.0-Flash-Lite:</b> ✗ “I am unable to determine the scientific name of the plant in the image.” ★ “The image does not provide enough detail to identify the plant.”</p> <p><b>GPT-4o-mini:</b> ✗ “I can’t identify the plant or provide its scientific name.” ★ “The image shows a blurred view of some plants and flowers, making it difficult to identify them accurately. Without clear visibility or additional context, determining the scientific name is not possible.”</p> <p><b>GPT-4o:</b> ✗ “Cannot determine from image.” ★ “The image shows a yellow flower, which is likely a common wildflower. However, the image is out of focus, making it difficult to identify the plant accurately. Without clear visual details, it’s not possible to determine the scientific name.”</p>
 <p data-bbox="331 1533 381 1554">habit</p>	<p><b>Qwen2-VL-72B-Instruct:</b> ✗ “Lamium maculatum” ★ “The plant in the image has heart-shaped leaves with serrated edges, which is characteristic of the plant species Lamium maculatum, commonly known as spotted dead-nettle.”</p> <p><b>Qwen2-VL-7B-Instruct:</b> ✗ “Lamium purpureum” ★ “The plant shown is identified as ‘Lamium purpureum’.”</p> <p><b>Gemini-2.0-Flash:</b> ✗ “Alliaria petiolata” ★ “The plant in the image has broad, toothed leaves and small white flowers. These are characteristics of Alliaria petiolata, commonly known as garlic mustard.”</p> <p><b>Gemini-2.0-Flash-Lite:</b> ✗ “Alliaria petiolata” ★ “The image shows Garlic Mustard (Alliaria petiolata).”</p> <p><b>GPT-4o-mini:</b> ✗ “I cannot determine the scientific name of the plant shown.” ★ “The plant in the image appears to have broad leaves and a bushy growth habit, which could suggest a variety of species. However, without specific identifying features, it’s difficult to determine the exact scientific name. Common plants with similar characteristics include various species of wildflowers or herbs. More context or detail would be needed for accurate identification.”</p> <p><b>GPT-4o:</b> ✗ “Alliaria petiolata” ★ “The plant in the image has large, jagged leaves and small white flowers, which are characteristics of the plant known as garlic mustard. Garlic mustard’s scientific name is Alliaria petiolata.”</p>
 <p data-bbox="331 1831 381 1852">leaf</p>	<p><b>Qwen2-VL-72B-Instruct:</b> ✗ “Passiflora edulis” ★ “The plant shown in the image has heart-shaped leaves and a climbing habit, which is characteristic of the plant species Passiflora edulis, commonly known as passion fruit.”</p> <p><b>Qwen2-VL-7B-Instruct:</b> ✗ “Ipomoea batatas” ★ “The plant shown in the image is a common plant that is often used for its leaves. The scientific name of this plant is ‘Ipomoea batatas’, which is commonly known as sweet potato.”</p> <p><b>Gemini-2.0-Flash:</b> ✗ “Ageratina adenophora” ★ “The plant in the image has a single, green leaf with serrated edges and prominent veins. Based on these characteristics, it closely resembles a plant in the genus Ageratina. Ageratina adenophora is a common species in this genus.”</p> <p><b>Gemini-2.0-Flash-Lite:</b> ✗ “Eupatorium” ★ “Based on the image, the plant appears to be a type of Eupatorium, possibly Eupatorium capillifolium, due to the leaf shape and overall appearance.”</p> <p><b>GPT-4o-mini:</b> ✗ “I cannot determine the scientific name of the plant shown.” ★ “The image shows a leaf that resembles certain species of plants, but without more context or specific identifying features, it’s difficult to determine the exact scientific name. Common plants with similar leaves include those from the Amaranthaceae family, such as Chenopodium species, but this is speculative.”</p> <p><b>GPT-4o:</b> ✗ “Sida rhombifolia” ★ “The leaf in the image has a serrated edge and a pointed tip, which are characteristic features of the plant known as Sida rhombifolia. This plant is commonly found in tropical and subtropical regions and is often identified by its rhomboid-shaped leaves.”</p>

1092 **D.3 Agronomic Visual Cognition**

1093 In Figure 12 we present an example question from the BPPQ dataset to clarify the exact prompt  
 1094 format used for model evaluation on agronomic visual cognition.

