

# What Large Language Models Know About Plant Molecular Biology

**Manuel Fernandez Burda<sup>1,2</sup>, Lucia Ferrero<sup>2</sup>, Nicolás Gaggion<sup>1,2</sup>, Camille Fonouni-Farde<sup>3,4</sup>, The MoBiPlant Consortium<sup>†</sup>, Martín Crespi<sup>3,4</sup>, Federico Ariel<sup>\*5,6</sup>, and Enzo Ferrante<sup>\*1,2</sup>**

<sup>1</sup>Instituto de Ciencias de la Computación (ICC), CONICET-Universidad de Buenos Aires, Buenos Aires, C1428EHA, Argentina.

<sup>2</sup>APOLO Biotech, Argentina.

<sup>3</sup>Institute of Plant Sciences Paris Saclay IPS2, CNRS, INRA, Université Evry, Université Paris-Saclay, Bâtiment 630, Orsay, 91405, France.

<sup>4</sup>Institute of Plant Sciences Paris-Saclay IPS2, Université de Paris, Bâtiment 630, Orsay, 91405, France.

<sup>5</sup>Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE), CONICET-Universidad de Buenos Aires, Buenos Aires, C1428EHA, Argentina.

<sup>6</sup>Universidad Nacional del Litoral, Santa Fe, Argentina.

*\*Correspondence to: fariel@fbmc.fcen.uba.ar and eferrante@dc.uba.ar*

## ABSTRACT

Large language models (LLMs) are rapidly permeating scientific research, yet their capabilities in plant molecular biology remain largely uncharacterized. Here, we present MOBIPLANT, the first comprehensive benchmark for evaluating LLMs in this domain, developed by a consortium of 112 plant scientists across 19 countries. MOBIPLANT comprises 565 expert-curated multiple-choice questions and 1,075 synthetically generated questions, spanning core topics from gene regulation to plant-environment interactions. We benchmarked seven leading chat-based LLMs using both automated scoring and human evaluation of open-ended answers. Models performed well on multiple-choice tasks (exceeding 75% accuracy), although most of them exhibited a consistent bias towards option A. In contrast, expert reviews exposed persistent limitations, including factual misalignment, hallucinations, and low self-awareness. Critically, we found that model performance strongly correlated with the citation frequency of source literature, suggesting that LLMs do not simply encode plant biology knowledge uniformly, but are instead shaped by the visibility and frequency of information in their training corpora. This understanding is key to guiding both the development of next-generation models and the informed use of current tools in the everyday work of plant researchers. MOBIPLANT is publicly available online [in this link](#).

## 1 Introduction

Large language models (LLMs) have demonstrated transformative potential across the research lifecycle, assisting with tasks such as automated literature surveys, data extraction, and experimental design guidance<sup>1</sup>. Researchers now use LLMs not only to improve writing quality, but also to interpret biological literature, propose novel research directions, assist with data analyses, and uncover functional patterns in complex biological data<sup>2</sup>. Despite their widespread adoption, these models still require careful expert validation to understand their capabilities and limitations, as domain-specific evaluations remain limited.

<sup>†</sup>The **MoBiPlant Consortium** members (affiliation information in Supplementary Table 1) contributed equally to this work: Alvarez José M., Arana María V., Arribas Hernández Laura, Attallah Carolina, Auge Gabriela, Ballaré Carlos L., Benhamed Moussa, Benoit Matthias, Blanco Flavio A., Blanco Nicolás E., Blanco-Herrera Francisca, Blazquez Miguel A., Blein Thomas, Bologna Nicolas G., Bouché Nicolas, Bourbousse Clara, Busch Wolfgang, Calixto Cristiane P.G., Calzadilla Pablo I., Carbonell Alberto, Carrari Fernando, Casal Jorge, Casati Paula, Cerdán Pablo D., Chorostecki Uciel, Conti Gabriela, Crevillén Pedro, Crocco Carlos D., Csorba Tibor, Cubas Pilar, Debat Humberto, de Carvalho-Niebel Fernanda, Delaux Pierre-Marc, Dezar Carlos, Drincovich Maria F., Estevez Jose M., Fankhauser Christian, Figueroa Carlos M., Fiol Diego F., Fragkostefanakis Sotirios, Frugier Florian, Fusari Corina M., Gago-Zachert Selma, García-Mata Carlos, Gastaldi Victoria, Godoy Herz Micaela A., Gras Diana E., Gudesblat Gustavo E., Gutiérrez Rodrigo A., Henriques Rossana, Ibañez Fernando, Iglesias María J., Kindgren Peter, Krapp Anne, Laxalt Ana M., Legascue María F., Legris Martina, Lia Verónica V., Liu Chang, Llorente Briardo, Lucero Leandro E., Maizel Alexis, Mammarella María F., Manavella Pablo A., Mansilla Natanael, Maroniche Guillermo A., Marquardt Sebastian, Martín-Pizarro Carmen, Mateos Julieta L., Mathur Saloni, Mencia Regina, Merchante Catharina, Moreno Javier E., Muschietti Jorge P., Niebel Andreas, Orellana Ariel, Pagnussat Luciana A., Pagnussat Gabriela C., Palatnik Javier F., Parmar Rajni, Petrillo Ezequiel, Posé David, Prochetto Santiago, Raynaud Cécile, Ricardi Martiniano M., Rivarola Maximo, Rodríguez Ramiro E., Rodríguez Melo Johan S., Romani Facundo, Roy Sonali, Sáez-Vásquez Julio, Segretin María E., Selma Sara, Signorelli Santiago, Soto Gabriela C., Staiger Dorothee, Tessi Tomas M., Tomassi Ariel H., Tonelli María L., Traubenik S., Vaucheret H., Verslues Paul E., Wang Dong, Welchen Elina, Yanovsky Marcelo J., Zanetti María E., Zavallo Diego and Zhang Wenli.

To date, most large-scale evaluations focus on general-purpose or biomedical domains, leaving gaps in specialized fields like plant molecular biology<sup>3</sup>. Notably, LLMs have demonstrated proficiency in logical reasoning and scientific benchmarks, often surpassing previous performance records. While benchmarks like SciBench<sup>4</sup>, LAB-Bench<sup>5</sup> and BrainBench<sup>6</sup> have been developed to assess LLMs' scientific knowledge and practical research capabilities, they primarily focus on general scientific disciplines or specific areas like biology, chemistry or neuroscience. But standardized, domain-specific evaluations remain scarce, particularly in plant molecular biology. To address this gap, we introduce the Plant Molecular Biology (MOBiPLANT) Benchmark, the first expert-curated plant-biology benchmark created by a consortium of more than 110 plant scientists residing in 19 different countries, combining manual and synthetically generated multiple-choice questions (MCQs) with rich human-evaluation of open-ended answers.

The few existing studies which survey the use of AI in research environments, underscore both excitement and caution, highlighting how researchers leverage chat assistants to accelerate literature review and experiment planning, yet worry about uncontrolled hallucinations and biorisk implications<sup>7</sup>. Meanwhile, frontier systems demonstrate that agentic LLMs can achieve or exceed expert performance on literature-search and contradiction-detection tasks<sup>2,8,9</sup>, suggesting that with the right evaluation frameworks, models could reliably support scientific discovery. In agriculture there have been approaches to leverage farmer queries through LLM-powered systems<sup>10</sup>. For genomics, initial studies reveal feasibility but also highlight persistent domain-specific challenges: for example, ChatGPT's outputs exhibit "plant blindness" and limited taxonomic diversity when answering botanical questions<sup>11</sup>; and LLMs, while extracting ecological data more than 50 times faster than human reviewers, require rigorous quality-control to avoid quantitative errors<sup>2</sup>.

Moreover, these models have been shown to misinterpret bioinformatics figures (such as reversing up- and down-regulated gene sets in differential expression plots) producing plausible yet incorrect biological narratives<sup>12</sup>. Another example is chat-based extraction of natural-product bioactivity from the literature, which yields high recall but introduces false positives, assigning activity values to compounds never tested<sup>13</sup>. Clinical workflows further demonstrate that, without careful oversight, LLMs can present over-confident hallucinations—such as fabricated drug interactions—or omit critical diagnostic details, reinforcing the necessity of human-in-the-loop validation<sup>14</sup>.

The proposed research pursues three interlocking objectives to rigorously assess LLM capabilities in plant molecular biology. First, we settled an expert Consortium composed of 112 scientists specialized in the field (Figure 1): the MoBiPlant Consortium. Second, we built an open and community-driven benchmark that unites expert-crafted, PhD-level MCQs (Expert MoBiPlant) with a synthetically expanded question set (Synthetic MoBiPlant), ensuring both the precision of domain-vetted items and the topical breadth afforded by controlled data augmentation. Third, we carried out a dual-mode evaluation: high-throughput automated scoring on the MCQs to establish a baseline accuracy, alongside in-depth human assessment of open-ended responses, examining not only correctness but also alignment with scientific consensus, potential of species bias, logical reasoning, and self-awareness.

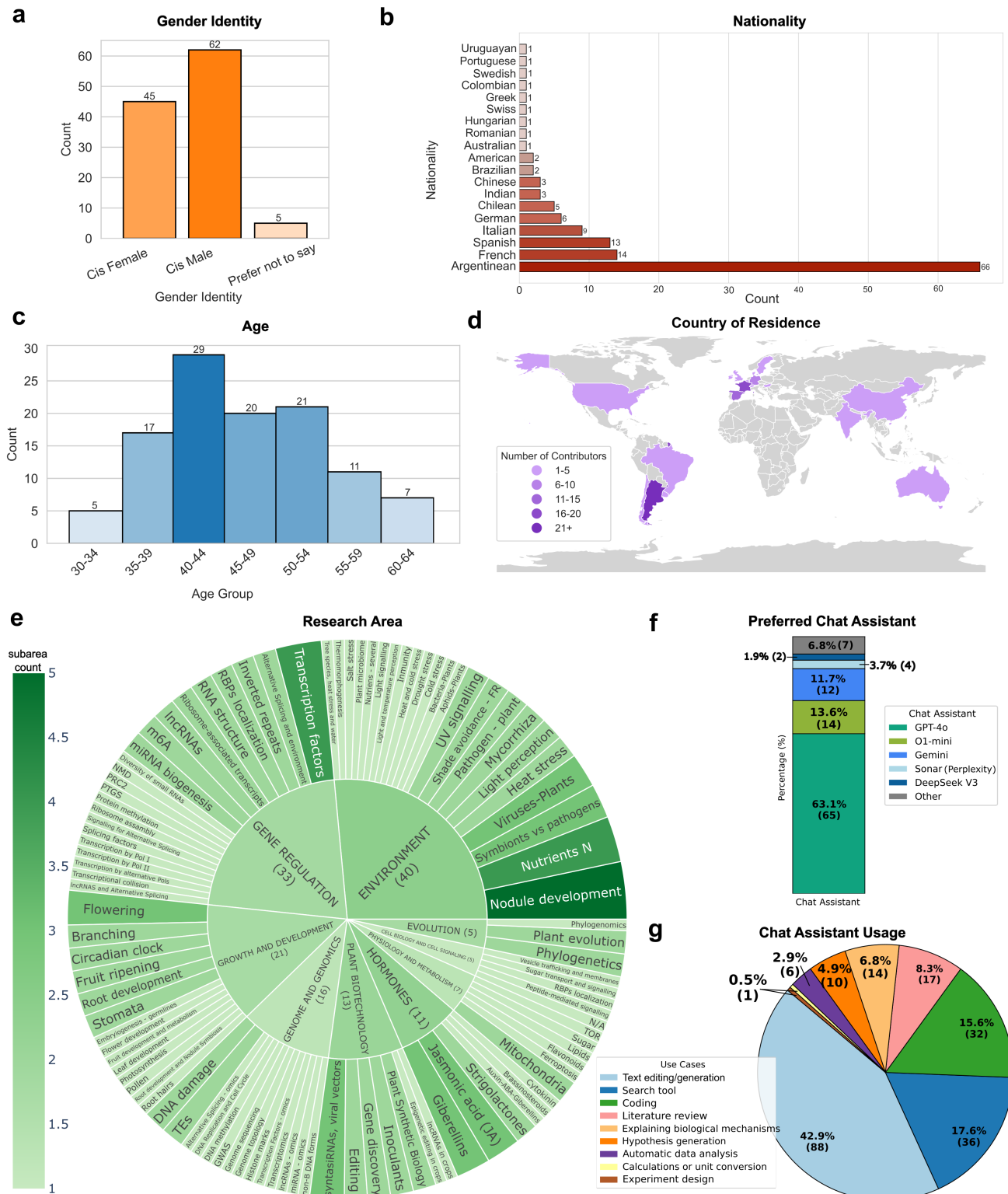
We evaluated some of the most popular LLMs used as chat assistants by scientists in plant molecular biology: Gemini<sup>15</sup> (1.5 Pro), ChatGPT<sup>16,17</sup> (both GPT-4o and o1-mini), DeepSeek<sup>18,19</sup> (both V3 and R1); a well-known open source chat assistant: LLaMA<sup>20</sup> (3.1 405B); and an assistant excelling in scientific benchmarks: Claude<sup>21</sup> (3.5 Sonnet). Through the MOBiPLANT framework, we uncovered model specialization patterns across domains and performance features. Finally, we offered a modular MCQ construction protocol for PhD-level question design, enabling the community to extend and adapt our benchmark to new scientific domains.

Our results highlight both the strengths and weaknesses of popular LLMs when tasked with understanding and reasoning about plant molecular biology. Despite demonstrating strong performance in structured MCQ tasks, these models still struggle with factual accuracy, hallucinations, and logical reasoning, particularly in the context of novel or less-represented scientific concepts. Notably, the fact that performance of LLMs correlates with the citation frequency of the source papers suggests that models may be better equipped to handle well-established concepts that are more likely to appear in their training data, while struggling with newer or less-represented material. Through our comprehensive benchmarking, we underscore the need for targeted improvements to enhance the reliability and scientific rigor of LLMs, particularly in specialized domains like plant molecular biology.

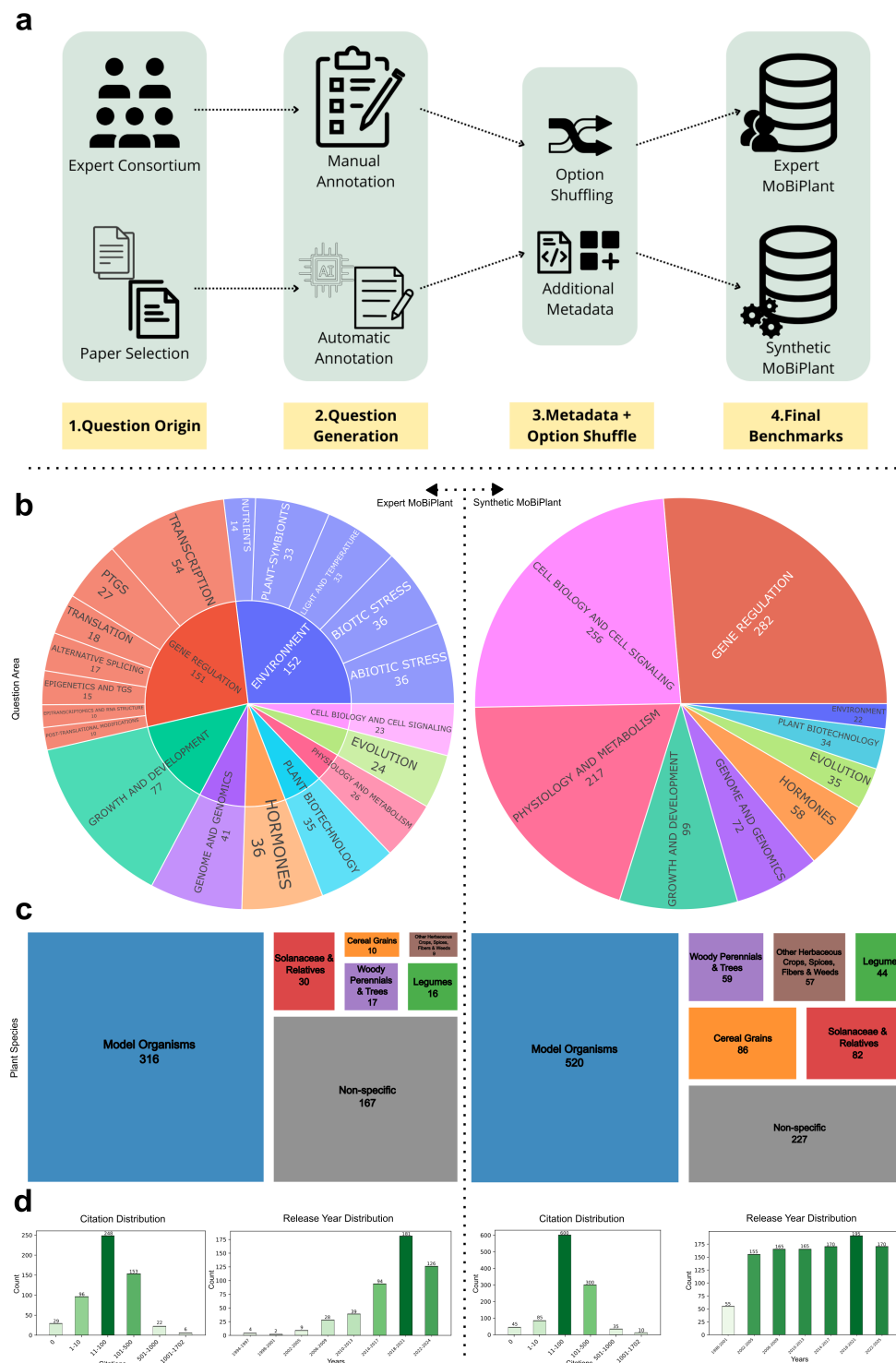
## 2 Results

### 2.1 MoBiPLANT allows assessment of LLMs understanding in plant molecular biology

Aiming to understand the limitations of popular LLMs when used by plant molecular researchers in their everyday routine, we created a diverse consortium of 112 plant molecular biology researchers from around the world, referred to as the MoBiPlant Consortium (Figure 1), which was in charge of elaborating complex questions and evaluating open-ended answers. Our evaluation methodology to assess LLMs' understanding of plant molecular biology consisted of a two-stage experimental design: (1) question generation and (2) answer evaluation. During the question generation stage, researchers from the MoBiPlant Consortium were tasked with selecting relevant, high-quality scientific papers and creating complex, domain-specific questions



**Figure 1. MOBiPLANT Consortium Description** including a) gender, b) nationality, c) age, d), country of residence and e) research area. It is worth mentioning that in ‘Nationality’ and ‘Research Area’, contributors might have reported more than one answer. f) In addition, we queried the Consortium which LLM was their favourite to use as assistant in research contexts, g) and how they employed them.