---

**Sample ID** – val\_BPPQ\_104


---

**Preamble:** You will be shown an image and asked a multiple-choice question about it. There are five options: A, B, C, D and E. Answer with the option letter only.

---

**Question:** What are these large black lumps of fungal material called?

- (A) Cleistothecia
- (B) Pycnidia
- (C) Sclerotia
- (D) Basidiospores
- (E) Don't know and don't want to guess



---

**Correct:** (C) Sclerotia

---

Figure 12: **Agronomic visual identification sample from BPPQ subset.** Model answers to multiple-choice questions are evaluated through *Accuracy* score and structured output answer parsing.

1095 **D.4 Agronomic Procedural Reasoning**

1096 We provide one example each of a multiple-choice and an open-ended procedural reasoning question  
 1097 drawn from wikiHow [52]. The first showcases a question from the *Missing Step* task (see Table 18),  
 1098 and the second presents a question from the *All Steps* task, in which the model is asked to generate  
 1099 step-by-step instructions to carry out a procedure. The open-ended example is followed by model  
 1100 responses and detailed G-Eval evaluations. Specifically, Table 19 shows the input question, expert  
 1101 answer, and evaluator analysis (first two evaluation steps). The models' answers with their detailed  
 1102 evaluations and scores are presented in subsequent tables: Qwen2-VL-72B-Instruct in Table 21,  
 1103 Qwen2-VL-7B-Instruct in Table 20, GPT-4o in Table 22, GPT-4o-mini in Table 23, Gemini-2.0-Flash  
 in Table 24 and Gemini-2.0-Flash-Lite in Table 25.

Table 18: **Example of a wikiHow Missing Step question.** The table shows the question prompt, where the missing step is indicated by “???”, the set of candidate options, and the correct answer used for accuracy evaluation.

Prompt	Sample ID – test_wikiHow_502
<i>Question</i>	These steps describe how to Get Rid of Plant Mites following the "Using a Chemical Killer" method. What is the missing step to fill in the gap in the procedure? Steps: 1. Apply miticide to infested plants to effectively kill mites and prevent nesting. 2. ??? 3. Use a dormant oil in the winter to kill nesting adults. 4. Apply a summer oil during the growing season to kill active infestations.
<i>Options</i>	A. Prune affected plants to prevent the spread of disease. B. Apply a fungicide to prevent root rot. C. Use neem oil as a repellent to deter pests. D. Wash plants in an insecticidal soap if the miticide proves ineffective.
<i>Correct Answer</i>	D. Wash plants in an insecticidal soap if the miticide proves ineffective.

1104

Table 19: **Qualitative result of G-Eval evaluation using GPT-4o as evaluator model.** The evaluator input prompt consists of a preamble, a question and an expert answer. The first two analysis steps of the G-Eval evaluation, which consist of the comprehension of the question and the expert answer, are executed offline only once. The next analysis steps and the final evaluation scores are computed for each tested model separately in Tables 20 to 25.

<b>Sample ID – test.wikiHow.471</b>	
<i>Preamble</i>	You will be given an open-ended question about a procedure. Respond fully and clearly to the question, avoiding unrelated information.
<i>Question</i>	How can I Grow Baby Spinach? Please provide detailed step-by-step instructions for Planting Spinach in an Outdoor Garden.
<i>Expert answer</i>	<ol style="list-style-type: none"> <li>1. Start planting spinach in early spring.</li> <li>2. Grow spinach in a raised 1 foot (30 cm) planter or directly in the ground.</li> <li>3. Place spinach in partial sun.</li> <li>4. Plant seeds 1 inch (2.5 cm) apart in rows 4 inches (10 cm) apart.</li> <li>5. Thin the plants once they start to come in.</li> <li>6. Harvest your spinach after 40 days or so.</li> </ol>
<b>G-Eval steps</b>	<b>Evaluator answer</b>
<i>1. Question analysis</i>	The question asks for detailed step-by-step instructions for planting spinach in an outdoor garden. The intended outcome is to successfully grow baby spinach, highlighting the planting process, care, and harvesting in an outdoor setting. Key considerations include timing, location, planting method, and care instructions.
<i>2. Expert answer analysis</i>	The expert answer outlines a clear procedure: start planting in early spring, choose a raised planter or ground, ensure partial sun exposure, plant seeds with specific spacing, thin plants as they grow, and harvest after about 40 days. The steps are logically ordered and cover essential aspects of growing spinach.

Table 20: **Qualitative result of Qwen2-VL-7B-Instruct answer on CABBAGE sample test.wikiHow.471 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 19, followed by a step-by-step G-Eval analysis and scoring. The analysis includes procedural accuracy, procedural flow, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

<b>Qwen2-VL-7B-Instruct answer – test.wikiHow.471</b>	
Growing baby spinach in an outdoor garden requires careful planning and attention to detail. Here are the step-by-step instructions to guide you through the process.	
<b>G-Eval steps</b>	<b>Evaluator answer</b>
<i>3. Model answer analysis</i>	The model’s answer introduces the task of growing baby spinach but does not provide any actual steps or details. It lacks any procedural content or specific actions to be taken, merely stating that it requires planning and attention to detail.
<i>4. Procedural Accuracy analysis</i>	The model’s answer does not cover any of the essential steps outlined in the expert answer. It fails to mention timing, planting location, sun exposure, seed spacing, thinning, or harvesting. Therefore, it does not achieve the intended outcome of guiding someone to grow spinach.
<i>5. Procedural Flow analysis</i>	The model’s answer lacks any procedural flow since it does not provide steps. It is not organized in a sequence that could guide the reader through the process of growing spinach.
<i>6. Conciseness analysis</i>	The model’s answer is concise but lacks content. It does not provide any steps or details, so while it is not verbose, it fails to communicate necessary information effectively.
<i>7. Final review</i>	The model’s answer is inadequate as it does not provide any of the required steps or details necessary to grow baby spinach, as outlined in the expert answer. It lacks procedural accuracy, flow, and conciseness because it fails to include any substantive information or instructions.
<i>G-Eval score</i>	{procedural_accuracy: 1.16, procedural_flow: 1.25, conciseness: 2.37}

Table 21: **Qualitative result of Qwen2-VL-72B-Instruct answer on CABBAGE sample test\_wikiHow\_471 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 19, followed by a step-by-step G-Eval analysis and scoring. The analysis includes procedural accuracy, procedural flow, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

Qwen2-VL-72B-Instruct answer – test_wikiHow_471	
<p>To grow baby spinach in an outdoor garden, follow these detailed step-by-step instructions:</p> <ol style="list-style-type: none"> <li>1. Choose a suitable location: Select a spot in your garden that receives partial shade to full sun. Spinach prefers cooler temperatures and can tolerate some shade, especially during the hottest parts of the day.</li> <li>2. Prepare the soil: Spinach thrives in well-drained, nutrient-rich soil with a pH between 6.0 and 7.5. Amend the soil with compost or well-rotted manure to improve its fertility and structure. Ensure the soil is loose and free of large rocks or debris.</li> <li>3. Sow the seeds: Plant spinach seeds directly into the garden in early spring or late summer, depending on your climate. Sow the seeds about 1/2 inch deep and 1 to 2 inches apart in rows spaced 12 to 18 inches apart. Water the seeds gently after planting to ensure good soil contact.