**Figure 2. MoBiPlant Overview.** **a)** MoBiPlant construction pipeline; **a.1)** The MoBiPlant Consortium was created and the source papers were selected, ensuring no overlap. **a.2)** Scientists manually created the Expert questions based on their expertise, always indicating a source paper where the answer is rooted, whereas synthetic ones were automatically generated from source papers via a frontier LLM, with question topics and plant species annotated at this stage. **a.3)** Answer options were shuffled, and metadata—such as plant species names—was standardized and cleaned. **a.4)** Both expert and synthetic benchmarks. **b)** Question area organization: the inner circle contains the main areas, while secondary areas (if applicable) are in the outer circle. **c)** Plant species distribution grouped by 6 delimiting categories. **d)** Source paper distribution according to release year and citation count.

based on the content of these papers (Figure 2.a). The consortium elaborated the questions, established correct answers, and then conducted a personal in-depth evaluation of LLM-generated open-ended responses. As a result, MoBiPlant is the first comprehensive, expert-driven benchmark that integrates rigorously vetted MCQs with open-ended assessments created and reviewed by domain specialists. Moreover, each question was accompanied by a correct answer and two plausible distractors, designed to form a MCQ set. These questions constituted the **Expert MoBiPlant** set.

To scale up our study and perform a more robust analysis, we constructed an additional set of 1075 synthetically generated questions derived from already published papers (Figure 2.a). More precisely, we employed a frontier LLM (see Methods 5.5) to generate synthetic MCQs based on existing scientific papers. In this process, the LLM was provided with a paper manuscript and prompted to produce questions, each with one correct answer and two incorrect alternatives. As we detail in Methods 5.5, an expert group validated a subset of these questions ensuring the set generated attained to the criteria. This set of LLM-generated questions was referred to as the **Synthetic MoBiPlant** set.

In the second stage, i.e. answer evaluation, we prompted alternative popular LLMs (see Methods 5.2) to respond to both sets of questions in two distinct formats:

- **MCQ format**, where the model selected the correct answer from a predefined list of options for both Expert and Synthetic sets. Evaluation was performed automatically, by simply comparing the selected answer against the correct answers, and computing the accuracy (proportion of correct predictions made by the model out of all predictions).
- **Open-ended format**, only for the Expert MoBiPlant set, allowing the models to generate full, unconstrained responses. A detailed manual evaluation was conducted by experts from the MoBiPlant Consortium, who assessed the quality and factuality of the model-generated answers. This evaluation was based on a set of pre-defined criteria (see Table 2).

## 2.2 LLMs Encode Plant Molecular Biology Knowledge

When benchmarking 7 of the most popular LLMs using the Expert MoBiPlant set in MCQ format, we see that all models achieve results over 75% accuracy (Figure 3.a). We find Claude 3.5 Sonnet to be the best performing model in our dataset with an overall score of 88.1%, highlighting its strong capabilities in agreement with previous observations on different scientific domains<sup>22</sup>.

The model ranking for questions coming from different research areas highlights topic specialization across LLMs (Figure 3.b). By plotting the shifts in models' order for all the areas benchmarked in Expert MoBiPlant, it is notable that there is a correlation between the two best scoring models, i.e. Claude 3.5 Sonnet and DeepSeek R1, in the global score chart and their rank in every area, securing the first two places in most categories. This however does not hold true for every other model, where the rankings get shuffled among areas. Interestingly, this topic specialization is clearly depicted in Plant Biotechnology, where Gemini, the lowest scoring model in our benchmark, achieves a score of 91.4%, securing the second place among top performances.

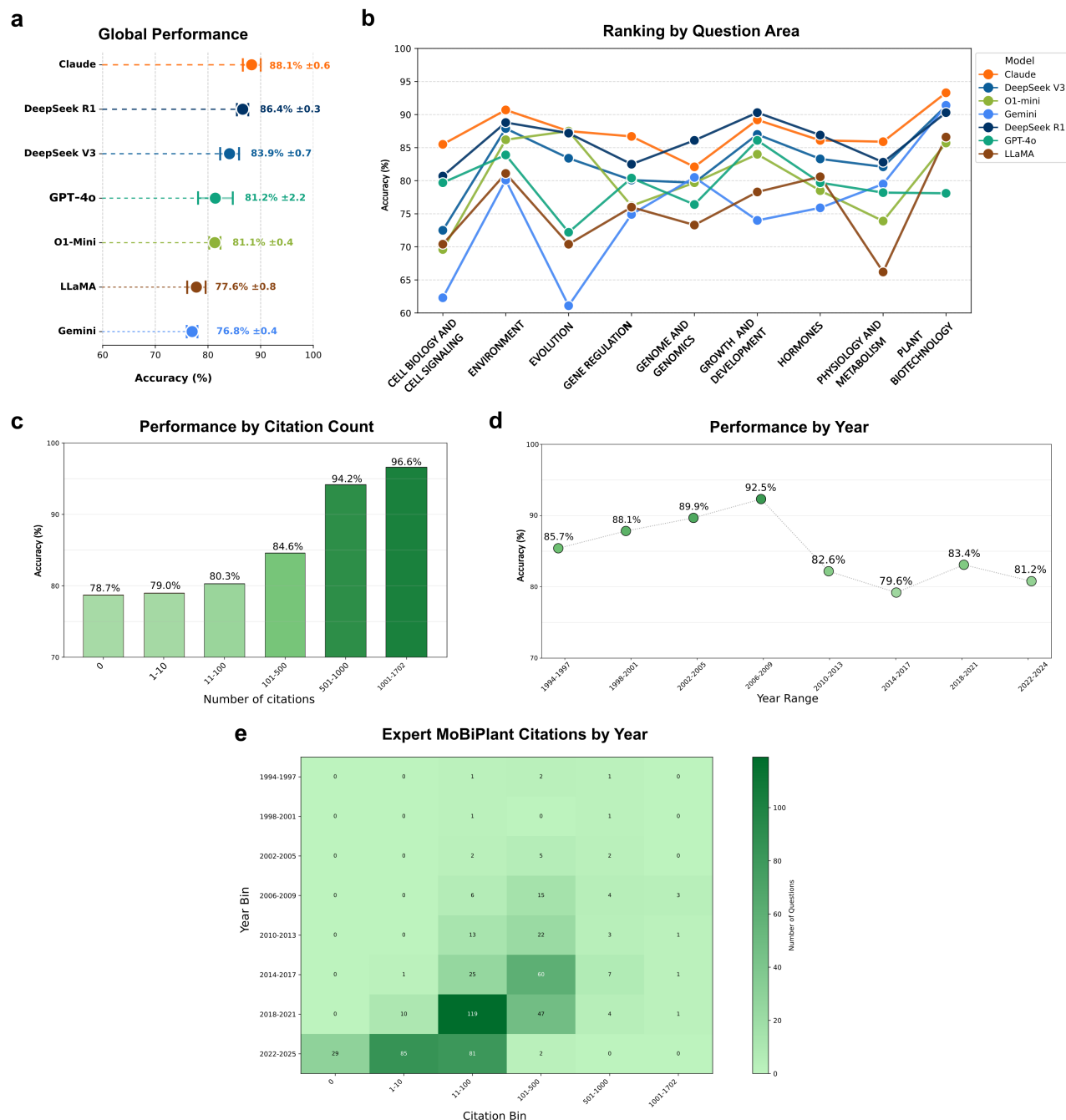
To better support our observations, we assessed the performance of the same LLMs on the Synthetic MoBiPlant set of MCQs, performing identical analyses. Overall, the scores are positively shifted (Supplementary Figure 1.a), granting diverse gains in model performance; from 0.03 (Claude) to 19.7 (LLaMA) in percentage points. The overall higher accuracy suggests that this set might be easier for models to answer, highlighting the importance of expert human intervention in creating challenging questions to evaluate model understanding. In addition, the leaderboard gets shuffled in this case with Deepseek V3, Claude and LLaMA leading. The same model specialization can be observed, as the model position significantly varies across areas (Supplementary Figure 1.b). Similar to the assessment of LLMs using the MoBiPlant Expert set (Figure 3.c), referring to the performance by citation count on Synthetic MoBiPlant set results in a clear positive correlation between the amount of citations and model accuracy (Supplementary Figure 1.c).

## 2.3 Established Domain Knowledge Shapes LLM Performance

To assess how citation impact and publication year influence model performance on MoBiPlant, we analyzed questions anchored to specific source papers. For each source paper, we extracted its publication year and citation count, and stratified the data into count bins (Figure 3.c-d). Our analysis revealed a pronounced correlation between citation frequency and model accuracy: studies with citation counts in the 501–1,702 range exhibited significantly higher accuracy compared to those in the 0–500 interval. This trend likely arises because highly cited works are more frequently represented in training corpora, either through direct inclusion or repeated paraphrasing, thereby reinforcing the model's familiarity with their content.

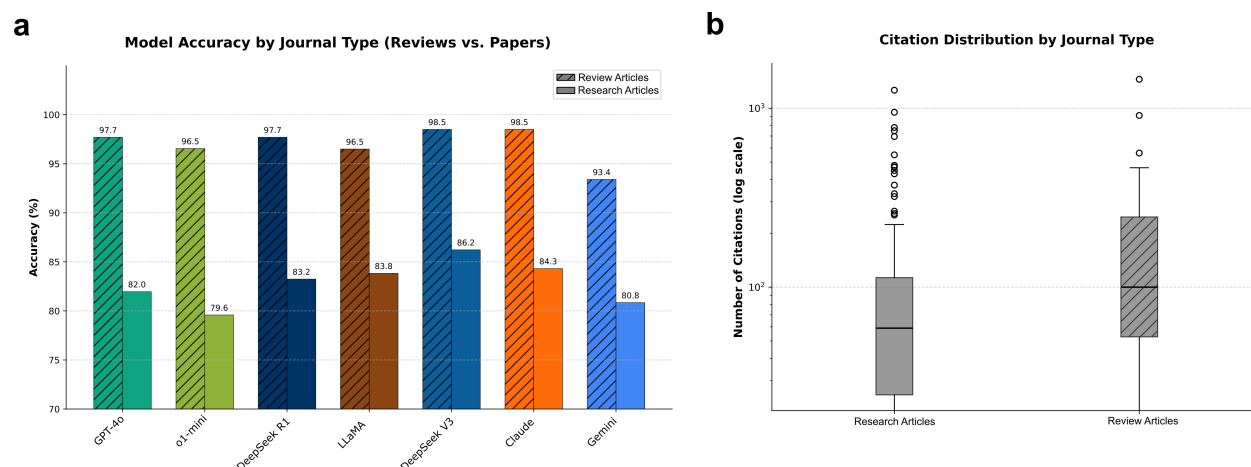
Interestingly, when papers were stratified by publication year, we observed a persistent trend: for articles published until 2009, the related questions were answered with higher accuracy (ranging from 87.5% to 92.5%) than those published more recently (since 2010, with the accuracy ranging from 79.6% to 83.4%). Considering the correlation between citations and accuracy (Figure 3.c), we assessed if older papers picked by the MoBiPlant members were more highly cited. However, this difference was not observed (Figure 3.e). Remarkably, the accuracy trend observed in the Synthetic MoBiPlant set





**Figure 3. Expert MoBiPlant MCQ benchmarking results.** **a)** The overall accuracy of different LLMs across Expert MoBiPlant. We report the mean value of accuracy among its standard deviation as error bars for all 3 independent repetitions of model answering, shuffling the option order on every repetition. **b)** Fluctuation of model ranking across question areas. Dots represent the model accuracy for different question areas. **c)** Performance by citation count and **d)** year, extracted from the paper that originated the question. The color intensity is proportional to the accuracy score on each plot. **e)** The represented publications broken down by citations and year of publication in Expert MoBiPlant accounted for with a heatmap, highlighting the intersected coverage between year of publication and amount of citations ranges.

(Supplementary Figure 1.d) closely mirrors that of the Expert MoBiPlant set. However, the difference in accuracy between older and more recent articles is less pronounced, ranging from 86.3% to 89% for articles published up to 2009, and from 85.1% to 86.8% for those published since 2010. Importantly, the distribution of citations across articles stratified by year of publication



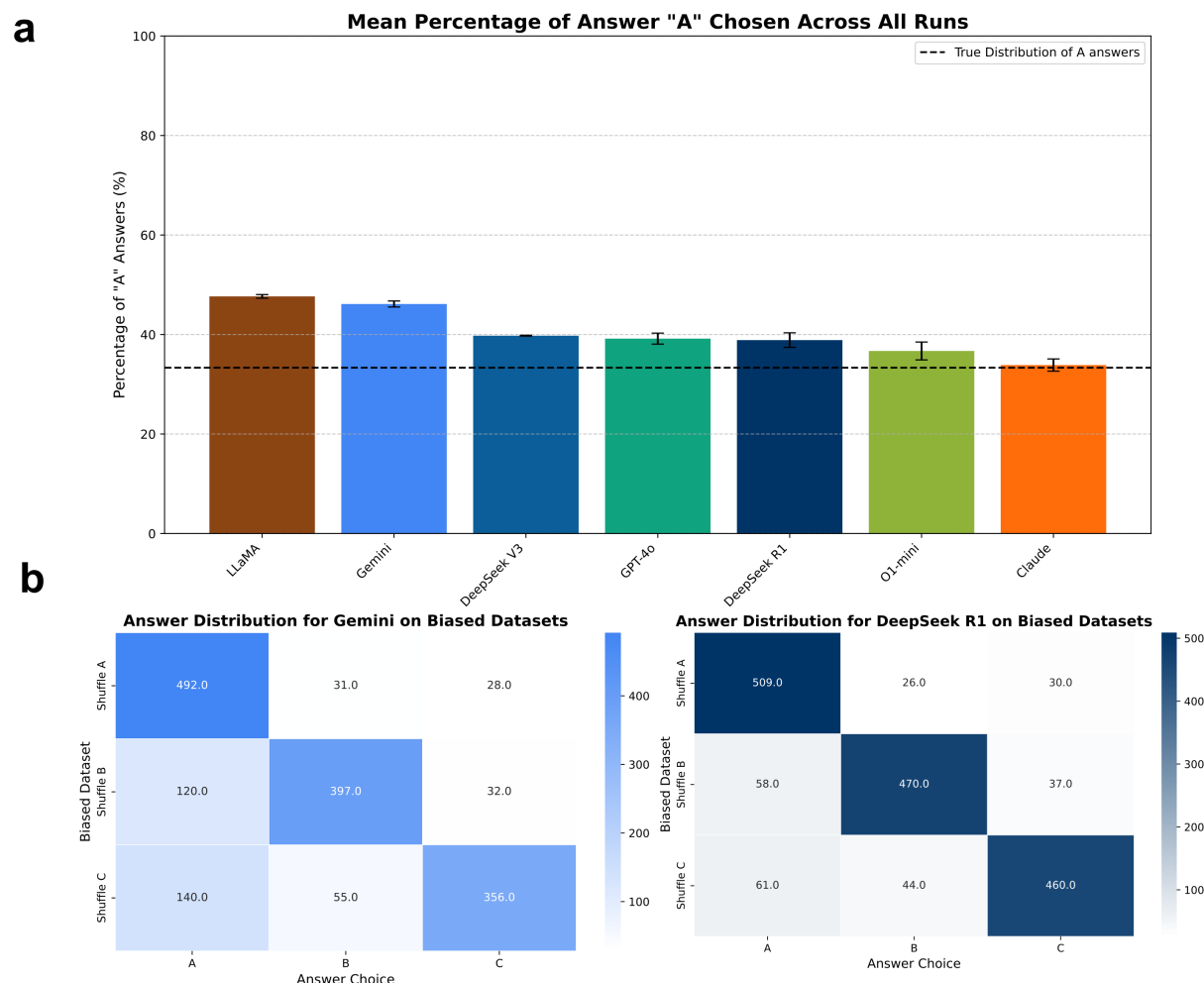
**Figure 4. Performance of LLMs based on the type of journal used as source of information.** Source journals are grouped by their type of publication, either review articles or research articles. **a)** Accuracy by model on each journal, where the dashed bar represents the mean accuracy over review journals and the solid bar represents the mean accuracy over the paper journals. **b)** The citation distribution among these two groups, represented in boxplots. Y-axis: citations on logarithmic scale. We further verified if there was a citation bias on these groups computing the Mann-Whitney U-test with a p-value of  $4.536e-13$ , suggesting a statistically significant difference in citation distributions between review and paper journals.

remains relatively uniform, with a more balanced number of articles per year compared to the Expert MoBiPlant set—thereby minimizing potential bias (Supplementary Figure 1.e). One plausible explanation for the higher accuracy of LLMs answers to questions based on older articles—despite these articles not necessarily being more highly cited—is that the knowledge they introduced has been indirectly cited and assimilated into more recent publications, becoming part of the established scientific consensus. To test this hypothesis, we segregated articles in the Synthetic MoBiPlant set into research papers and review articles. Since review articles aim to summarize, synthesize, and critically analyze the current state of knowledge on a given topic, they offer an ideal lens through which to examine the consolidation and diffusion of scientific information. By splitting the journals into two groups: review articles and research articles (papers), we calculated the accuracy on each of them (Figure 4.a). Interestingly, reviews surpass papers by 10 to 15 percentage points on each model, albeit the slight—yet significant—difference in citation distributions between reviews and papers (Figure 4.b). These findings suggest that review articles tend to reflect more consolidated knowledge which is better assimilated by LLMs. Accordingly, review publications often synthesize well-established conceptual frameworks that form part of the shared understanding in plant molecular biology, resulting in clearer and more universally accepted ideas. Taken together, our analyses of citation trends and the integration of key concepts in review articles underscore the role of canonical knowledge in shaping LLM performance in the domain of plant molecular biology.