</li> <li>4. Provide adequate water: Keep the soil consistently moist but not waterlogged. Water the plants deeply and regularly, especially during dry spells. Mulching around the plants can help retain moisture and suppress weeds.</li> <li>5. Fertilize: Apply a balanced, slow-release fertilizer or a side dressing of compost around the plants once they have germinated. This will provide the necessary nutrients for healthy growth.</li> <li>6. Monitor for pests and diseases: Keep an eye out for common spinach pests such as aphids, flea beetles, and downy mildew. Use appropriate pest management techniques, such as insecticidal soap or neem oil, to control pests and diseases.</li> <li>7. Harvest: Baby spinach can be harvested when the leaves are 2 to 3 inches long. Use scissors or a sharp knife to cut the leaves close to the soil, leaving the plant's crown intact. This will allow the plant to continue growing and producing new leaves for subsequent harvests.</li> </ol> <p>By following these steps, you can successfully grow baby spinach in your outdoor garden and enjoy a bountiful harvest of fresh, tender leaves.</p>	
G-Eval steps	Evaluator answer
3. <i>Model answer analysis</i>	The model's answer is more detailed, covering additional aspects like soil preparation, watering, fertilizing, and pest management. It also includes specific details about soil pH, seed depth, and spacing, which are not mentioned in the expert answer. The model provides a more comprehensive guide to growing spinach.
4. <i>Procedural Accuracy analysis</i>	The model's answer includes all the essential steps mentioned in the expert answer, such as choosing the location, planting seeds, and harvesting. It also adds more detailed steps like soil preparation, watering, fertilizing, and pest management, which enhance the procedural accuracy by covering more aspects of growing spinach.
5. <i>Procedural Flow analysis</i>	The model's answer is logically structured, starting from location selection, soil preparation, planting, watering, fertilizing, pest management, and finally harvesting. This sequence is clear and follows a logical flow, although it includes more steps than the expert answer.
6. <i>Conciseness analysis</i>	The model's answer is more verbose than the expert answer, with additional information on soil preparation, watering, fertilizing, and pest management. While these details contribute to a comprehensive guide, they make the answer less concise compared to the expert's concise instructions.
7. <i>Final review</i>	The model's answer is procedurally accurate, covering all essential steps and adding more detailed instructions that enhance the overall guide to growing spinach. The flow is logical and well-organized, although it includes more steps than the expert answer. However, the additional details make the answer less concise, leading to a lower score in conciseness. Overall, the model provides a comprehensive and accurate procedure for growing baby spinach.
G-Eval score	{procedural_accuracy: 4.60, procedural_flow: 4.34, conciseness: 2.92}