## 2.4 Answer Order Influences LLM Performance On MCQs

As explained in the Methods section, answer options in the MCQs were randomly shuffled into three groups to minimize positional bias. While approximately 33% of the correct answers were assigned to option A in each group, five out of seven from the evaluated models selected more than 37% of the time the option A, with LLaMA reaching over 45% (Figure 5.a). Remarkably, O1-mini and Claude 3.5 Sonnet did not exhibit any bias towards the option A. This observation suggests that when uncertain, the rest of the LLMs tend to default to option A, potentially inflating their performance metrics or obscuring their true accuracy. Considering that LLMs previously showed bias towards the order of answers in MCQs<sup>23,24</sup>, we conducted a parallel experiment on models' election distribution (A,B or C) to explore potential option-bias risks. We generated 3 distinct versions of Expert MoBiPlant, reordering the options so that the correct option was fixed in a particular choice (the first having every correct option in A, the second in B and the third in C). Then, we evaluated one of the best models and one of the worst models in Expert MoBiPlant (Figure 5.b, Supplementary Table 2). Overall, we see that both models have a bias towards option A (although DeepSeek R1 seems to be more robust to these shuffles while still showing flaws). We extend on this by denoting: 1) accuracy tops up when the correct option is set on A (even surpassing the vanilla Expert MoBiPlant performance) and that when the correct option is not on A, accuracy consistently drops (Supplementary Table 2); 2) model's choice tends to be A even when it is not the correct option; on the second (all corrects are B) and third (all corrects are C) versions Gemini chooses three

times more the option A than its other wrong counterpart (while DeepSeek shows a less pronounced tendency). This settles the ground for posterior discussion about complementing MCQ evaluation with robust human reviewing.



**Figure 5. LLMs pose option bias on Expert MoBiPlant. a)** The mean proportion of “A” responses by each LLM on the Expert MoBiPlant set over the three random shuffles. Standard deviations are denoted as error bars for each model. The distribution of correct “A” answers (%33,33) is highlighted with a dotted line. **b)** We created three alternate versions of MoBiPlant, where for each we moved all the correct answers to a specific position (the first had all correct responses transferred to “A”, the second to “B” and the third to “C”). We show the distribution of answers for these biased datasets for Gemini (left) and DeepSeek R1 (right).

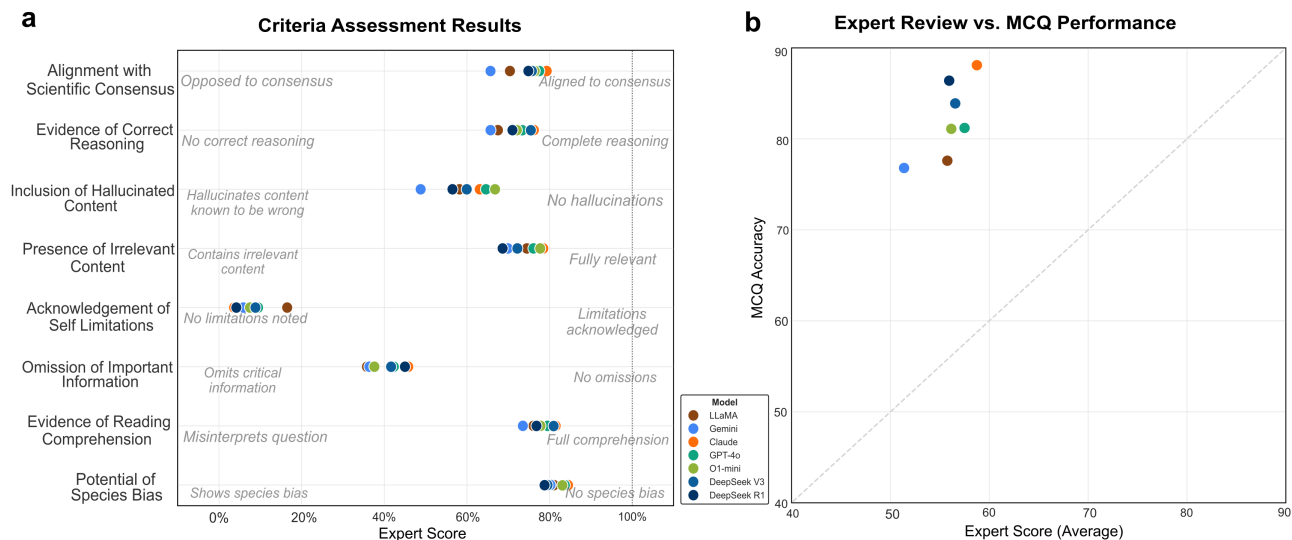
## 2.5 Expert Reviews Uncover Key Gaps In LLM-generated Answers

To perform a more in-depth analysis of LLMs strengths and failure modes in plant molecular biology understanding, we prompted the models to produce open-ended answers to the Expert MoBiPlant set. Performance was assessed based on a set of criteria inspired by previous LLMs human evaluation methodologies in the medical domain<sup>25,26</sup> (see Methods 5.4 for a detailed description of the human annotation procedure).

The open-ended design provides critical insights that structured benchmarks might obscure: it reveals how models construct biological arguments rather than simply selecting answers, exposing tendencies toward hallucination in underspecified scenarios and latent biases in phylum knowledge.

After the Consortium had reviewed the outputs, we created a numeric mapping that assigned each possible review a value from 0 to 100, where the values closer to zero indicate bad habits (e.g. hallucinated content) and the values near 100 indicate good outcomes (e.g. no hallucinated content in answer) on a given criterion (Figure 6.a). Remarkably, all models exhibit strong





**Figure 6. Expert evaluation on open ended answers.** **a)** Human-judged scores across eight evaluation criteria. Each dot represents a model's score (0–100%) on one criterion, with higher values indicating more positive performance (e.g., more complete reasoning, fewer hallucinations). **b)** Model's mean human-evaluation score (x-axis, macro averaged across criteria) against its automatic evaluation accuracy (y-axis). The Pearson correlation for these scores is 0.71, demonstrating positive correlation between automatic and human-based assessments (p-value: 0.075).

reading comprehension and maintain high topical relevance. Evidence of reading comprehension ranges from 74.0% (Gemini) to 80.6% (Claude), demonstrating that these models reliably interpret specialized prompts and extract critical details. Likewise, presence of irrelevant content scores falls between 69.1% (DeepSeek R1) and 78.2% (Claude), indicating that most generated text remains on point (higher scores indicate less irrelevant content). Potential species bias is similarly low, with values from 80.1% (DeepSeek R1) up to 83.6% (Claude), signifying broadly applicable knowledge across taxa on MOBIPLANT (higher scores indicate less species-bias).

In contrast, performance on factual alignment, reasoning, hallucination avoidance, and information completeness remains more moderate. Alignment with scientific consensus spans 65.7% (Gemini) to 79.3% (Claude), showing occasional deviations from established plant-science consensus, where applicable. Evidence of correct reasoning ranges from 65.7% (Gemini) to 73.9% (Claude and DeepSeek V3), indicating that causal or logical justification remains imperfect. Models score between 48.8% (Gemini) and 66.8% (o1-mini) on inclusion of hallucinated content—i.e., the proportion of responses free from fabricated claims—and only 35.8% (LLaMA) to 45.8% (Claude) on omission of important information, meaning critical details are frequently missing (lower scores indicate more omission of important information).

Remarkably, self-awareness is uniformly low across alternative LLMs: acknowledgement of limitations ranges from just 4.2% (Claude) to 16.5% (LLaMA). Interestingly, all models scored on an interval of 4.2% to 8.3% with exception of LLaMA with a remarkable 16.5%, showing stronger self-awareness in comparison to its competitors. In scientific settings, this overconfidence risks the unqualified presentation of uncertain or speculative material, skewing and potentially harming research if used without caution.

Although based on different scoring frameworks, we compared automatic evaluation accuracy based on MCQs, with human assessment scores of open-ended answers (macro-averaged across criteria; Figure 6.b). The two approaches show a positive Pearson correlation of 0.71 (p-value = 0.075), highlighting both their alignment and complementary nature in evaluating model performance. Moreover, complementary evaluations gain prominence when judging LLM's capabilities within the plant biology domain. We show that additional axes of assessment, when conducted by experts within the field, discover significant defects of chat assistants that often go unnoticed when employing automated-only evaluation rubrics, such as MCQ.

## 2.6 Persistent Misconceptions and Hallucinations Limit LLMs In Plant Science Applications

Several notable failure modes emerged during open-ended assessments of model performance in plant biology (Table 1). *Cross-domain bias* was observed, with models incorrectly applying mechanisms from animal or human biology to plant systems. *Species confusion* occurred when models conflated traits or gene functions across distinct plant species, leading to biologically inaccurate generalizations. This is common when extrapolating well characterized mechanisms in model organisms (e.g.

*Arabidopsis thaliana*) to other species. Errors in *acronym interpretation* were also common, where plant-specific genes were mistaken by completely different genes but similar in characters. Additionally, *terminology gaps* were evident, misunderstanding specialized plant physiology terms. The models often relied on *outdated knowledge*, failing to incorporate recent advances in emerging areas. In several instances, models *hallucinated molecular pathways*, combining unrelated pathways or involving incorrect genes in a given molecular mechanism. Finally, there was a marked *environmental misunderstanding*, with the models struggling to reason about field conditions, stress responses, such as the complex temperature- or drought-related physiological adaptations. Collectively, these failure patterns highlight key challenges for the accurate deployment of LLMs in plant-specific biological contexts.

**Table 1. Examples of common failure modes sighted on the open-ended assessment of Expert MoBiPlant.** Additional examples can be examined in Supplementary Table 3.

Error	Model	Prompt	Model response	Error explanation
Cross-domain bias & Species confusion	Claude 3.5 Sonnet	What types of DNA motifs bound by transcription factors are enriched at the boundaries of TADs in plant species?	In plants, particularly in <i>Arabidopsis</i> , CTCF-like binding motifs and certain transcription factor binding sites are enriched at TAD boundaries, though the specific motifs differ from those found in animals. Notable enriched motifs include TCP-binding sites and AT-rich sequences recognized by AT-hook proteins. The plant-specific transcription factor TCP family appears to play an important role in TAD boundary formation, similar to how CTCF functions in animals, though the mechanism is not identical. Additionally, active gene regulatory elements and certain epigenetic marks are frequently found at these boundaries.	TAD-like domains have rarely been observed in <i>Arabidopsis</i> . The insulator protein CTCF required for TAD/loop formation in animals is not found in plants.
		How is hyponasty regulated in <i>Arabidopsis thaliana</i> ?	Hyponastic growth in <i>Arabidopsis thaliana</i> , the upward bending of leaves, is primarily regulated by the plant hormone ethylene. Ethylene signaling promotes asymmetrical cell elongation on the adaxial (upper) side of the petiole, leading to the upward curvature. This process is influenced by various factors including light, submergence, and other hormones like auxin and gibberellins, which can interact with ethylene signaling pathways. Key components in ethylene-mediated hyponasty include ethylene receptors, CTR1, EIN2, EIN3/EIL1, and downstream targets like ERFs that regulate gene expression related to cell expansion.	The hyponasty response is upward leaf movement driven by higher cell expansion on the abaxial side vs. the adaxial. This difference depends on auxin biosynthesis, transport, and distribution, regulated by R/FR light and temperature.
Hallucinated pathways & Terminology gaps	LLaMA 3.1 405B	What proteins have been identified as molecular partners of the <i>Arabidopsis</i> lncRNA ASCO? In what molecular mechanism are they all involved in?	The <i>Arabidopsis</i> lncRNA ASCO has been identified to interact with the protein complex composed of the nuclear speckle RNA-binding protein (NSR), alternative splicing regulator SR45, serine/arginine-rich splicing factor 45a (SR45a), and the splice factor U2AF65b. All of these proteins are involved in the regulation of alternative splicing. ASCO mediates the interaction between NSR and SR45, promoting the alternative splicing of NSR-targeted pre-mRNAs.	The lncRNA ASCO interacts with splicing factors but not with SR42 nor U2AF65b. Furthermore, ASCO has not been shown to mediate the interaction between NSR and SR45.
Cross-domain bias & Hallucinated pathways	GPT-4o	In <i>Medicago truncatula</i> , GA suppression of infection involves the degradation of which proteins acting in the GA signaling pathway?	In <i>Medicago truncatula</i> , gibberellin (GA) suppression of infection involves the degradation of DELLA proteins, which act as repressors in the GA signaling pathway. DELLA proteins are negative regulators that inhibit growth and development processes, and their degradation by the GA signaling mechanism allows for the promotion of growth and alleviation of the suppression caused by stress responses, including pathogen infection. The action of GA in degrading these proteins helps facilitate effective plant-microbe interactions, particularly in symbiotic associations.	Mixes results across species and pathways; DELLA proteins can have positive effects in specific contexts. GA-mediated DELLA degradation typically reduces infection.

Continued on next page

Error	Model	Prompt	Model response	Error explanation
<b>Acronym misinterpretation &amp; Species confusion &amp; Hallucinated pathways</b>	DeepSeek-R1	Which processes are regulated by the transcription factor MIB2 during thermomorphogenesis in tomato?	The transcription factor MIB2 (MYB-INTERACTING bHLH 2) in tomato regulates thermomorphogenesis by modulating auxin biosynthesis and signaling under high temperature. It directly activates YUCCA3 (YUC3) and auxin-responsive genes (e.g., <i>IAA19</i> ), enhancing cell elongation and stem growth. MIB2 also interacts with heat shock factors to coordinate thermoresponsive growth, balancing heat adaptation with stress tolerance, facilitating hypocotyl elongation and thermomorphogenic responses while mitigating oxidative stress.	MULTIPLE INFLORESCENCE BRANCH 2 (MIB2) modulates inflorescence branching in response to high temperature in tomato. The answer extrapolates <i>Arabidopsis</i> pathways and claims HSP promoter binding not supported by evidence.
	o1-mini	Which lncRNA is involved in the thermomorphogenic response in <i>Arabidopsis thaliana</i> and how?	The long noncoding RNA ELF3-AS is involved in the thermomorphogenic response in <i>Arabidopsis thaliana</i> by regulating the expression of <i>ELF3</i> . ELF3-AS modulates <i>ELF3</i> activity, influencing growth and development under elevated temperatures.	ELF3-AS1 is an oncogenic lncRNA in bladder cancer (human context). In <i>Arabidopsis</i> , high temperatures are sensed via thermo-inducible phase separation of EARLY FLOWERING 3 (ELF3), which is a protein, not a lncRNA.

It has been previously reported that LLMs often struggle to provide accurate references when responding to scientific queries<sup>27</sup>, such as correctly identifying citation attributes like Title or Authors. In the Expert MoBiPlant set, members of the Consortium included the correct reference corresponding to the most reliable article addressing each question. During the evaluation phase, we prompted each LLM to provide a supporting reference for its answer. Although these references were reviewed by experts to assess their accuracy, the diversity of errors made a systematic evaluation challenging. In most cases, the DOI and title provided by the models were either incorrect or entirely fictitious. As illustrated in Supplementary Table 4, even for the most cited article in the dataset—and accordingly, showing remarkably good performance by LLMs delivering the related reference—the articles referred by the models exhibited various types of errors. These included: (i) the correct title paired with an unrelated DOI; (ii) the correct title with a non-existent DOI; (iii) a fabricated title with a fabricated DOI; among other misleading combinations.

## 3 Discussion

### 3.1 Domain-specific Benchmarking Is Critical For Assessing LLM Performance

As mentioned above, LLMs have been extensively evaluated in several high-stakes domains, with the medical field being one of the most prominent examples<sup>25,26</sup>. In fact, with domain-specific training, LLMs have reached or even surpassed the passing thresholds on multiple-choice exams designed for human medical students, demonstrating their ability to retrieve and reason over well-established, textbook-based knowledge. However, the situation is markedly different in the domain of plant molecular biology. While basic concepts in botany or elementary plant molecular biology may be found in standard textbooks (somewhat analogous to consolidated anatomical knowledge in medicine), the majority of relevant and up-to-date information resides in scientific literature, such as peer-reviewed research papers. This calls for a dedicated, domain-specific evaluation frameworks in plant molecular biology like MoBiPlant. Accurately assessing LLM performance in this field requires the involvement of subject-matter experts capable of judging the model's capacity to understand, synthesize, and apply cutting-edge research findings.

### 3.2 Automatic MCQ Evaluation And Expert Human Analysis Offer Complementary -Yet Correlated- Insights

Our results reveal that while state-of-the-art models achieve high multiple-choice accuracy (>75%) and strong comprehension in plant biology (74–81 scoring points), they still suffer from moderate factual misalignment (66–79 scoring points), frequent hallucinations (49–67 scoring points), and poor self-awareness (<17 scoring points). Human evaluations correlate positively with automatic MCQ scores while core model behaviours are addressed: open-ended judgments of reasoning, omission, and content relevance are substantially lower (Figure 6.b).

MCQ formats provide straightforward accuracy measures but can mask potential hallucination, lack of self-awareness and model biases. Additionally, uncovered failure modes highlight the risk of open-ended queries. Recent work shows that MCQ evaluation often underestimates open-ended capabilities and suffers from “first-token” misalignments between predicted option letters and actual text answers<sup>28</sup>. Moreover, relying solely on MCQ for evaluation risks “fake alignment” where high scores may hide poor factual grounding when models generate unconstrained text<sup>29</sup>. The experiment conducted on the previous section, where we show models election bias on plant molecular biology questions, sets the need for robust supporting evaluation frameworks.

Automated metrics moderately correlate with human judgments in knowledge-intensive settings. In our plant-biology domain, models are clustered by MCQ performance but scattered widely in expert criteria (hallucinations, reasoning, omissions), underscoring the need for

complementary human assessments or more nuanced automated proxies. Moreover, we show that human evaluation provides richer and complementary information to that obtained through MCQ benchmarks, as models with similar MCQ scores can differ by 10–20 scoring points in expert-judged performance when it comes to assessments of reasoning quality, omission rates, and content relevance for example.

### 3.3 LLM Understanding Scales With Paper Citations

We observe a pronounced uplift in model accuracy for questions derived from highly cited sources (501–1,702 citations) versus sparsely cited works—paralleling findings that over-represented data drive LLM competencies<sup>30</sup>. These effects hint at the influence of training-data composition: LLMs internalize frequent patterns but struggle with under-represented or recent material.

Crucially, performance is heavily skewed by training-data biases: by utilizing Synthetic MoBiPlant we show that questions originated from review articles consistently pose higher accuracy scores than those from primary research articles (Figure 4), highlighting the fact that review articles discuss well-established ideas within the field that are more likely to be reflected in the models’ training data and thus more readily recognized.

Although general species-bias scores remain high (80–84%, high scores indicating low bias), critical omissions (36–46%) and hallucinations (49–67%) indicate uneven coverage of specialized plant-biology knowledge. Domain benchmarks like BioLLMBench confirm similar gaps in bioinformatics tasks<sup>31</sup>, motivating targeted corpus expansion.

## 4 Conclusions

LLMs are increasingly being used by a broad audience—from high school and undergraduate students to early-career and established researchers. While their applications continue to expand, the tendency to rely on LLMs as a substitute for critical document-based research—receiving synthesized answers presented as factual—poses significant risks. In the field of plant molecular biology, in particular, careful validation of LLM-generated content is still needed to prevent the dissemination of misconceptions and inaccurate information.

Our comprehensive benchmarking reveals that while current LLMs excel at parsing and responding to structured plant biology prompts—achieving high MCQ accuracy and strong reading comprehension—they consistently struggle with deeper scientific rigor and self-awareness. Factual alignment remains only moderate, hallucinated content persists at nontrivial rates, and essential details are often omitted. Crucially, all models exhibit a pronounced inability to recognize or communicate their own uncertainties, risking the unqualified presentation of speculative or incomplete information. Interestingly, while ChatGPT (particularly GPT-4o) is by far the most widely adopted model among researchers in our consortium—as reported in our survey of preferred assistants—it does not consistently rank as the top performer in our benchmark evaluations. Although it achieves solid results in both MCQ and open-ended formats, models such as Claude 3.5 Sonnet and DeepSeek R1 tend to outperform it across key evaluation dimensions. Furthermore, Claude 3.5 Sonnet and O1-mini were the only LLMs assessed that did not exhibit any bias towards option A in MCQs.

Our findings underscore the limitations of evaluating LLMs solely on closed-form tasks and highlight the value of expert human assessment in revealing hidden failure modes. To advance toward trustworthy scientific assistants, it will be crucial to integrate retrieval augmentation or search tools in LLM-based agentic workflows to ground outputs in authoritative plant databases, applying domain-specific fine-tuning to reinforce structured knowledge, and incorporating uncertainty calibration mechanisms that enable models to flag low-confidence assertions.

By adopting this multi-pronged strategy, future models can not only bolster their factual accuracy and completeness but also develop the capacity necessary for safe and reliable deployment in research scenarios. The path forward lies in blending rigorous evaluation frameworks with targeted architectural enhancements to realize the full potential of chat assistants in plant biology.

## 5 Methods

### 5.1 Dataset Design and Curation

The construction of the MoBiPlant Benchmark involved the creation of a consortium of 112 researchers in plant molecular biology (the MoBiPlant Consortium), selected to represent diverse geographical regions, career stages, institutional affiliations, and demographic backgrounds. Contributors spanned disciplines encompassing molecular mechanisms (gene regulation, genome and genomics, cell biology and signaling), organismal processes (growth and development, hormones, physiology and metabolism, evolution), and applied domains (biotechnology, environment). Each expert authored five multiple-choice questions (MCQs) aligned with their specialization, adhering to rigorous guidelines ensuring question clarity, plausibility of distractors, and balanced option lengths to mitigate selection biases. A complete protocol detailing question design principles is provided in Supplementary Material Section F.

To elevate the dataset size, we further boosted question count using a frontier LLM which was prompted to generate 1075 additional question-answer pairs from human-selected papers. We detail the extraction, processing and creation in Section 5.5.

Each question in MoBiPlant includes rich metadata detailing the plant species involved, the specific sub-discipline it relates to, and the source publication where the correct answer can be verified (with the exception of eleven more general questions, where their authors claimed that multiple sources were required). These fields enhance the utility and transparency of the benchmark by enabling topic-based filtering, ensuring traceability to original sources, and offering a clearer view of the content’s scope. Figure 2.b-d presents the distribution of these attributes, illustrating the representation of species, topic areas, and source diversity across the dataset.

### 5.2 Model Selection

To conduct this evaluation, we selected a mixture of open and closed models, excelling in several general benchmarks<sup>22,32,33</sup> and validated by the MoBiPlant Consortium top chat-assistant choices (Figure 1.g). There are several biologically tuned language models centered in



biomedical knowledge<sup>2,34-37</sup>, but they are comparatively small to the models widely adopted as first-hand large chat assistants by the community. This is why we focused on benchmarking the following selection comprising GPT-4o<sup>16</sup>, Claude 3.5 Sonnet<sup>21</sup>, Llama 3.1 405B<sup>20</sup>, Gemini 1.5 Pro<sup>15</sup> and Deepseek V3<sup>18</sup> as well as recently released reasoning models, featuring o1-mini<sup>17</sup> and Deepseek R1<sup>19</sup>.

All models were instantiated with a value of 0.7 for temperature and a max generation length of 4096 tokens to ensure an equal evaluation environment and mimic that of assistant usage settings.

### 5.3 Automated MCQ Evaluation

To extensively assess model performance, we employ two parallel evaluation designs: Automated MCQ evaluation and Human Evaluation (see next section for a detailed description of the Human Evaluation Protocol).

For automatic evaluation we directly computed the accuracy of model answers over the set of MCQ questions. This method presents models with questions and predefined answer choices, followed by automated extraction of their selections for comparison against ground-truth answers. While this approach offers advantages—such as providing a structured framework for systematic evaluation and simplifying answer extraction by constraining responses to predefined choices—it introduces three main limitations: 1) exploitation of shortcuts/biases<sup>38</sup>: models may exploit statistical patterns or suboptimal distractors (e.g. option order) to guess correct answers without genuine comprehension, artificially inflating performance metrics; 2) limited real-world applicability<sup>39,40</sup>: MCQ formats may inadequately reflect real-world applications, where open-ended queries are the norm rather than constrained selections; and 3) choice-order sensitivity<sup>41</sup>: performance can vary based on the order of answer choices, introducing inconsistencies unrelated to model capability.

To address concerns of model over-reliance on statistical patterns and choice-order sensitivity, we generated three permutations of the MCQ dataset, each containing questions with randomly shuffled answer options. By evaluating model performance across all permutations, we quantified variability in accuracy attributable to option order using the standard deviation of scores between shuffled versions. This approach directly measures robustness to positional bias while ensuring reproducibility across stochastic configurations.

We designed a chain-of-thought prompt<sup>42</sup> to elicit reasoning in model output generation. The prompt instructs the model to reason through the correct answer in order to enhance answer generation. This approach is natural and aligns the real-world application of MCQs, encouraging the model to generate a step-by-step reasoning before pointing to the final answer between `<answer>` tags, which is later extracted using string-matching techniques. Since this methodology introduces possible parsing errors in model generation we report the valid accuracy when scoring with this method. This is, the accuracy over all examples in which a model answer could be parsed.

### 5.4 Human Evaluation

To address the limitations of automated multiple-choice assessments, we implemented a human expert evaluation protocol adapted from existing clinical knowledge language model validation frameworks<sup>25,26</sup>. This approach begins with a priori definition of domain-specific evaluation criteria, developed collaboratively with plant molecular biologists to reflect field-specific priorities. Two core criteria were established in addition to the existing: (1) Species bias, assessing whether model answers overgeneralize findings beyond the queried species (e.g., attributing *Arabidopsis thaliana* mechanisms to non-model plants), and (2) Consensus alignment with established plant science dogmas and theories, evaluating agreement with peer-reviewed evidence while accounting for contested or evolving concepts. This resulted in a total of 8 criteria for human evaluation, as defined in Table 2.

For species bias, evaluators flagged answers containing unsupported taxonomic extrapolations, a critical concern when employing these systems in understudied species. Alignment with scientific consensus was scored on a ternary scale: alignment (consistent with established literature), opposition (contradicting peer-reviewed evidence), or lack-of-alignment (addressing unresolved or actively debated topics). This granular scoring acknowledges the dynamic nature of plant biology, where emerging methodologies (e.g., the advent of CRISPR-based gene editing) routinely challenge paradigms. The detailed description of every criterion with the available options that were provided to the expert reviewers is included in Table 2.

For evaluating model performance in open-ended contexts mirroring real-world scientific inquiry, we generated responses to all MoBiPlant questions using the same prompts across models. Instructions explicitly directed models to provide concise, self-contained answers without supplying the predefined multiple-choice options. The idea was to simulate situations where users do not know the answers in advance, helping to reveal subtle behaviors like guessing, overgeneralizing, or missing details that MCQs often hide.

To ensure evaluator rigor, we implemented a blind workflow: model answers were anonymized, stripped of metadata (e.g., model names, question area), and randomly shuffled across questions. Crucially, since experts in one area may not have the detailed knowledge needed to evaluate work in another, each contributor evaluated responses only to their own original questions. We used the SuperAnnotate Platform<sup>43</sup> to coordinate the annotation process involving the 112 researchers from the MoBiPlant Consortium.

To standardize the manual evaluation protocol, we established a normalized scoring framework. Qualitative criteria (e.g., species bias, consensus alignment) were mapped to normalized numerical scores ranging from 0 (indicating undesirable performance) to 1 (indicating optimal performance). This transformation preserves the interpretability of expert judgments while facilitating quantitative correlation analyses. For instance, responses contradicting scientific consensus were assigned a score of 0, whereas alignment with peer-reviewed evidence scored 1; ambiguous or context-dependent answers received intermediate values. Detailed mappings are provided alongside the criteria description in Table 2.

### 5.5 Synthetic Data Generation

The production of synthetic datasets has been of uprising use over the recent years<sup>44-50</sup>, with primarily focus on model distillation and model evaluation. In this case, we propose a framework for systematically extracting multiple-choice questions from a curated set of manually selected publications within plant molecular biology throughout the years 2001 to 2025.



**Table 2. Criteria set used for human evaluation on MOBIPLANT.**

Criterion	Description	Rating Options & Explanations	Score
<b>Evidence of Correct Reasoning</b>	Assesses whether the response follows logical reasoning to arrive at a correct conclusion.	<b>Yes:</b> The reasoning is logical, well-structured, and leads to a correct conclusion. <b>No:</b> The reasoning is flawed, contains logical fallacies, or does not support the conclusion.	Yes: <b>1</b> No: <b>0</b>
<b>Evidence of Reading Comprehension</b>	Evaluates whether the response fully understands and addresses the question, including its context and nuances.	<b>Yes:</b> Fully comprehends and correctly addresses the entire question. <b>Partially:</b> Shows partial understanding but misses important details. <b>No:</b> Misinterprets or ignores key aspects of the question.	Yes: <b>1</b> Partially: <b>0.5</b> No: <b>0</b>
<b>Alignment with Scientific Consensus</b>	Determines if the response aligns with the widely accepted scientific understanding.	<b>Aligned to consensus:</b> Reflects the prevailing scientific view, dogmas and/or theories. <b>Opposed to consensus:</b> Contradicts widely accepted scientific knowledge. <b>No consensus:</b> The topic lacks a clear, widely agreed-upon stance.	Aligned: <b>1</b> Opposed: <b>0</b> No consensus: (don't compute)
<b>Omission of Important Information</b>	Checks if critical biological details are missing, which could affect the accuracy of the response.	<b>Yes, great biological significance:</b> Major omission that alters understanding. <b>Yes, little biological significance:</b> Minor omission with limited impact. <b>No:</b> Includes all necessary information.	Yes, great: <b>0</b> Yes, little: <b>0.5</b> No: <b>1</b>
<b>Inclusion of Hallucinated Content</b>	Identifies whether the response contains incorrect or unsupported information.	<b>Yes, known wrong:</b> Claims fabricated/incorrect information. <b>Yes, unverifiable:</b> Includes unverifiable claims. <b>No:</b> Based on verifiable information.	Yes, known wrong: <b>0</b> Yes, unverifiable: <b>0.5</b> No: <b>1</b>
<b>Inclusion of Irrelevant Content</b>	Determines if the response includes off-topic or unnecessary details.	<b>Yes:</b> Contains extraneous information. <b>No:</b> Stays focused on the question.	Yes: <b>0</b> No: <b>1</b>
<b>Potential of Species Bias</b>	Assesses whether the response shows bias toward or against specific species in a biological context.	<b>Yes:</b> Exhibits species bias (e.g., general claim untrue across species, or answers about a species using other species). <b>No:</b> Does not contain biased claims. <b>Not applicable:</b> The question does not involve species comparisons.	Yes: <b>1</b> No: <b>0</b> Not applicable: (don't compute)
<b>Acknowledgement of Limitations</b>	Evaluates whether the response recognizes uncertainties or limitations where relevant.	<b>Yes:</b> Explicitly acknowledges relevant self-limitations. <b>No:</b> Omits relevant uncertainties despite evident limitations. <b>Not applicable:</b> The answer does not require discussing limitations.	Yes: <b>1</b> No: <b>0</b> Not applicable: (don't compute)

For the selection of published papers, a group of specialists selected subsets of relevant work documents in the following journals: (1) research articles from *New Phytologist*, *Plant Journal*, *Plant Physiology*, *The Plant Cell*, *Molecular Plant*, *Journal of Experimental Botany*, and *Nature Plants*; and (2) review articles from *Trends in Plant Science*, and *Current Opinion on Plant Biology* (see Supplementary Table 5: MoBiPlant Journal distribution). This curation ended with a total of 215 documents (at least one per year and journal, if already existing) with fine-grained information within the field.

Consequently, we prompt a frontier LLM (Gemini 2.5 Pro) to act as an MCQ extractor when a work is given as pdf. The process is similar to what is solicited to the human collaborators: generate 5 multiple-choice questions with one correct answer and two incorrect answers (distractors). The instructions intend to be as similar as possible to the guidelines prompted to humans (see Supplementary Material Section G). The idea was to ensure a clear and unambiguous question that is directly linked with a unique fact or conclusion from the given work, with its respective correct answer and distractors generated by tweaking the information on the correct answer in a way that keeps semantic and terminology-related feasibility. In addition, the instructions include the metadata tagging of each question (whether there is a plant species involved in the query and its respective area), ensuring compatibility with the human-generated MCQ. By generating 5 questions for every one of the 215 studies collected on the previous step, we ended up with 1075 additional multiple-choice questions tagged and compatible with the previous format, which constitute the Synthetic MoBiPlant dataset.

For the validation of this methodology, we randomly sampled 20 questions ensuring equal representation from all journals for human revision, which confirmed the suitability of the generated questions. This was performed by the same experts that proposed the studies.

## 6 Data and Code Availability

All datasets used in this study, Expert MoBiPlant and Synthetic MoBiPlant are available via HuggingFace ([manufernandezbur/MoBiPlant](https://huggingface.co/manufernandezbur/MoBiPlant)). The code used to process data, generate datasets and perform the analyses described in this paper is available at [GitHub](https://github.com/manufernandezbur/MoBiPlant), including model instantiation, benchmarking and data analysis. Additional instructions for reproducing the results are provided in the repository's README file.

## 7 Ethical Considerations

This study evaluates the capabilities and limitations of large language models (LLMs) in the context of plant molecular biology using a benchmark developed by a global consortium of domain experts. The research was conducted in alignment with ethical principles of transparency, fairness, and inclusivity.

**Expert Contributions and Consent:** The MoBiPlant Benchmark was created with the voluntary participation of 112 researchers across 19 countries. All contributors agreed to participate and were fully aware of the project's goals.

**Diversity and Representation:** The study prioritized inclusive participation, ensuring representation across gender identities, geographic regions, institutional affiliations, and career stages. Demographic information was collected and reported in aggregate form (Figure 1) to inform on dataset composition and potential biases.

**Synthetic Data Use:** The study includes synthetic data generated using a frontier LLM. These data were derived from publicly available scientific literature and used strictly for evaluation purposes. A human-in-the-loop validation process was implemented to mitigate the risk of embedding flawed or biased information. No copyrighted or proprietary content was used beyond what is permissible under fair use for scientific research.

**Model Evaluation and Bias Awareness:** Evaluated models are commercial and open-source LLMs that operate as general-purpose assistants. We acknowledge that these models may reflect biases inherent in their training data, including overrepresentation of certain species, geographic regions, or scientific paradigms. As such, our evaluation explicitly includes metrics for species bias and alignment with scientific consensus. We emphasize that LLM outputs should not be regarded as authoritative or error-free, particularly in understudied taxa or emergent research areas.

**Use of AI in Scientific Assessment:** We recognize the ethical implications of using AI-generated outputs in scientific settings. While LLMs offer utility in accelerating literature review and hypothesis generation, our findings underscore the continued need for expert oversight. To prevent misuse, we do not advocate for the replacement of domain expertise with AI, but rather for informed, cautious integration of these tools under appropriate validation frameworks.

**Transparency and Reproducibility:** All prompts, evaluation protocols, and scoring rubrics are provided in the Methods section and Supplementary Materials to support transparency and reproducibility. The MoBiPlant dataset is made freely available for academic research under an open license, enabling community-led extensions and audits.

## 8 Acknowledgements

The FA lab is funded by AXA Research Fund, ICGEB, Agencia I+D+i, and the IRP NOCOSYM (CNRS). The EF lab is supported by Agencia I+D+i, a Human Frontiers grant, the Google Award for Inclusion Research and a Googler Initiated Grant. MFB was supported by the AI Safety Argentina Scholarship (AISAR) program. APOLO Biotech covered the computing expenses. We thank SuperAnnotate for giving us access to their annotation interface.