Table 22: **Qualitative result of GPT-4o answer on CABBAGE sample test\_wikiHow\_471 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 19, followed by a step-by-step G-Eval analysis and scoring. The analysis includes procedural accuracy, procedural flow, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

GPT-4o answer - test_wikiHow_471	
<p>Growing baby spinach in an outdoor garden is a rewarding process that can provide you with fresh, nutritious greens. Here's a detailed step-by-step guide to help you grow baby spinach successfully:</p> <p><b>Step 1: Choose the Right Variety.</b>  <b>Select Seeds:</b> Choose a variety of spinach that is well-suited for baby leaf production. Varieties like 'Baby's Leaf', 'Corvair', or 'Space' are popular choices for baby spinach.</p> <p><b>Step 2: Timing and Location.</b>  <b>Planting Time:</b> Spinach is a cool-season crop. Plant seeds in early spring as soon as the soil is workable or in late summer for a fall crop. Spinach can tolerate light frosts, which makes it ideal for spring and fall planting.  <b>Sunlight:</b> Choose a location that receives full sun to partial shade. Spinach prefers cooler temperatures, so partial shade can help in warmer climates.</p> <p><b>Step 3: Prepare the Soil.</b>  <b>Soil Requirements:</b> Spinach grows best in well-drained, nutrient-rich soil with a pH between 6.0 and 7.0.  <b>Soil Preparation:</b> Loosen the soil to a depth of at least 12 inches and mix in plenty of organic matter, such as compost or well-rotted manure, to improve soil fertility and drainage.</p> <p><b>Step 4: Planting the Seeds.</b>  <b>Sowing Seeds:</b> Sow seeds directly into the garden bed. Plant seeds about 1/2 inch deep and 1 inch apart in rows that are 12 to 18 inches apart.  <b>Watering:</b> Water the seeds gently to ensure good seed-to-soil contact and to keep the soil moist until germination.</p> <p><b>Step 5: Care and Maintenance.</b>  <b>Thinning:</b> Once seedlings are about 2 inches tall, thin them to about 2 to 4 inches apart to allow room for growth.  <b>Watering:</b> Keep the soil consistently moist but not waterlogged. Spinach requires regular watering, especially during dry spells.  <b>Fertilizing:</b> If necessary, apply a balanced fertilizer once or twice during the growing season to promote healthy growth.  <b>Weeding:</b> Keep the bed free of weeds, which can compete with spinach for nutrients and water.</p> <p><b>Step 6: Pest and Disease Management.</b>  <b>Pests:</b> Watch out for pests such as aphids or leaf miners. Handpick or use organic insecticidal soap if infestations occur.  <b>Diseases:</b> Ensure good air circulation and avoid overhead watering to prevent fungal diseases.</p> <p><b>Step 7: Harvesting.</b>  <b>When to Harvest:</b> Baby spinach leaves are ready to harvest when they reach about 2 to 3 inches in length, typically 30 to 40 days after planting.  <b>How to Harvest:</b> Use scissors or garden shears to cut the outer leaves, allowing the inner leaves to continue growing for subsequent harvests.</p> <p>By following these steps, you can enjoy a bountiful harvest of fresh baby spinach from your garden. Growing spinach is not only easy but also provides a delicious and healthy addition to your meals.</p>	
G-Eval steps	Evaluator answer
3. <i>Model answer analysis</i>	The model's answer provides a more detailed and expanded procedure, including choosing the right variety, timing, and location, preparing the soil, planting seeds, care and maintenance, pest and disease management, and harvesting. It includes additional details such as soil pH, specific seed varieties, and pest management.
4. <i>Procedural Accuracy analysis</i>	The model's answer covers all the essential steps from the expert answer but adds additional steps and details. It includes choosing the right variety, preparing the soil, and pest management, which are not mentioned in the expert answer. The critical components like planting time, location, seed spacing, thinning, and harvesting are all present, though with more detail.
5. <i>Procedural Flow analysis</i>	The model's answer is logically structured and follows a clear sequence: selecting seeds, timing and location, soil preparation, planting, care, pest management, and harvesting. This flow is detailed and logical, though it introduces additional steps not present in the expert answer, which may affect direct comparability.
6. <i>Conciseness analysis</i>	The model's answer is more verbose than the expert answer, with seven steps compared to six in the expert answer. Each step is expanded with additional details, making it less concise. While the information is useful, it is more detailed than necessary for the basic task of growing spinach.
7. <i>Final review</i>	The model's answer is fully accurate, covering all essential steps and adding valuable details. It follows a logical sequence but introduces additional steps, affecting direct comparability with the expert answer. The answer is more verbose, providing extensive information beyond the basic requirements, which impacts conciseness.
G-Eval score	{procedural_accuracy: 4.26, procedural_flow: 3.98, conciseness: 2.84}