## References

1. Peppin, A. *et al.* The reality of AI and biorisk, DOI: [10.48550/arXiv.2412.01946](https://doi.org/10.48550/arXiv.2412.01946). [2412.01946\[cs\]](https://arxiv.org/abs/2412.01946).
2. Simon, E., Swanson, K. & Zou, J. Language models for biological research: a primer. *Nat. Methods* **21**, 1422–1429, DOI: [10.1038/s41592-024-02354-y](https://doi.org/10.1038/s41592-024-02354-y). Publisher: Nature Publishing Group.
3. Gougherty, A. V. & Clipp, H. L. Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. *npj Biodivers.* **3**, 13, DOI: [10.1038/s44185-024-00043-9](https://doi.org/10.1038/s44185-024-00043-9).
4. Wang, X. *et al.* SciBench: Evaluating college-level scientific problem-solving abilities of large language models, DOI: [10.48550/arXiv.2307.10635](https://doi.org/10.48550/arXiv.2307.10635). [2307.10635\[cs\]](https://arxiv.org/abs/2307.10635).
5. Laurent, J. M. *et al.* LAB-bench: Measuring capabilities of language models for biology research, DOI: [10.48550/arXiv.2407.10362](https://doi.org/10.48550/arXiv.2407.10362). [2407.10362\[cs\]](https://arxiv.org/abs/2407.10362).
6. Luo, X. *et al.* Large language models surpass human experts in predicting neuroscience results. *Nat. Hum. Behav.* **9**, 305–315, DOI: [10.1038/s41562-024-02046-9](https://doi.org/10.1038/s41562-024-02046-9).
7. Naddaf, M. How are researchers using AI? survey reveals pros and cons for science. *Nature* DOI: [10.1038/d41586-025-00343-5](https://doi.org/10.1038/d41586-025-00343-5). Bandiera\_abtest: a Cg\_type: News Publisher: Nature Publishing Group Subject\_term: Machine learning, Publishing, Scientific community.
8. Skarlinski, M. D. *et al.* Language agents achieve superhuman synthesis of scientific knowledge, DOI: [10.48550/arXiv.2409.13740](https://doi.org/10.48550/arXiv.2409.13740). [2409.13740\[cs\]](https://arxiv.org/abs/2409.13740).
9. Hatch, V. BioChatter: making large language models accessible for biomedical research.
10. Didwania, K., Seth, P., Kasliwal, A. & Agarwal, A. AgriLLM: Harnessing transformers for farmer queries, DOI: [10.48550/arXiv.2407.04721](https://doi.org/10.48550/arXiv.2407.04721). [2407.04721\[cs\]](https://arxiv.org/abs/2407.04721).
11. Geitmann, A. & Bidhendi, A. J. Plant blindness and diversity in AI language models. *Trends Plant Sci.* **28**, 1095–1097, DOI: [10.1016/j.tplants.2023.06.016](https://doi.org/10.1016/j.tplants.2023.06.016).
12. Wang, J., Ye, Q., Liu, L., Guo, N. L. & Hu, G. Scientific figures interpreted by ChatGPT: strengths in plot recognition and limits in color perception. *npj Precis. Oncol.* **8**, 1–8, DOI: [10.1038/s41698-024-00576-z](https://doi.org/10.1038/s41698-024-00576-z). Publisher: Nature Publishing Group.
13. Kalmer, T. L. *et al.* Assessing the ability of ChatGPT to extract natural product bioactivity and biosynthesis data from publications, DOI: [10.1101/2024.08.01.606186](https://doi.org/10.1101/2024.08.01.606186). Pages: 2024.08.01.606186 Section: New Results.
14. Griot, M., Hemptinne, C., Vanderdonck, J. & Yuksel, D. Large language models lack essential metacognition for reliable medical reasoning. *Nat. Commun.* **16**, 642, DOI: [10.1038/s41467-024-55628-6](https://doi.org/10.1038/s41467-024-55628-6).
15. Team, G. *et al.* Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, DOI: [10.48550/arXiv.2403.05530](https://doi.org/10.48550/arXiv.2403.05530). [2403.05530\[cs\]](https://arxiv.org/abs/2403.05530).
16. OpenAI *et al.* GPT-4 technical report, DOI: [10.48550/arXiv.2303.08774](https://doi.org/10.48550/arXiv.2303.08774). [2303.08774\[cs\]](https://arxiv.org/abs/2303.08774).
17. OpenAI *et al.* OpenAI o1 system card, DOI: [10.48550/arXiv.2412.16720](https://doi.org/10.48550/arXiv.2412.16720). [2412.16720\[cs\]](https://arxiv.org/abs/2412.16720).
18. DeepSeek-AI *et al.* Deepseek-v3 technical report, DOI: [10.48550/arXiv.2412.19437](https://doi.org/10.48550/arXiv.2412.19437). [2412.19437\[cs\]](https://arxiv.org/abs/2412.19437).
19. DeepSeek-AI *et al.* DeepSeek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning, DOI: [10.48550/arXiv.2501.12948](https://doi.org/10.48550/arXiv.2501.12948). [2501.12948\[cs\]](https://arxiv.org/abs/2501.12948).
20. Grattafiori, A. *et al.* The llama 3 herd of models, DOI: [10.48550/arXiv.2407.21783](https://doi.org/10.48550/arXiv.2407.21783). [2407.21783\[cs\]](https://arxiv.org/abs/2407.21783).
21. Team, A. The claude 3 model family: Opus, sonnet, haiku.
22. Rein, D. *et al.* GPQA: A graduate-level google-proof q&a benchmark, DOI: [10.48550/arXiv.2311.12022](https://doi.org/10.48550/arXiv.2311.12022). [2311.12022\[cs\]](https://arxiv.org/abs/2311.12022).
23. Zheng, C., Zhou, H., Meng, F., Zhou, J. & Huang, M. Large language models are not robust multiple choice selectors.
24. Pezeshkpour, P. & Hruschka, E. Large language models sensitivity to the order of options in multiple-choice questions. In Duh, K., Gomez, H. & Bethard, S. (eds.) *Findings of the Association for Computational Linguistics: NAACL 2024*, 2006–2017, DOI: [10.18653/v1/2024.findings-naacl.130](https://doi.org/10.18653/v1/2024.findings-naacl.130) (Association for Computational Linguistics).
25. Feng, S. Y., Khetan, V., Sacaleanu, B., Gershman, A. & Hovy, E. CHARD: Clinical health-aware reasoning across dimensions for text generation models, DOI: [10.48550/arXiv.2210.04191](https://doi.org/10.48550/arXiv.2210.04191). [2210.04191\[cs\]](https://arxiv.org/abs/2210.04191).
26. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180, DOI: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2).
27. Byun, C., Vasicek, P. & Seppi, K. This reference does not exist: An exploration of LLM citation accuracy and relevance. In Blodgett, S. L. *et al.* (eds.) *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, 28–39, DOI: [10.18653/v1/2024.hcinlp-1.3](https://doi.org/10.18653/v1/2024.hcinlp-1.3) (Association for Computational Linguistics).
28. Wang, X. *et al.* "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models, DOI: [10.48550/arXiv.2402.14499](https://doi.org/10.48550/arXiv.2402.14499). [2402.14499\[cs\]](https://arxiv.org/abs/2402.14499).

29. Wang, Y. *et al.* Fake alignment: Are LLMs really aligned well?, DOI: [10.48550/arXiv.2311.05915](https://doi.org/10.48550/arXiv.2311.05915). [2311.05915\[cs\]](https://arxiv.org/abs/2311.05915).
30. Zevallos, R., Farrús, M. & Bel, N. Frequency balanced datasets lead to better language models. In Bouamor, H., Pino, J. & Bali, K. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2023*, 7859–7872, DOI: [10.18653/v1/2023.findings-emnlp.527](https://doi.org/10.18653/v1/2023.findings-emnlp.527) (Association for Computational Linguistics).
31. Sarwal, V. *et al.* A benchmark for large language models in bioinformatics, DOI: [10.1101/2023.12.19.572483](https://doi.org/10.1101/2023.12.19.572483). Pages: 2023.12.19.572483, Section: New Results.
32. Chiang, W.-L. *et al.* Chatbot arena: An open platform for evaluating LLMs by human preference, DOI: [10.48550/arXiv.2403.04132](https://doi.org/10.48550/arXiv.2403.04132). [2403.04132\[cs\]](https://arxiv.org/abs/2403.04132).
33. Wang, Y. *et al.* MMLU-pro: A more robust and challenging multi-task language understanding benchmark, DOI: [10.48550/arXiv.2406.01574](https://doi.org/10.48550/arXiv.2406.01574). [2406.01574\[cs\]](https://arxiv.org/abs/2406.01574).
34. Labrak, Y. *et al.* BioMistral: A collection of open-source pretrained large language models for medical domains, DOI: [10.48550/arXiv.2402.10373](https://doi.org/10.48550/arXiv.2402.10373). [2402.10373\[cs\]](https://arxiv.org/abs/2402.10373).
35. Luo, R. *et al.* BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinforma.* **23**, bbac409, DOI: [10.1093/bib/bbac409](https://doi.org/10.1093/bib/bbac409).
36. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240, DOI: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682). [1901.08746\[cs\]](https://arxiv.org/abs/1901.08746).
37. Yang, X., Gao, J., Xue, W. & Alexandersson, E. PLLaMa: An open-source large language model for plant science, DOI: [10.48550/arXiv.2401.01600](https://doi.org/10.48550/arXiv.2401.01600). [2401.01600\[cs\]](https://arxiv.org/abs/2401.01600).
38. Zheng, C., Zhou, H., Meng, F., Zhou, J. & Huang, M. On large language models' selection bias in multi-choice questions, DOI: [10.48550/arXiv.2309.03882](https://doi.org/10.48550/arXiv.2309.03882). Version: 2, [2309.03882\[cs\]](https://arxiv.org/abs/2309.03882).
39. Li, W. *et al.* Can multiple-choice questions really be useful in detecting the abilities of LLMs?, DOI: [10.48550/arXiv.2403.17752](https://doi.org/10.48550/arXiv.2403.17752). Version: 3, [2403.17752\[cs\]](https://arxiv.org/abs/2403.17752).
40. Singh, S. *et al.* It is too many options: Pitfalls of multiple-choice questions in generative AI and medical education, DOI: [10.48550/arXiv.2503.13508](https://doi.org/10.48550/arXiv.2503.13508). [2503.13508\[cs\]](https://arxiv.org/abs/2503.13508).
41. Guan, B., Roosta, T., Passban, P. & Rezagholizadeh, M. The order effect: Investigating prompt sensitivity in closed-source LLMs, DOI: [10.48550/arXiv.2502.04134](https://doi.org/10.48550/arXiv.2502.04134). [2502.04134\[cs\]](https://arxiv.org/abs/2502.04134).
42. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models, DOI: [10.48550/arXiv.2201.11903](https://doi.org/10.48550/arXiv.2201.11903). [2201.11903\[cs\]](https://arxiv.org/abs/2201.11903).
43. SuperAnnotate | streamline AI data workflows.
44. Nugroho, B., Widjaja, D., Wijaya, N. & Sutoyo, R. The feasibility of using synthetic-generated dataset for training sentiment analysis model. 1–6, DOI: [10.1109/ICITDA64560.2024.10810031](https://doi.org/10.1109/ICITDA64560.2024.10810031).
45. Arora, A. & Arora, A. Machine learning models trained on synthetic datasets of multiple sample sizes for the use of predicting blood pressure from clinical data in a national dataset. *PLOS ONE* **18**, e0283094, DOI: [10.1371/journal.pone.0283094](https://doi.org/10.1371/journal.pone.0283094).
46. Harsha, C., Phogat, K. S., Dasaratha, S., Puranam, S. A. & Ramakrishna, S. Synthetic data generation using large language models for financial question answering. In Chen, C.-C. *et al.* (eds.) *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, 76–95 (Association for Computational Linguistics).
47. Amin, K. *et al.* Private prediction for large-scale synthetic text generation, DOI: [10.48550/arXiv.2407.12108](https://doi.org/10.48550/arXiv.2407.12108). [2407.12108\[cs\]](https://arxiv.org/abs/2407.12108).
48. Braga, M., Kasela, P., Raganato, A. & Pasi, G. Synthetic data generation with large language models for personalized community question answering, DOI: [10.48550/arXiv.2410.22182](https://doi.org/10.48550/arXiv.2410.22182). Version: 1, [2410.22182\[cs\]](https://arxiv.org/abs/2410.22182).
49. Barr, A. A., Quan, J., Guo, E. & Sezgin, E. Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data. *Front. Artif. Intell.* **8**, 1533508, DOI: [10.3389/frai.2025.1533508](https://doi.org/10.3389/frai.2025.1533508). Publisher: Frontiers.
50. Kim, S. *et al.* Evaluating language models as synthetic data generators, DOI: [10.48550/arXiv.2412.03679](https://doi.org/10.48550/arXiv.2412.03679). [2412.03679\[cs\]](https://arxiv.org/abs/2412.03679).

## Supplementary Material

### A The MoBiPlant Consortium

**Supplementary Table 1. The MoBiPlant Consortium.** All authors that comprised the confection and evaluation of MOBIPLANT are listed with their corresponding institution(s).

Author's Name(s)	Author's Surname(s)	Author's Institution
José M.	Alvarez	Centro de Biotecnología Vegetal, Facultad de Ciencias de la Vida, Universidad Andrés Bello, 8370186 Santiago, Chile; 2 Agencia Nacional de Investigación y Desarrollo-Millennium Nucleus in Data Science for Plant Resilience (Phytolearning), 8370186 Santiago, Chile.
María Verónica	Arana	Instituto de Investigaciones Forestales y Agropecuarias Bariloche, Instituto Nacional de Tecnología Agropecuaria, Estación Experimental Bariloche – Consejo Nacional de Investigaciones Científicas y Técnicas (INTA EEA Bariloche-CONICET), San Carlos de Bariloche, Río Negro R8403DVZ, Argentina.
Laura	Arribas Hernández	Instituto de Hortofruticultura Subtropical y Mediterránea La Mayora (IHSM), Consejo Superior de Investigaciones Científicas - Universidad de Málaga (CSIC-UMA)
Carolina	Attallah	APOLO Biotech SAS; Facultad de Bioquímica y Ciencias Biológicas - Universidad Nacional del Litoral
Gabriela	Auge	Instituto de Agrobiotecnología y Biología Molecular (IABIMO), INTA-CONICET, Hurlingham (CP 1686), Argentina
Carlos L.	Ballaré	Instituto de Investigaciones Fisiológicas y Ecológicas Vinculadas a la Agricultura-Consejo Nacional de Investigaciones Científicas y Técnicas, Facultad de Agronomía, Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires C1417DSE, Argentina
Moussa	Benhamed	Université de Paris Cité, Institute of Plant Sciences Paris-Saclay (IPS2), 91190 Gif-sur-Yvette, France.
Matthias	Benoit	Laboratoire des Interactions Plantes-Microbes-Environnement (LIPME), INRAE, CNRS, Université de Toulouse, INRAE, CNRS, Castanet-Tolosan 31326, France
Flavio Antonio	Blanco	Instituto de Biotecnología y Biología Molecular, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Centro Científico y Tecnológico-La Plata, Consejo Nacional de Investigaciones Científicas y Técnicas, 1900-La Plata, Argentina
Nicolás E.	Blanco	Centre of Photosynthetic and Biochemical Studies (CEFOBI-CONICET), Faculty of Biochemical Science and Pharmacy, Rosario National University, S2002LRK Rosario, Argentina
Francisca	Blanco-Herrera	Centro de Biotecnología Vegetal, Facultad de Ciencias de la Vida. Universidad Andrés Bello
Miguel A.	Blazquez	Instituto de Biología Molecular y Celular de Plantas (CSIC-U, Valencia, SpainPV)
Thomas	Blein	Université Paris-Saclay, CNRS, INRAE, Université Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91190 Gif-sur-Yvette, France. 2 Université Paris Cité, CNRS, INRAE, Institute of Plant Sciences Paris-Saclay (IPS2), 91190 Gif-sur-Yvette, France.
Nicolas G.	Bologna	Centre for Research in Agricultural Genomics (CRAG)
Nicolas	Bouché	INRAE

*Continued on next page*



Author's Name(s)	Author's Surname(s)	Author's Institution
<b>Clara</b>	<b>Bourbousse</b>	Sorbonne Université, CNRS, INSERM, Development, Adaptation and Ageing, Dev2A, F-75005 Paris, France - Sorbonne Université, CNRS, Inserm, Institut de Biologie Paris-Seine, IBPS, F75005 Paris, France
<b>Wolfgang</b>	<b>Busch</b>	Salk Institute for Biological Studies
<b>Cristiane Paula Gomes</b>	<b>Calixto</b>	Institute of Biosciences, University of Sao Paulo
<b>Pablo Ignacio</b>	<b>Calzadilla</b>	Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, CEA, CNRS, Gif-sur-Yvette Cedex, France
<b>Alberto</b>	<b>Carbonell</b>	Instituto de Biología Molecular y Celular de Plantas
<b>Fernando</b>	<b>Carrari</b>	Instituto de Fisiología, Biología Molecular y Neurociencias-Buenos Aires (UBA-CONICET), Buenos Aires, Argentina. Cátedra de Genética. Facultad de Agronomía. Universidad de Buenos Aires, Argentina.
<b>Jorge</b>	<b>Casal</b>	Instituto de Investigaciones Fisiológicas y Ecológicas Vinculadas a la Agricultura, Facultad de Agronomía, Universidad de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas, 1417 Buenos Aires, Argentina 2 Fundacion Instituto Leloir, Instituto de Investigaciones Bioquímicas de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas, 1405 Buenos Aires, Argentina
<b>Paula</b>	<b>Casati</b>	Centro de Estudios Fotosintéticos y Bioquímicos (CEFOBI-CONICET/UNR)
<b>Pablo Diego</b>	<b>Cerdán</b>	Fundación Instituto Leloir, IIBBA-CONICET, Buenos Aires, Argentina; 2-Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina
<b>Uciel</b>	<b>Chorostecki</b>	Universitat Internacional de Catalunya
<b>Gabriela</b>	<b>Conti</b>	Instituto de Agrobiotecnología y Biología Molecular - INTA/CONICET. Argentina. Facultad de Agronomía, Universidad de Buenos Aires. Argentina.
<b>Pedro</b>	<b>Crevillén</b>	Centro de Biotecnología y Genómica de Plantas (CBGP), Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA/CSIC), Pozuelo de Alarcón (Madrid), Spain
<b>Carlos Daniel</b>	<b>Crocco</b>	Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Instituto de Investigaciones Fisiológicas y Ecológicas Vinculadas a la Agricultura (IFEVA), Facultad de Agronomía, Universidad de Buenos Aires, Buenos Aires, Argentina.
<b>Tibor</b>	<b>Csorba</b>	Hungarian University of Agriculture and Life Sciences
<b>Pilar</b>	<b>Cubas</b>	Centro Nacional de Biotecnología-CSIC
<b>Fernanda</b>	<b>de Carvalho-Niebel</b>	LIPME, INRAE, CNRS, Université de Toulouse, Castanet-Tolosan, France.
<b>Humberto</b>	<b>Debat</b>	National Institute of Agricultural Technology, Argentina. 2.University of Toronto, Canada.
<b>Carlos</b>	<b>Dezar</b>	CONICET - APOLO Biotech
<b>Maria Fabiana</b>	<b>Drincovich</b>	Centro de Estudios Fotosintéticos y Bioquímicos (CEFOBI), CONICET-Universidad Nacional de Rosario
<b>Jose Manuel</b>	<b>Estevez</b>	Fundación Instituto Leloir and IIBBA-CONICET. Av. Patricias Argentinas 435, Buenos Aires C1405BWE, Argentina. 2) 7ANID - Millennium Science Initiative Program - Millennium Institute for Integrative Biology (iBio) Santiago, Chile. 3) Centro de Biotecnología Vegetal, Facultad de Ciencias de la Vida, Universidad Andres Bello, Santiago, Chile.
<b>Carlos María</b>	<b>Figueroa</b>	Instituto de Agrobiotecnología del Litoral, UNL, CONICET, FBCB, Santa Fe, Argentina