Table 23: **Qualitative result of GPT-4o-mini answer on CABBAGE sample test\_wikiHow\_471 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 19, followed by a step-by-step G-Eval analysis and scoring. The analysis includes procedural accuracy, procedural flow, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

GPT-4o-mini answer – test_wikiHow_471	
<p>Step-by-Step Instructions for Planting Baby Spinach in an Outdoor Garden</p> <p>Step 1: Choose the Right Time to Plant.  Timing: Spinach is a cool-season crop. Plant it in early spring (about 2-4 weeks before the last expected frost) or in late summer for a fall harvest (about 6-8 weeks before the first expected frost).</p> <p>Step 2: Select a Suitable Location.  Sunlight: Choose a location that receives full sun (at least 6 hours of sunlight per day) or partial shade, especially in warmer climates.  Soil: Spinach prefers well-draining, fertile soil with a pH between 6.0 and 7.0.</p> <p>Step 3: Prepare the Soil.  Clear the Area: Remove any weeds, rocks, or debris from the planting area.  Amend the Soil: Work in organic matter such as compost or well-rotted manure to enrich the soil. This will improve drainage and nutrient content.  Tilling: Loosen the soil to a depth of about 12 inches to help with root development.</p> <p>Step 4: Sow the Seeds.  Seed Selection: Choose a variety of spinach suitable for baby greens, such as ‘Baby Leaf’ or ‘Tyee’.  Sowing: Make shallow furrows in the soil about 1/2 inch deep and space them 12-18 inches apart. Sow the seeds about 1-2 inches apart in the furrows.  Covering: Gently cover the seeds with soil and pat down lightly to ensure good seed-to-soil contact.  Watering: Water the area gently to moisten the soil without washing away the seeds.</p> <p>Step 5: Care for the Plants.  Watering: Keep the soil consistently moist, especially during germination (which takes about 7-14 days). Avoid letting the soil dry out completely.  Thinning: Once seedlings are about 2 inches tall, thin them to 3-4 inches apart to allow for proper growth. You can use the thinned seedlings in salads.  Fertilizing: If your soil is not very rich, consider applying a balanced fertilizer according to package instructions, usually a few weeks after germination.</p> <p>Step 6: Harvesting.  Timing: Baby spinach can be harvested when the leaves are about 3-4 inches long, typically 4-6 weeks after planting.  Method: Use scissors to cut the outer leaves, allowing the inner leaves to continue growing. You can also pull the entire plant if you prefer.</p> <p>Step 7: Post-Harvest Care.  Cleaning: Rinse the harvested spinach leaves in cold water to remove any soil or debris.  Storage: Store the spinach in a plastic bag in the refrigerator, where it can last for about a week.</p> <p>By following these steps, you can successfully grow baby spinach in your outdoor garden, enjoying fresh and nutritious greens right from your own backyard!</p>	
G-Eval steps	Evaluator answer
3. <i>Model answer analysis</i>	The model’s answer is more detailed, covering: 1) Timing for spring and fall, 2) Location with sun and soil pH, 3) Soil preparation, 4) Sowing seeds with spacing, 5) Plant care with watering, thinning, and fertilizing, 6) Harvesting method and timing, 7) Post-harvest care. It includes additional details on soil preparation and post-harvest care not present in the expert answer.
4. <i>Procedural Accuracy analysis</i>	The model’s answer includes all the essential steps from the expert answer but adds more detail, such as soil preparation and post-harvest care, which are not in the expert version. It covers the necessary steps like timing, planting, spacing, thinning, and harvesting accurately, but with added depth.
5. <i>Procedural Flow analysis</i>	The model’s procedural flow is logical and clear, starting with timing, location, soil preparation, planting, care, harvesting, and post-harvest care. It follows a coherent sequence that aligns with the expert’s basic structure but expands on it with additional steps.
6. <i>Conciseness analysis</i>	The model’s answer is more verbose than the expert’s, with additional steps and details. While it provides comprehensive information, it is less concise due to the extra details on soil preparation and post-harvest care. The number of steps is greater, and each step contains more information than the expert answer.
7. <i>Final review</i>	The model’s answer is procedurally accurate, covering all necessary steps and adding useful details on soil preparation and post-harvest care, which enhance the instructions. The flow is logical and clear, following a coherent sequence that aligns with the expert’s basic structure. However, the model’s answer is less concise due to the additional details, resulting in a lower conciseness score. Overall, the model provides a thorough and well-organized procedure for growing baby spinach.
G-Eval score	{procedural_accuracy: 4.36, procedural_flow: 4.44, conciseness: 3.09}

Table 24: **Qualitative result of Gemini-2.0-Flash answer on CABBAGE sample test\_wikiHow\_471 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 19, followed by a step-by-step G-Eval analysis and scoring. The analysis includes procedural accuracy, procedural flow, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