*Continued on next page*

Author's Name(s)	Author's Surname(s)	Author's Institution
<b>Diego Fernando</b>	<b>Fiol</b>	Instituto de Investigaciones Biológicas, IIB-CONICET-Universidad Nacional de Mar Del Plata, Mar del Plata, Buenos Aires, Argentina
<b>Sotirios</b>	<b>Fragkostefanakis</b>	Institute of Molecular Biosciences, Goethe University Frankfurt, Frankfurt, Germany
<b>Crhristian</b>	<b>Fankhauser</b>	Center for Integrative Genomics, Faculty of Biology and Medicine, University of Lausanne
<b>Florian</b>	<b>Frugier</b>	Institute of Plant Sciences Paris-Saclay (IPS2), University Paris-Saclay, CNRS, INRAE University Paris-Cité, Univ. d'Evry, Gif-sur-Yvette, France
<b>Corina M.</b>	<b>Fusari</b>	Centro de Estudios Fotosintéticos y Bioquímicos (CEFOBI-CONICET-UNR), Suipacha 570, Rosario, S2000LRJ, Argentina
<b>Selma</b>	<b>Gago-Zachert</b>	Institute of Biochemistry and Biotechnology, Martin Luther University Halle-Wittenberg, Charles Tanford Protein Centre, Kurt-Mothes-Str. 3A, 06120 Halle (Saale), Germany
<b>Carlos</b>	<b>García-Mata</b>	IIB-CONICET-UNMDP
<b>Victoria</b>	<b>Gastaldi</b>	Instituto de Biología Molecular y Celular de plantas (IBMCP)
<b>Micaela Amalia</b>	<b>Godoy Herz</b>	Universidad de Buenos Aires (UBA), Facultad de Ciencias Exactas y Naturales, Departamento de Fisiología, Biología Molecular y Celular and CONICET-UBA, Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE), (C1428EHA), Buenos Aires, Argentina
<b>Diana Ester</b>	<b>Gras</b>	APOLO BIOTECH S.A. CONICET
<b>Gustavo Eduardo</b>	<b>Gudesblat</b>	Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Departamento de Fisiología, Biología Molecular y Celular. Instituto de Biociencias, Biotecnología y Biología Traslacional, iB3. Buenos Aires, Argentina. 2) CONICET, Buenos Aires, Argentina.
<b>Rodrigo A.</b>	<b>Gutiérrez</b>	Facultad de Ciencias Biológicas, Institute for Integrative Biology (iBio), Center for Genome Regulation (CRG), Institute of Ecology and Biodiversity (IEB), Pontificia Universidad Católica de Chile, Santiago, Chile
<b>Rossana</b>	<b>Henriques</b>	School of Biological, Earth and Environmental Sciences, University College Cork, North Mall Campus, Cork, Ireland / Environmental Research Institute, University College Cork, Lee Road, Cork, Ireland
<b>Fernando</b>	<b>Ibañez</b>	INIAB, Instituto de Investigaciones Agrobiotecnológicas (CONICET-UNRC). Ruta 36 Km 601, Río Cuarto, Argentina.
<b>María José</b>	<b>Iglesias</b>	Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE-UBA-CONICET) and Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, C1428EGA Buenos Aires, Argentina
<b>Peter</b>	<b>Kindgren</b>	Umea Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, 90187 Umea, Sweden
<b>Anne</b>	<b>Krapp</b>	Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin for Plant Sciences (IJPB), 78000, Versailles, France
<b>Ana María</b>	<b>Laxalt</b>	Instituto de Investigaciones Biológicas, CONICET, Universidad Nacional de Mar del Plata
<b>María Florencia</b>	<b>Legascue</b>	APOLO Biotech
<b>Martina</b>	<b>Legris</b>	University of Neuchâtel
<b>Verónica Viviana</b>	<b>Lia</b>	Instituto de Ecología, Genética y Evolución de Buenos Aires (IEGEBA, UBA-CONICET)

*Continued on next page*

Author's Name(s)	Author's Surname(s)	Author's Institution
<b>Chang</b>	<b>Liu</b>	Department of Epigenetics, Institute of Biology, University of Hohenheim, Garbenstrasse 30, 70599 Stuttgart, Germany
<b>Briardo</b>	<b>Llorente</b>	Australian Research Council Centre of Excellence in Synthetic Biology, School of Natural Sciences, Macquarie University, Sydney, Australia. Australian Genome Foundry, Sydney, Australia.
<b>Leandro Exequiel</b>	<b>Lucero</b>	Instituto de Agrobiotecnología del Litoral (CONICET-UNL), Universidad Nacional del Litoral, Colectora 6 Ruta Nacional 168 km 0, 3000, Santa Fe, Argentina; Facultad de Humanidades y Ciencias. Universidad Nacional del Litoral, Colectora Ruta Nacional 168 km 0, 10 3000, Santa Fe, Argentina
<b>Alexis</b>	<b>Maizel</b>	Center for Organismal Studies - Heidelberg University
<b>María Florencia</b>	<b>Mammarella</b>	APOLO Biotech
<b>Pablo Andrés</b>	<b>Manavella</b>	Instituto de Hortofruticultura Subtropical y Mediterránea (IHSM) 'La Mayora', Universidad de Málaga-Consejo Superior de Investigaciones Científicas (UMA-CSIC), 29010 Málaga, Spain
<b>Natanael</b>	<b>Mansilla</b>	APOLOBIOTECH, FBCB-UNL
<b>Guillermo A</b>	<b>Maroniche</b>	Facultad de Ciencias Agrarias, CONICET, Universidad Nacional de Mar del Plata, Argentina
<b>Sebastian</b>	<b>Marquardt</b>	Lund University
<b>Carmen</b>	<b>Martín-Pizarro</b>	Instituto de Hortofruticultura Subtropical y Mediterránea "La Mayora" (IHSM "La Mayora"), Universidad de Málaga-Consejo Superior de Investigaciones Científicas (UMA-CSIC), Málaga 29010, Spain.
<b>Julieta Lisa</b>	<b>Mateos</b>	Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE-UBA-CONICET) and Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, C1428EGA Buenos Aires, Argentina, and (2) RNA Biology and Molecular Physiology, Faculty of Biology, Bielefeld University, 33615 Bielefeld, Germany
<b>Saloni</b>	<b>Mathur</b>	National Institute of Plant Genome Research (NIPGR), Aruna Asaf Ali marg, New Delhi, India 110067
<b>Regina</b>	<b>Mencia</b>	Instituto de Agrobiotecnología del Litoral (CONICET-UNL), Cátedra de Biología Celular y Molecular, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral, 3000 Santa Fe, Argentina
<b>Catharina</b>	<b>Merchante</b>	Instituto de Hortofruticultura Subtropical y Mediterránea La Mayora IHSM-UMA-CSIC
<b>Javier Edgardo</b>	<b>Moreno</b>	Instituto de Agrobiotecnología del Litoral, Universidad Nacional del Litoral – CONICET, Facultad de Bioquímica y Ciencias Biológicas, Centro Científico Tecnológico CONICET Santa Fe, Colectora Ruta Nacional No. 168 km. 0, Paraje El Pozo, Santa Fe 3000, Argentina
<b>Jorge P.</b>	<b>Muschietti</b>	Instituto de Investigaciones en Ingeniería Genética y Biología Molecular, Dr. Héctor Torres (INGEBI-CONICET), Vuelta de Obligado 2490, C1428ADN, Buenos Aires, Argentina and 2 Departamento de Biodiversidad y Biología Experimental, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Int. Güiraldes 2160, Ciudad Universitaria, Pabellón II, C1428EGA Buenos Aires, Argentina.
<b>Andreas</b>	<b>Niebel</b>	CNRS
<b>Ariel</b>	<b>Orellana</b>	Universidad Andrés Bello
<b>Luciana Anabella</b>	<b>Pagnussat</b>	Facultad de Ciencias Agrarias, Universidad Nacional de Mar del Plata (FCA, UNMdP), INBIOTEC-CONICET, Mar del Plata, Buenos Aires, Argentina

*Continued on next page*

Author's Name(s)	Author's Surname(s)	Author's Institution
<b>Gabriela C.</b>	<b>Pagnussat</b>	Institutop de Investigaciones Biológicas IIB-Conicet Universidad nacional de Mar del Plata
<b>Javier F.</b>	<b>Palatnik</b>	Instituto de Biología Molecular y Celular de Rosario, CONICET, Universidad Nacional de Rosario, Rosario, Argentina.
<b>Rajni</b>	<b>Parmar</b>	Tennessee State University, Nashville, TN, USA
<b>Ezequiel</b>	<b>Petrillo</b>	Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE), CONICET-Universidad de Buenos Aires, C1428EHA, Buenos Aires, Argentina
<b>Delaux</b>	<b>Pierre-Marc</b>	Laboratoire de Recherche en Sciences Végétales (LRSV), Université de Toulouse, CNRS, UPS, INP, Toulouse, Castanet-Tolosan, 31320, France.
<b>David</b>	<b>Posé</b>	Instituto de Hortofruticultura Subtropical y Mediterránea "La Mayora" (IHSM "La Mayora"), Universidad de Málaga-Consejo Superior de Investigaciones Científicas (UMA-CSIC), Málaga 29010, Spain.
<b>Santiago</b>	<b>Prochetto</b>	Laboratory of Plant Physiology and Molecular Genetics, Université libre de Bruxelles, 1050 Brussels, Belgium. InBioS-PhytoSystems, Translational Plant Biology, University of Liège, 4000 Liège, Belgium
<b>Cécile</b>	<b>Raynaud</b>	Institute of Plant Sciences of Paris-Saclay
<b>Martiniano María</b>	<b>Ricardi</b>	Instituto de Fisiología, Biología Molecular y Neurociencias (IFIBYNE), CONICET-Universidad de Buenos Aires, Buenos Aires, Argentina
<b>Maximo</b>	<b>Rivarola</b>	ApoloBiotech - Conicet
<b>Ramiro Esteban</b>	<b>Rodriguez</b>	Instituto de Biología Molecular y Celular de Rosario (IBR, CONICET-UNR)
<b>Johan Stiben</b>	<b>Rodríguez Melo</b>	APOLO BIOTECH
<b>Facundo</b>	<b>Romani</b>	Department of Plant Sciences, University of Cambridge
<b>Sonali</b>	<b>Roy</b>	Tennessee State University
<b>Julio</b>	<b>Sáez-Vásquez</b>	CNRS, Laboratoire Génome et Développement des Plantes (LGDP), UMR 5096, 66860 Perpignan, France b) Univ. Perpignan Via Domitia, LGDP, UMR5096, 66860 Perpignan, France
<b>María Eugenia</b>	<b>Segretin</b>	Instituto de Investigaciones en Ingeniería Genética y Biología Molecular "Dr. Héctor N. Torres" (INGEBI-CONICET) // Departamento de Fisiología, Biología Molecular y Celular, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires FBM
<b>Sara</b>	<b>Selma</b>	VIB Center for Plant Systems Biology, 9052 Ghent, Belgium
<b>Santiago</b>	<b>Signorelli</b>	School of Agriculture, Universidad de la Republica; School of Molecular Sciences, The University of Western Australia
<b>Gabriela Cynthia</b>	<b>Soto</b>	Instituto de Genética (IGEAF, INTA), Buenos Aires, Argentina2)Instituto de Agrobiotecnología y Biología Molecular (IABIMO, CONICET-INTA), Buenos Aires, Argentina.
<b>Dorothee</b>	<b>Staiger</b>	RNA Biology and Molecular Physiology, Faculty of Biology, Bielefeld University
<b>Tomas Maria</b>	<b>Tessi</b>	Centre for organism al studies (COS) - Universität Heidelberg
<b>Ariel Hernán</b>	<b>Tomassi</b>	Instituto de Biología Molecular y Celular de Plantas (IBMCP), CSIC-Universitat Politècnica de València, Av. de los Naranjos s/n, 46022 Valencia, Spain
<b>María Laura</b>	<b>Tonelli</b>	UNRC-INIAB, CONICET
<b>Soledad</b>	<b>Traubenik</b>	Institute of Plant Sciences Paris-Saclay (IPS2), CNRS, INRAE, Université Paris-Saclay, 91190 Gif sur Yvette, France

*Continued on next page*

Author's Name(s)	Author's Surname(s)	Author's Institution
Hervé	Vaucheret	Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin for Plant Sciences (IJPB), 78000, Versailles, France.
Paul E.	Verslues	Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan
Dong	Wang	Institute of Advanced Agricultural Sciences, College of Life Science, Nanchang University
Elina	Welchen	CONICET and Universidad Nacional del Litoral
Marcelo Javier	Yanovsky	Instituto de Investigaciones Bioquímicas de Buenos Aires (CONICET - Fundación Instituto Leloir)ir
María Eugenia	Zanetti	Instituto de Biotecnología y Biología Molecular, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Centro Científico y Tecnológico-La Plata, Consejo Nacional de Investigaciones Científicas y Técnicas, 1900-La Plata, Argentina.
Diego	Zavallo	APOLO Biotech
Wenli	Zhang	State Key Laboratory of Crop Genetics & Germplasm Enhancement and Utilization, Nanjing Agricultural University, No.1 Weigang, Nanjing, Jiangsu 210095

## B MoBiPLANT Examples

The following are 3 example questions drawn from **Expert MoBiPlant**.

What proteins have been identified as molecular partners of the Arabidopsis lncRNA ASCO? In what molecular mechanism are they all involved in?

**Options:**

- A)The proteins identified so far as interactors of the lncRNA ASCO are NSRa, PRP8 and SmD1b, which are involved in splicing.
  - B)The proteins identified so far as interactors of the lncRNA ASCO are LHP1 and VIM1, which are involved in epigenetics.
  - C)The proteins identified so far as interactors of the lncRNA ASCO are GRP7, PRP64 and SmD1a, which are involved in splicing.
- Correct answer:** A)

How many subunits make up the RNA polymerase I complex in Arabidopsis thaliana?

**Options:**

- A)In Arabidopsis, RNA Pol I consists of 12 protein subunits common to RNA Pol II and Pol III the two others being RNA Pol I-specific subunits.
  - B)In Arabidopsis, RNA Pol I consists of 14 protein subunits: 12 are RNA Pol I-specific protein subunits and two others are common to RNA Pol II and Pol III.
  - C)In Arabidopsis, RNA Pol I consists of 12 subunits common to those of all nuclear RNA polymerases (RNA Pol I–Pol V). 5 are common to RNA Pol II and Pol III and the others are RNA Pol I-specific subunits.
- Correct answer:** C)

The PCR2 complex has conserved functions in plants and animals. Which histone modification is the result of the action of PCR2 in Arabidopsis thaliana?

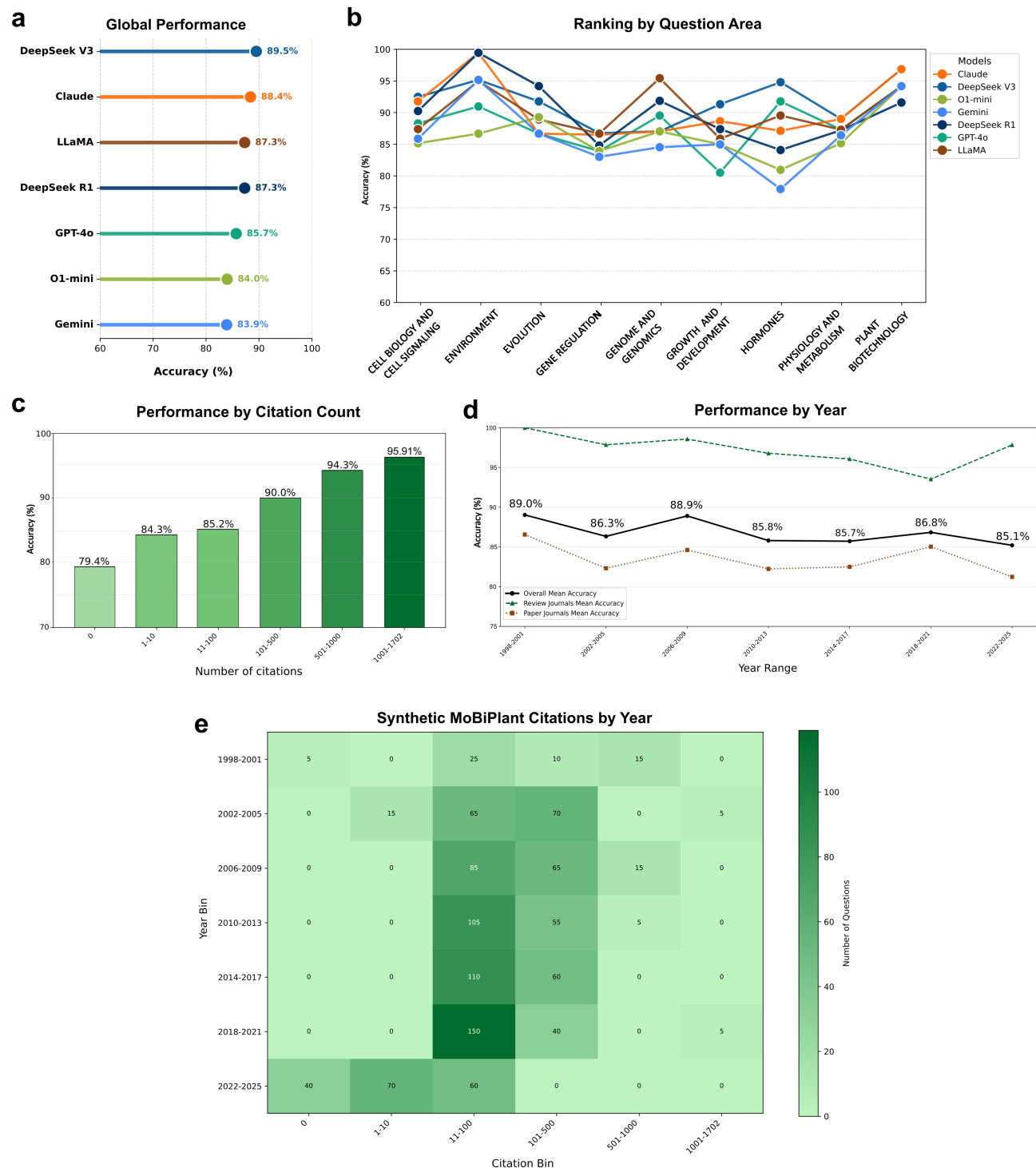
**Options:**

- A)In Arabidopsis PRC2 mediates the deposition of histone H3 lysine 36 trimethylation (H3K36me3).
  - B)In Arabidopsis PRC2 mediates the deposition of histone H3 lysine 27 trimethylation (H3K27me3).
  - C)In Arabidopsis PRC2 mediates the deposition of histone H3 lysine 4 trimethylation (H3K4me3).
- Correct answer:** B)

## C Synthetic MoBiPLANT Results

We report the evaluation results of Synthetic MoBiplant set in Supplementary Figure 1.