<b>Gemini-2.0-Flash answer – test_wikiHow_471</b>	
To successfully grow baby spinach in your outdoor garden, follow these detailed step-by-step instructions:	
<ol style="list-style-type: none"> <li>1. Choose the Right Time: Spinach thrives in cool weather. Plant spinach in early spring (as soon as the soil can be worked) or in late summer/early fall for a fall/winter harvest. Aim for planting when daytime temperatures are consistently between 50°F and 70°F (10°C and 21°C).</li> <li>2. Select a Suitable Location: Choose a spot in your garden that receives at least 3-6 hours of sunlight per day. Spinach can tolerate partial shade, especially in warmer climates. Ensure the location has well-draining soil.</li> <li>3. Prepare the Soil: Spinach prefers fertile, well-drained soil with a pH between 6.0 and 7.0. Amend the soil with compost or well-rotted manure before planting to improve its fertility and drainage. Remove any rocks, weeds, and debris from the planting area. Rake the soil surface smooth.</li> <li>4. Sow the Seeds: Sow spinach seeds directly into the prepared soil. Plant seeds ½ inch deep and 1 inch apart in rows that are 12-18 inches apart. You can also broadcast the seeds over a wider area and then thin the seedlings later.</li> <li>5. Water Gently: After sowing the seeds, water the soil gently to avoid disturbing the seeds. Keep the soil consistently moist but not waterlogged until the seedlings emerge.</li> <li>6. Thin Seedlings: Once the seedlings have developed their first true leaves (about 2-3 weeks after germination), thin them to about 3-6 inches apart. This will give the remaining plants enough room to grow and mature. You can eat the thinned seedlings as microgreens.</li> <li>7. Water Regularly: Spinach needs consistent moisture to grow well. Water the plants regularly, especially during dry periods. Aim to keep the soil evenly moist but not soggy. Water deeply at the base of the plants to avoid wetting the foliage, which can lead to disease.</li> <li>8. Fertilize (Optional): Spinach is a heavy feeder and can benefit from supplemental fertilization. If your soil is not very fertile, you can side-dress the plants with a balanced organic fertilizer or compost tea every 2-3 weeks. Follow the instructions on the fertilizer package.</li> <li>9. Control Weeds: Keep the planting area free of weeds, which can compete with the spinach plants for nutrients and water. Hand-pull weeds carefully to avoid disturbing the spinach roots. You can also mulch around the plants with straw or shredded leaves to help suppress weed growth and retain soil moisture.</li> <li>10. Monitor for Pests and Diseases: Keep an eye out for common spinach pests such as aphids, flea beetles, and leaf miners. Control pests with insecticidal soap or neem oil. Prevent fungal diseases by providing good air circulation and avoiding overhead watering. Remove any infected leaves promptly.</li> <li>11. Harvest: Baby spinach can be harvested when the leaves are young and tender, typically about 3-6 weeks after planting. Cut the outer leaves with scissors or a knife, leaving the inner leaves to continue growing. You can harvest spinach multiple times from the same plant. Harvest in the morning when the leaves are crisp and cool.</li> <li>12. Succession Planting: To enjoy a continuous supply of baby spinach, sow new seeds every 2-3 weeks. This is especially important in the spring and fall when temperatures are ideal for spinach growth.</li> </ol>	
G-Eval steps	Evaluator answer
3. <i>Model answer analysis</i>	The model's answer is more detailed, with 12 steps. It includes additional information on planting time, location selection, soil preparation, seed sowing, watering, thinning, regular watering, optional fertilization, weed control, pest monitoring, harvesting, and succession planting. The steps are logically ordered and cover a broader range of gardening practices.
4. <i>Procedural Accuracy analysis</i>	The model's answer covers all the essential steps from the expert answer, including planting time, location, seed sowing, thinning, and harvesting. It also adds more detailed guidance on soil preparation, watering, fertilization, weed control, and pest management, which enhances the procedural accuracy by providing a comprehensive approach to growing spinach.
5. <i>Procedural Flow analysis</i>	The model's answer presents the steps in a clear and logical sequence, starting from preparation to planting and maintenance, and finally harvesting. The flow is coherent and practical, with each step logically following the previous one. The additional details do not disrupt the flow and instead provide a more thorough guide.
6. <i>Conciseness analysis</i>	The model's answer is more verbose than the expert's, with additional steps and details. While it includes more information, it maintains focus on the essentials of growing spinach. The extra details, such as pest control and fertilization, contribute to a more comprehensive guide without unnecessary verbosity.
7. <i>Final review</i>	The model's answer is highly accurate, covering all essential steps and adding useful additional details that enhance the procedure's effectiveness. The procedural flow is well-organized and logical, making it easy to follow. While the model's answer is more detailed than the expert's, it remains focused on essential information, earning a high score for conciseness. Overall, the model provides a comprehensive and well-structured guide for growing baby spinach.
G-Eval score	{procedural_accuracy: 4.72, procedural_flow: 4.69, conciseness: 3.72}

Table 25: **Qualitative result of Gemini-2.0-Flash-Lite answer on CABBAGE sample test.wikiHow\_471 using G-Eval with GPT-4o as evaluator.** This table presents the response generated by the model to the same question shown in Table 19, followed by a step-by-step G-Eval analysis and scoring. The analysis includes procedural accuracy, procedural flow, and conciseness assessments, as well as a final review synthesised by GPT-4o. G-Eval steps 3 and onward are computed per model and reflect performance relative to the expert reference.

<b>Gemini-2.0-Flash-Lite answer – test.wikiHow_471</b>	
<p>Here's how to grow baby spinach in your outdoor garden:</p> <ol style="list-style-type: none"> <li>1. Choose the Right Time and Location:            Timing: Spinach is a cool-season crop. The best time to plant is in early spring (as soon as the ground can be worked) or in the fall (6-8 weeks before the first expected frost). Avoid planting in the heat of summer, as it will bolt (go to seed) quickly and become bitter.            Location: Choose a spot that gets at least 4-6 hours of sunlight per day. Spinach can tolerate partial shade, especially in warmer climates. Ensure the area has well-draining soil.</li> <li>2. Prepare the soil:            Soil Test: Before planting, it's helpful to get your soil tested to determine its pH and nutrient levels. Spinach prefers a slightly alkaline soil with a pH of 6.5 to 7.0.            Amend the Soil: Amend the soil with compost or well-rotted manure to improve drainage and fertility. Spinach is a heavy feeder, so rich soil is essential. Work the compost into the top 6-8 inches of soil.</li> <li>3. Planting Spinach Seeds:            Direct Sowing: Spinach is best grown from seed. Sow seeds directly into the garden. Sowing Depth and Spacing: Plant seeds about 1/2 inch deep and 1 inch apart. If you're growing baby spinach, you can sow seeds more densely. Sow seeds in rows that are 12-18 inches apart. Watering: After planting, water the soil gently but thoroughly to keep the soil consistently moist until the seeds germinate. Germination usually takes 7-14 days.</li> <li>4. Caring for Your Spinach:            Watering: Water regularly, especially during dry periods. Aim to keep the soil consistently moist, but not waterlogged. Water at the base of the plants to avoid wetting the leaves, which can promote disease.            Thinning: Once the seedlings have a few true leaves, thin them to about 3-4 inches apart for mature spinach or closer for baby spinach. You can eat the thinnings!            Fertilizing: Side-dress the plants with a balanced fertilizer or compost tea a few weeks after thinning and again a few weeks later. This will provide the nutrients spinach needs to grow well.            Weeding: Keep the area around the spinach plants free of weeds. Weeds compete for nutrients, water, and sunlight.            Pest and Disease Control: Spinach is relatively pest-resistant, but watch out for common garden pests like aphids and leaf miners. You can use insecticidal soap or neem oil to control pests if necessary. Avoid overhead watering to prevent fungal diseases.</li> <li>5. Harvesting Baby Spinach:            Harvesting Time: You can start harvesting baby spinach leaves about 3-4 weeks after planting. Harvest the outer leaves when they are about 3-4 inches long. You can harvest the entire plant if you prefer.            Harvesting Method: Cut the leaves at the base, leaving the inner leaves to continue growing. This allows for multiple harvests from the same plant.            Continuous Harvest: For a continuous harvest, sow seeds every 2-3 weeks.</li> <li>6. Additional Tips:            Succession Planting: Plant spinach in succession (every few weeks) to extend your harvest season.            Mulching: Apply a layer of mulch (such as straw or shredded leaves) around the plants to help retain moisture, suppress weeds, and regulate soil temperature.            Bolting Prevention: To prevent bolting, keep the plants well-watered and shaded during hot weather. Choose bolt-resistant varieties.            Variety Selection: Consider different spinach varieties. Some are more bolt-resistant, while others have different leaf shapes and textures.</li> </ol>	
<b>G-Eval steps</b>	<b>Evaluator answer</b>
3. <i>Model answer analysis</i>	The model's answer is more detailed, covering timing, location, soil preparation, planting, care, harvesting, and additional tips. It includes specific instructions for soil testing, amending, watering, thinning, fertilizing, pest control, and continuous harvesting.
4. <i>Procedural Accuracy analysis</i>	The model's answer includes all the essential steps from the expert answer and adds more detailed instructions on soil preparation, care, and additional tips. It covers the planting process comprehensively and accurately.
5. <i>Procedural Flow analysis</i>	The model's answer is logically organized into sections: choosing time and location, preparing soil, planting, caring, harvesting, and additional tips. The flow is clear and sequential, aligning well with the expert answer's structure but with more detail.
6. <i>Conciseness analysis</i>	The model's answer is more verbose than the expert answer, with additional details and tips. While it is comprehensive, it could be seen as less concise due to the extra information, even though it efficiently covers the necessary steps.
7. <i>Final review</i>	The model's answer provides a comprehensive and detailed procedure for growing baby spinach, covering all essential steps and adding valuable details on soil preparation, care, and harvesting. It maintains a logical flow and sequence, aligning well with the expert answer while offering more depth. However, the additional information makes it less concise compared to the expert's more succinct instructions.
G-Eval score	{procedural_accuracy: 4.76, procedural_flow: 4.51, conciseness: 3.28}