**Supplementary Figure 1. Synthetic MoBiPlant benchmarking results.** **a)** The overall accuracy over the entire synthetic set. **b)** Fluctuation of model ranking across question areas. Each line and collection of dots represent the model accuracy in the area categories. **c)** Correlation between source's amount of citations and mean model performance, with accuracy scores scaled to the 70–100 range. **d)** Accuracy plotted against publication year. The color intensity in the plot reflects the accuracy score for each data point. The source's release year and citation count used in this analysis were extracted from the original work that motivated the research question.

### D Answer Distribution

Supplementary Table 2 shows the answer distribution of Gemini 1.5 Pro and DeepSeek-R1 in the option bias experiment.

**Supplementary Table 2. Answer distributions for option bias experiment in MOBiPLANT.** For the original shuffles (no option bias), mean accuracy is reported across all three shuffles.

Dataset & Model		A	B	C	Format Error	Answer Accuracy
Option bias in A	GroundTruth	565	0	0	0	-
	Gemini 1.5 Pro	505	31	29	0	89.1
	DeepSeek R1	509	26	30	0	90.1
Option bias in B	GroundTruth	0	565	0	0	-
	Gemini 1.5 Pro	120	397	32	16	71.9
	DeepSeek R1	58	470	37	0	82.9
Option bias in C	GroundTruth	0	0	565	0	-
	Gemini 1.5 Pro	140	55	356	14	64.3
	DeepSeek R1	61	44	460	0	81.0
Original Gemini 1.5 Pro	First shuffle	262	163	140	0	76.8
	Second shuffle	256	164	145	0	
	Third shuffle	264	154	147	0	
Original DeepSeek R1	First shuffle	225	168	168	4	86.1
	Second shuffle	208	168	174	15	
	Third shuffle	226	155	174	10	

### E Failure Modes

In Supplementary Table 3 we report additional examples of failure modes encountered on the models evaluated on MOBiPLANT. We extend the definition of the found set of failure modes on the following list:

- **Cross-domain bias:** Apply animal or human biology mechanisms to plants incorrectly.
- **Species confusion:** Mix up information across different plant species.
- **Acronym misinterpretation:** Expand plant-specific abbreviations into wrong or non-existent gene names.
- **Terminology gaps:** Misunderstand specialized plant physiology terms (e.g. thermomorphogenesis).
- **Outdated knowledge:** Lack recent literature on emerging plant research areas.
- **Hallucinated pathways:** Combine unrelated defense or signaling pathways or involve incorrect genes in a given molecular mechanism.
- **Environmental misunderstanding:** Struggle with field conditions, temperature stress, and abiotic factors.

**Supplementary Table 3. Additional failure-mode examples from the open-ended assessment of Expert MoBiPlant.**

Error	Model	Prompt	Model response	Error explanation
Cross-domain bias & Species confusion	DeepSeek V3	What types of DNA motifs bound by transcription factors are enriched at the boundaries of TADs in plant species?	At plant TAD boundaries, motifs bound by architectural proteins are claimed to be enriched, with CTCF-like sequences and conserved elements recognized by plant factors; <i>Arabidopsis</i> studies are cited as evidence for critical roles in 3D genome organization.	CTCF homologs have not been identified in plants; CTCF is not required for plant 3D boundaries.

Continued on next page

Error	Model	Prompt	Model response	Error explanation
<b>Acronym misinterpretation &amp; Species confusion &amp; Hallucinated pathways</b>	DeepSeek R1	Which processes are regulated by the transcription factor MIB2 during thermomorphogenesis in tomato?	MIB2 (MYB-INTERACTING bHLH 2) in tomato is said to regulate thermomorphogenesis by modulating auxin biosynthesis/signaling under high temperature, directly activating <i>YUC3</i> and auxin-responsive genes (e.g., <i>IAA19</i> ); it is also claimed to interact with heat-shock factors to coordinate growth and stress tolerance.	MULTIPLE INFLORESCENCE BRANCH 2 (MIB2) modulates inflorescence branching at high temperature in tomato. Uses Arabidopsis pathways to extrapolate across species.
<b>Terminology gaps &amp; Cross-domain bias &amp; Environmental misunderstanding</b>	o1-mini	Which lncRNA is involved in the thermomorphogenic response in <i>Arabidopsis thaliana</i> and how?	The long noncoding RNA ELF3-AS is involved in the thermomorphogenic response in <i>Arabidopsis thaliana</i> by regulating the expression of the ELF3 gene, a crucial component in temperature signaling pathways. ELF3-AS modulates ELF3 activity, thereby influencing plant growth and development in response to elevated temperatures.	Claims ELF3-AS regulates <i>ELF3</i> during thermomorphogenesis. ELF3-AS1 is an oncogenic human lncRNA (bladder cancer). In <i>Arabidopsis</i> , high temperature is sensed via thermo-inducible phase separation of the protein EARLY FLOWERING 3 (ELF3); ELF3 is not a lncRNA.
<b>Cross-domain bias &amp; Hallucinated pathways</b>	GPT-4o	In <i>Medicago truncatula</i> , GA suppression of infection involves the degradation of which proteins acting in the GA signaling pathway?	In <i>Medicago truncatula</i> , gibberellin (GA) suppression of infection involves the degradation of DELLA proteins, which act as repressors in the GA signaling pathway. DELLA proteins are negative regulators that inhibit growth and development processes, and their degradation by the GA signaling mechanism allows for the promotion of growth and alleviation of the suppression caused by stress responses, including pathogen infection. The action of GA in degrading these proteins helps facilitate effective plant-microbe interactions, particularly in symbiotic associations.	Mixing results in different species, and signaling pathways; DELLA proteins can have positive effects in specific development processes. GA-mediated DELLA degradation results in a reduction of infection
<b>Hallucinated pathways</b>	DeepSeek R1	What is the link between TCP15, auxins and cytokinins during gynoecium development in <i>Arabidopsis thaliana</i> ?	TCP15, a transcription factor in <i>Arabidopsis thaliana</i> , integrates auxin and cytokinin (CK) signaling to regulate gynoecium development. TCP15 promotes auxin biosynthesis by directly activating <i>YUC4</i> (a key auxin biosynthesis gene), establishing auxin maxima critical for medial tissue formation. Concurrently, TCP15 represses cytokinin signaling by downregulating type-B ARR genes (e.g., <i>ARR1</i> , <i>ARR12</i> ), which mediate CK responses. This dual regulation balances auxin-CK antagonism, ensuring proper apical-basal patterning and style/stigma differentiation. Disrupting TCP15 leads to cytokinin overactivation, auxin deficiency, and abnormal gynoecium morphology.	TCP15 links cytokinin and gibberellin responses; CK induces <i>TCP15</i> , which inhibits auxin-biosynthesis genes, affecting inner gynoecium tissues.
<b>Hallucinated pathways</b>	DeepSeek V3	Which photoreceptors phosphorylate the protein PHYTOCHROME KINASE SUBSTRATE 4 in response to blue light in <i>Arabidopsis</i> ?	In <i>Arabidopsis</i> , the phosphorylation of PHYTOCHROME KINASE SUBSTRATE 4 (PKS4) in response to blue light is primarily mediated by the cryptochrome photoreceptors, CRY1 and CRY2. These blue light receptors interact with PKS4 and regulate its phosphorylation, which is crucial for photomorphogenic responses.	Phototropins.

Continued on next page

Error	Model	Prompt	Model response	Error explanation
<b>Acronym misinterpretation</b>	LLaMA 3.1 405B	How does the lncRNA APOLO regulate the <i>RHD6</i> locus in <i>Arabidopsis thaliana</i> in response to low temperatures?	The lncRNA APOLO (Alternative Polyadenylation-Linked Odd Locus Organizer) in <i>Arabidopsis thaliana</i> regulates the <i>RHD6</i> (ROOT HAIR DEFECTIVE 6) gene locus in response to low temperatures through a mechanism involving chromatin looping. APOLO acts as a scaffold for the recruitment of the Polycomb Repressive Complex 1 (PRC1) to the <i>RHD6</i> locus, resulting in the repression of <i>RHD6</i> gene expression under cold conditions. This repression is mediated by the physical interaction between APOLO and the PRC1 subunit LHP1 (LIKE HETEROCHROMATIN PROTEIN 1), and the reduction of <i>RHD6</i> expression contributes to the adaptation of plant growth to low temperatures.	APOLO = AUXIN REGULATED PROMOTER LOOP.
<b>Cross-domain bias</b>	GPT-4o	How does the lncRNA APOLO regulate the locus of the <i>RHD6</i> gene in <i>Arabidopsis thaliana</i> in response to low temperatures?	In <i>Arabidopsis thaliana</i> , the long non-coding RNA (lncRNA) APOLO plays a crucial role in regulating the <i>RHD6</i> gene locus in response to low temperatures through a process involving chromatin modulation. Under cold conditions, APOLO interacts with the chromatin at the <i>RHD6</i> locus and recruits the Polycomb Repressive Complex 1 (PRC1), which leads to the deposition of histone modifications associated with repressive chromatin states, such as H2Aub (ubiquitinated histone H2A). This recruitment ultimately influences the expression of <i>RHD6</i> , a gene involved in root hair development, by altering the chromatin structure to a more closed conformation, thereby affecting the plant's adaptive responses to cold stress.	In <i>Arabidopsis</i> , plant PRC1 component LHP1 is involved in the recognition and spreading of H3K27me3. The mistake arises from the knowledge about metazoan PRC1.
<b>Outdated knowledge</b>	DeepSeek R1	What proteins have been identified as molecular partners of the <i>Arabidopsis</i> lncRNA ASCO? In what molecular mechanism are they all involved in?	The <i>Arabidopsis</i> lncRNA ASCO interacts with Nuclear Speckle RNA-binding proteins (AtNSRs), including NSRa and NSRb, which are involved in alternative splicing. ASCO modulates these proteins by sequestering them in nuclear speckles, thereby altering their ability to bind and regulate the splicing of target pre-mRNAs. This interaction impacts auxin-mediated alternative splicing events, influencing lateral root development and stress responses. The mechanism involves ASCO acting as a competing RNA to disrupt NSR activity, which is critical for integrating splicing regulation with auxin signaling pathways in plant development.	In 2020, it was shown that in addition to NSR proteins, the lncRNA ASCO interacts with the splicing factors PRP8 and SmD1b.
<b>Hallucinated pathways</b>	LLaMA 3.1 405B	What proteins have been identified as molecular partners of the <i>Arabidopsis</i> lncRNA ASCO? In what molecular mechanism are they all involved in?	The <i>Arabidopsis</i> lncRNA ASCO has been identified to interact with the protein complex composed of the nuclear speckle RNA-binding protein (NSR), alternative splicing regulator SR45, and serine/arginine-rich splicing factor 45a (SR45a) and the splice factor U2AF65b. All of these proteins are involved in the regulation of alternative splicing. ASCO mediates the interaction between NSR and SR45, promoting the alternative splicing of NSR-targeted pre-mRNAs.	The lncRNA ASCO interacts with splicing factors but not with SR42 nor U2AF65b. Furthermore, ASCO has not been shown to mediate the interaction between NSR and SR45.

Continued on next page

Error	Model	Prompt	Model response	Error explanation
Species confusion & Hallucinated pathways	DeepSeek R1	What is the impact of 1 $\mu$ M bioactive gibberellin (GA) application on epidermal infection by symbiotic rhizobia in <i>Medicago truncatula</i> ?	The application of 1 $\mu$ M bioactive gibberellin (GA) suppresses epidermal infection by symbiotic rhizobia in <i>*Medicago truncatula*</i> by inhibiting the formation of infection threads, a critical step in nodulation. This occurs because GA antagonizes key symbiotic signaling pathways, such as those involving DELLA proteins, which are essential for rhizobial infection. Experimental studies show that exogenous GA reduces nodule numbers and infection events, while GA-deficient mutants exhibit enhanced nodulation. The inhibitory effect of GA highlights its role as a negative regulator of early symbiotic interactions in legumes.	"suppresses -> significantly reduces. DELLA signalling is not a symbiotic pathway. The question specified "GA application", infection"" and ""in <i>Medicago truncatula</i> """, specific GA concentrations can reduce nodule number in <i>Medicago</i> but GA-deficient mutants have not been shown to exhibit enhanced nodulation
Acronym misinterpretation	GPT-4o	Which proteins facilitate UVR8 redimerization in <i>Arabidopsis thaliana</i> ?	In <i>Arabidopsis thaliana</i> , the proteins RUP1 (ROOT UVB SENSITIVE 1) and RUP2 (ROOT UVB SENSITIVE 2) facilitate the redimerization of the UVR8 photoreceptor. UVR8 is a UV-B light receptor that monomerizes upon UV-B exposure, leading to the activation of a signaling pathway that helps the plant respond to UV-B stress. RUP1 and RUP2 play crucial roles in the UVR8 signaling pathway by promoting the redimerization of UVR8, thus attenuating the UV-B signal and contributing to the reestablishment of the photoreceptor's inactive state.	RUP = REPRESSOR OF UV-B PHOTOMORPHOGENESIS.
Species confusion & Hallucinated pathways	LLaMA 3.1 405B	What is the impact of 1 $\mu$ M GA on epidermal infection by symbiotic rhizobia in <i>Medicago truncatula</i> ?	Application of 1 $\mu$ M bioactive gibberellin (GA) inhibits epidermal infection by symbiotic rhizobia in <i>Medicago truncatula</i> . This is because gibberellins play a regulatory role in the nodulation process of legumes, and elevated levels of GA can suppress rhizobial infections and nodule formation.	"inhibits-> significantly reduces. The question specified ""infection"" and ""in <i>Medicago truncatula</i> """, GAs can have different effects (positive or negative) on the other stages of nodulation depending on concentrations and species. The causal link is reversed, the infection process being part of the nodule formation process.
Species confusion & Hallucinated pathways	DeepSeek R1	In <i>Lotus japonicus</i> , what is the impact of 1 $\mu$ M GA on root hair deformation in response to Nod Factors (NF)?	In <i>*Lotus japonicus*</i> , application of 1 $\mu$ M bioactive gibberellin (GA) enhances root hair deformation in response to Nod factors (NF), suggesting a synergistic role of GA in promoting early symbiotic signaling. This was demonstrated in a study where GA treatment amplified NF-induced root hair curling and calcium spiking, key steps in rhizobial infection. The effect is likely mediated through GA's interaction with the DELLA protein SLR1, which modulates transcriptional responses to NF. However, higher GA concentrations (e.g., 10 $\mu$ M) suppress nodulation, indicating dose-dependent regulation.	Completely abolishes. GA causes complete loss of NF-induced root hair deformation in <i>Lotus</i> . SLR1 is a rice DELLA. The question is not about nodulation.

Continued on next page



Error	Model	Prompt	Model response	Error explanation
<b>Hallucinated pathways</b>	Gemini 1.5 Pro	What is the impact of 1 $\mu$ M GA on epidermal infection by symbiotic rhizobia in <i>Medicago truncatula</i> ?	Application of 1 $\mu$ M bioactive gibberellin (GA) inhibits epidermal infections by symbiotic rhizobia in * <i>Medicago truncatula</i> *. This concentration significantly reduces the number of infection threads formed after rhizobial inoculation, indicating that GA negatively regulates this early stage of the symbiotic interaction. This effect is observed even in the presence of Nod factors, suggesting that GA acts downstream of Nod factor signaling.	“ <i>Inhibits</i> ” → significantly reduces. GAs likely act downstream of Nod factor perception but directly control Nod factor signaling.
<b>Hallucinated pathways</b>	Claude 3.5 Sonnet	What is the impact of 1 $\mu$ M GA on epidermal infection by symbiotic rhizobia in <i>Medicago truncatula</i> ?	Application of 1 $\mu$ M bioactive GA (GA3) inhibits rhizobial infection in <i>Medicago truncatula</i> by reducing both the number of infection threads and nodule formation. GA treatment interferes with early symbiotic signaling and suppresses the expression of key nodulation genes like NSP1, NSP2, and ERN1 that are essential for infection thread development. This negative effect of GA on rhizobial infection is mediated through DELLA proteins, which are degraded upon GA perception.	“ <i>Inhibits</i> ” → significantly reduces. Causal link reversed (infection is part of nodulation). The question is about <i>infection</i> ; GA can have positive effects on nodulation elsewhere. “ <i>Suppresses</i> ” → significantly reduces.
<b>Hallucinated pathways</b>	GPT-4o	What is the impact of 1 $\mu$ M GA on epidermal infection by symbiotic rhizobia in <i>Medicago truncatula</i> ?	The application of 1 $\mu$ M bioactive gibberellin (GA) has been shown to negatively impact the epidermal infection of <i>Medicago truncatula</i> by symbiotic rhizobia. This effect is likely due to the influence of gibberellins on the early stages of root nodule formation, which are crucial for successful rhizobial infection and symbiosis establishment. Gibberellins can affect the expression of specific genes and the hormonal balance within the plant, thereby modulating the plant’s susceptibility to rhizobial infection. Such hormonal modulation could interfere with signaling pathways essential for the recognition and infection processes during symbiosis.	Causal link reversed (infection $\subset$ nodulation). Last claim not wrong but heavily extrapolated.
<b>Cross-domain bias &amp; Hallucinated pathways</b>	Gemini 1.5 Pro	In <i>Medicago truncatula</i> , GA suppression of <i>infection</i> involves degradation of which proteins?	In * <i>Medicago truncatula</i> *, gibberellic acid (GA) suppression of infection by the root-knot nematode * <i>Meloidogyne hapla</i> * involves the degradation of DELLA proteins, which are negative regulators of the GA signaling pathway. This degradation is mediated by the F-box protein SKP1, leading to the enhanced expression of GA-responsive genes that contribute to infection susceptibility.	Misinterprets question (rhizobia vs. nematode). Last claim has never been shown in <i>Medicago</i> .
<b>Hallucinated pathways</b>	Claude 3.5 Sonnet	In <i>Medicago truncatula</i> , GA suppression of infection involves degradation of which proteins?	In <i>Medicago truncatula</i> , GA signaling involves the degradation of DELLA proteins, specifically MtDELLA1 and MtDELLA2. When GA levels increase, these DELLA proteins are targeted for degradation via the 26S proteasome pathway, which suppresses rhizobial infection. The DELLAs normally act as positive regulators of nodulation by promoting infection thread formation, so their GA-induced degradation results in reduced nodulation. This mechanism helps explain how elevated GA levels can inhibit symbiotic interactions between <i>M. truncatula</i> and rhizobia.	Also MtDELLA3. “ <i>Suppresses</i> ” → reduces. “ <i>Nodulation</i> ” → infection (per question).
<b>Hallucinated pathways</b>	o1-mini	In <i>Medicago truncatula</i> , GA suppression of infection involves degradation of which proteins?	In * <i>Medicago truncatula</i> *, gibberellin (GA) suppression of infection involves the degradation of DELLA proteins, which are key repressors in the GA signaling pathway. When GA levels increase, DELLA proteins are ubiquitinated and subsequently degraded via the 26S proteasome, facilitating the activation of growth and immune responses that suppress pathogen infection.	GA-mediated DELLA degradation reduces infection.

Continued on next page

Error	Model	Prompt	Model response	Error explanation
<b>Hallucinated pathways</b>	DeepSeek V3	In <i>Medicago truncatula</i> , GA suppression of infection involves degradation of which proteins?	In * <i>Medicago truncatula</i> *, GA (gibberellin) suppression of infection involves the degradation of DELLA proteins, which act as negative regulators in the GA signaling pathway. DELLA proteins inhibit plant growth and development, and their degradation is promoted by GA through the ubiquitin-proteasome system, allowing for the activation of GA-responsive genes. This process is crucial for the plant's response to symbiotic interactions and pathogen resistance.	"Symbiotic infection." (Vague/incorrect scope).
<b>Outdated knowledge &amp; Hallucinated pathways</b>	LLaMA 3.1 405B	In <i>Medicago truncatula</i> , the promoter of which TF gene controlling infection thread progression is bound by GFP-della1-Δ18?	In <i>Medicago truncatula</i> , the promoter of the NIN gene (Nodule Inception) is directly bound by the GFP-della1-Delta18 protein. This gene encodes a transcription factor essential for the regulation of the infection thread progression during symbiotic nitrogen fixation in legume plants.	<i>NIN</i> → <i>ERN1</i> .
<b>Hallucinated pathways</b>	GPT-4o	In <i>Medicago truncatula</i> , the promoter of which TF gene controlling infection thread progression is bound by GFP-della1-Δ18?	In <i>Medicago truncatula</i> , the promoter of the gene ENOD11, which encodes an important transcription factor involved in the regulation of infection thread progression during nodulation, is directly bound by the GFP-della1-Delta18 protein. This interaction highlights the role of DELLA proteins in the symbiotic relationship between legumes and nitrogen-fixing bacteria, as they are crucial for the regulation of gene expression in response to rhizobial infection.	<i>ERN1</i> . <i>ENOD11</i> is not a TF.
<b>Hallucinated pathways</b>	o1-mini	In <i>Medicago truncatula</i> , the promoter of which TF gene controlling infection thread progression is bound by GFP-della1-Δ18?	In * <i>Medicago truncatula</i> *, the promoter of the *NODULE INCEPTION (NIN)* gene, which encodes a crucial transcription factor regulating infection thread progression, is directly bound by the GFP-della1-Δ18 protein. This interaction plays a significant role in the symbiotic nodulation process.	<i>NSP1</i> → <i>ERN1</i> .
<b>Hallucinated pathways</b>	DeepSeek V3	In <i>Medicago truncatula</i> , the promoter of which TF gene controlling infection thread progression is bound by GFP-della1-Δ18?	In * <i>Medicago truncatula</i> *, the promoter of the *NODULE INCEPTION (NIN)* gene, which encodes a crucial transcription factor regulating infection thread progression, is directly bound by the GFP-della1-Δ18 protein. This interaction plays a significant role in the symbiotic nodulation process.	<i>NIN</i> → <i>ERN1</i> .
<b>Outdated knowledge &amp; Hallucinated pathways</b>	LLaMA 3.1 405B	In <i>Lotus japonicus</i> , what is the impact of 1 μM GA on root hair deformation in response to NF?	In <i>Lotus japonicus</i> , the application of 1 μM bioactive gibberellin (GA) has been shown to enhance root hair deformation in response to Nod Factors (NF). This effect is likely due to the role of gibberellins in regulating plant cell growth and differentiation. The interaction between GA and NF signaling pathways can lead to increased root hair deformation, which is an important step in the initiation of symbiotic relationships between legume plants and rhizobia.	"Enhance": Completely abolish. Complete loss of root hair deformation in <i>Lotus</i> .

Continued on next page

Error	Model	Prompt	Model response	Error explanation
<b>Terminology gaps</b>	Gemini 1.5 Pro	In <i>Lotus japonicus</i> , what is the impact of 1 $\mu$ M GA on NF-induced root hair deformation?	Application of 1 $\mu$ M bioactive gibberellin (GA) inhibits root hair deformation induced by Nod factors (NF) in * <i>Lotus japonicus</i> *. While NF alone triggers root hair curling and swelling, the simultaneous application of GA suppresses these morphological changes, indicating an antagonistic interaction between GA and NF signaling pathways in regulating root hair development during nodulation.	Swelling not shown. <i>development</i> → <i>deformation</i> .
<b>Hallucinated pathways</b>	GPT-4o	In <i>Lotus japonicus</i> , what is the impact of 1 $\mu$ M GA on NF-induced root hair deformation?	In <i>Lotus japonicus</i> , the application of 1 $\mu$ M bioactive gibberellin (GA) has been shown to enhance root hair deformation in response to Nod Factors (NF). Gibberellins are plant hormones that play a key role in various developmental processes, including the promotion of root hair elongation and deformation. The presence of GA enhances the plant's ability to respond to NF, which are signaling molecules produced by rhizobia to initiate symbiotic interactions leading to nitrogen fixation. Thus, GA treatment can facilitate the symbiotic relationship by promoting the morphological changes necessary for effective nodule formation and function.	"Enhance": Completely abolish. GA causes complete loss of NF-induced deformation in <i>Lotus</i> . The question did not deal with the effect of GA on nodulation.
<b>Outdated knowledge &amp; Hallucinated pathways</b>	LLaMA 3.1 405B	What is the infection-thread formation phenotype of DELLA-deficient pea <i>la cry-s</i> double mutants vs. wild type?	In DELLA-deficient pea <i>la cry-s</i> double mutants, the infection thread formation is increased compared to wild-type. The absence of DELLA proteins, which are negative regulators of gibberellin signaling, results in enhanced root growth and increased nodulation. This is due to the increased expression of genes involved in nodule formation and the suppression of defense responses against rhizobia, leading to more efficient infection thread formation and nodule development.	"Increased": Significantly reduced. "results enhanced root growth ...": Not shown. Last claim is wrong
<b>Hallucinated pathways</b>	Gemini 1.5 Pro	What is the infection-thread formation phenotype of DELLA-deficient pea <i>la cry-s</i> double mutants vs. wild type?	Pea <i>la cry-s</i> double mutants, which are DELLA-deficient, exhibit more rapid infection thread formation compared to wild-type plants. The absence of DELLA proteins, which are negative regulators of gibberellin signaling, leads to enhanced susceptibility to rhizobial infection, resulting in faster infection thread development. This suggests that gibberellins play a role in negatively regulating nodulation in pea.	"rapid": Significant decrease. "enhanced susceptibility": Decreased. No data on formation rate. Gibberellins roles vary by nodulation stage.
<b>Hallucinated pathways</b>	Claude 3.5 Sonnet	What is the infection-thread formation phenotype of DELLA-deficient pea <i>la cry-s</i> double mutants vs. wild type?	DELLA-deficient <i>la cry-s</i> double mutant pea plants show excessive infection thread formation compared to wild-type plants during rhizobial infection. These mutants form approximately twice as many infection threads as wild-type plants, indicating that DELLA proteins normally act to restrict or limit infection thread formation during the early stages of rhizobial symbiosis. This phenotype demonstrates that DELLA proteins play an important role in negatively regulating the early infection events of rhizobial symbiosis.	"Enhanced": Significantly reduced. Imagined quantitative data. "negatively regulating": Positively.

Continued on next page

Error	Model	Prompt	Model response	Error explanation
<b>Hallucinated pathways</b>	GPT-4o	What is the infection-thread formation phenotype of DELLA-deficient pea <i>la cry-s</i> double mutants vs. wild type?	DELLA-deficient pea <i>la cry-s</i> double mutants exhibit an enhanced infection thread formation phenotype compared to wild-type plants. DELLA proteins are known to act as growth repressors, and their deficiency in these mutants leads to increased gibberellin signaling, which promotes root nodule formation and consequently enhances symbiotic interactions with rhizobia. This results in an increased number of infection threads, which are essential for the successful colonization of root nodules by nitrogen-fixing bacteria.	"Enhanced": Significantly reduced. "Repress"/causal link reversed (infection is part of nodulation).
<b>Hallucinated pathways</b>	o1-mini	What is the infection-thread formation phenotype of DELLA-deficient pea <i>la cry-s</i> double mutants vs. wild type?	DELLA-deficient pea <i>la cry-s</i> double mutants exhibit an enhanced infection thread formation phenotype compared to wild-type plants, indicating that DELLA proteins normally act to restrict or regulate the formation of infection threads during symbiotic nodulation. This suggests that the absence of DELLA proteins allows for more robust or frequent infection thread development, potentially facilitating improved symbiotic interactions with rhizobia.	"Enhanced": Significantly reduced. "Restrict": Induce. "More robust": Less robust. "Potentially facilitating improved symbiotic interactions with rhizobia": Impairing.
<b>Species confusion &amp; Hallucinated pathways</b>	DeepSeek R1	What is the infection-thread formation phenotype of DELLA-deficient pea <i>la cry-s</i> double mutants vs. wild type?	In DELLA-deficient <i>*la cry-s*</i> double mutants of pea, infection thread formation is significantly impaired compared to wild-type plants. DELLA proteins, which are positive regulators of gibberellin signaling, are essential for rhizobial infection and nodulation. Their deficiency disrupts the coordination of signaling pathways required for infection thread initiation and progression, leading to fewer or aborted infection threads. Wild-type plants exhibit robust infection thread development due to DELLA-mediated activation of nodulation-related genes, such as those involving NSP2 and cell wall modifications. This phenotype highlights DELLA's critical role in facilitating early symbiotic interactions.	Not shown in pea; partially true in <i>Medicago</i> .

**Supplementary Table 4. Examples of common failure modes sighted on the open-ended assessment of Expert MoBiPlant for the references indicated by LLMs to support their statements.**

Prompt	Right Reference	LLM	Source Provided by the LLM	Title	Authorship	DOI
What was the first example of a long noncoding RNA capturing a miRNA by target mimicry in any living organism and what is it involved in? Indicate the corresponding organism and describe the associated molecular mechanism.	Franco-Zorrilla, J., Valli, A., Todesco, M. <i>et al.</i> Target mimicry provides a new mechanism for regulation of microRNA activity. <i>Nat Genet</i> 39, 1033–1037 (2007) ( <a href="https://doi.org/10.1038/ng2079">https://doi.org/10.1038/ng2079</a> ).	O1-mini	Target mimicry provides a new mechanism for regulation of microRNA activity in plants ( <a href="https://www.pnas.org/content/105/38/14593">https://www.pnas.org/content/105/38/14593</a> ).	The title is partially right	Authorship is not stated	The link does not exist

*Continued on next page*

Prompt	Right Reference	LLM	Source Provided by the LLM	Title	Authorship	DOI
What was the first example of a long noncoding RNA capturing a miRNA by target mimicry in any living organism and what is it involved in? Indicate the corresponding organism and describe the associated molecular mechanism.	Franco-Zorrilla, J., Valli, A., Todesco, M. <i>et al.</i> Target mimicry provides a new mechanism for regulation of microRNA activity. <i>Nat Genet</i> 39, 1033–1037 (2007) ( <a href="https://doi.org/10.1038/ng2079">https://doi.org/10.1038/ng2079</a> ).	GPT-4o	A noncoding RNA acts as a microRNA decoy to regulate phosphate homeostasis ( <a href="https://www.science.org/doi/10.1126/science.1154427">https://www.science.org/doi/10.1126/science.1154427</a> ).	The title is wrong (the paper does not exist)	Authorship is not stated	The link does not exist
What is the infection thread formation phenotype of DELLA-deficient pea <i>la cry-s</i> double mutants compared with wild-type?	Erin L. McAdam, James B. Reid, Eloise Foo, Gibberellins promote nodule organogenesis but inhibit the infection stages of nodulation, <i>J. Exp. Bot.</i> 69(8):2117–2130 (2018) ( <a href="https://doi.org/10.1093/jxb/ery046">https://doi.org/10.1093/jxb/ery046</a> ).	Claude	DELLA proteins regulate arbuscule formation in arbuscular mycorrhizal symbiosis (Floss <i>et al.</i> , PNAS 2013) ( <a href="https://www.pnas.org/content/110/51/E5025">https://www.pnas.org/content/110/51/E5025</a> ).	The title cited exists but is not the correct one	Authorship corresponds to the cited article	The link corresponds to the cited article but is not the most suitable one
What is the infection thread formation phenotype of DELLA-deficient pea <i>la cry-s</i> double mutants compared with wild-type?	Erin L. McAdam, James B. Reid, Eloise Foo, Gibberellins promote nodule organogenesis but inhibit the infection stages of nodulation, <i>J. Exp. Bot.</i> 69(8):2117–2130 (2018) ( <a href="https://doi.org/10.1093/jxb/ery046">https://doi.org/10.1093/jxb/ery046</a> ).	DeepSeek R1	Floss <i>et al.</i> , “DELLAs regulate rhizobial infection in legumes,” <i>Science</i> (2013) ( <a href="https://www.science.org/doi/10.1126/science.1244467">https://www.science.org/doi/10.1126/science.1244467</a> ).	The title is wrong (the paper does not exist)	Authorship is from another article	The link does not exist
Which proteins facilitate UVR8 redimerization in <i>Arabidopsis thaliana</i> ?	M. Heijde & R. Ulm. Reversion of the <i>Arabidopsis</i> UV-B photoreceptor UVR8 to the homodimeric ground state. <i>PNAS</i> 110(3):1113–1118 (2013) ( <a href="https://doi.org/10.1073/pnas.1214237110">https://doi.org/10.1073/pnas.1214237110</a> ).	Gemini	Heilmann, B., & Jenkins, G. I. (2013). UV-B photoreceptor-mediated signalling in plants. <i>Biochem. Soc. Trans.</i> 41(6):1745–1749 ( <a href="https://pubmed.ncbi.nlm.nih.gov/24256252/">https://pubmed.ncbi.nlm.nih.gov/24256252/</a> ).	The title corresponds to a related review article but is not the correct one	Authorship belongs to a related article published in <i>Plant Physiology</i> (wrong journal) in 2013, but it is not correct for the cited article	The link directs to a completely unrelated article on PubMed
Which proteins facilitate UVR8 redimerization in <i>Arabidopsis thaliana</i> ?	M. Heijde & R. Ulm. Reversion of the <i>Arabidopsis</i> UV-B photoreceptor UVR8 to the homodimeric ground state. <i>PNAS</i> 110(3):1113–1118 (2013) ( <a href="https://doi.org/10.1073/pnas.1214237110">https://doi.org/10.1073/pnas.1214237110</a> ).	DeepSeek V3	RUP1 and RUP2 are essential for UVR8 redimerization in <i>Arabidopsis</i> ( <a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4415188/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4415188/</a> ).	The title is wrong (the paper does not exist)	Authorship is not stated	The link leads to another article unrelated to the topic



## F MCQ Collection Guidelines

### Guidelines Resource

**Note:** please do not use LLMs of any kind in the making of the questions. Remember that this is a study to evaluate LLMs capabilities against human expert knowledge.

**Clear Question Stem:** write concise and unambiguous questions. Ensure each question focuses on a single concept or idea.

**Area:** think of questions within your area of expertise, particularly in those highlighted in the last email. For inspiration, we recommend reading your own publications or related publications and identifying which question/s are answered by the insights provided by that work. Be aware that questions should not be asking specific information from a document. For example:

× The question: "Which are the proteins identified in this study?" is not correctly formulated.

✓ A correctly formulated question would be: "What proteins have been identified as molecular partners of the Arabidopsis lncRNA ASCO?"

**Explicit fashion:** Specify exact concepts, processes, or elements being asked about. Include relevant context or constraints (without revealing the true answer). It should be clear what constitutes a complete answer for that question. For example:

× Non-explicit Question: "Are the LBD transcription factors involved in hormone crosstalk?" The question doesn't indicate which plant species to focus on, it doesn't specify which hormones to consider, it lacks temporal or developmental context and it could be answered with a simple yes/no without providing meaningful information.

✓ Explicit Question: "Which long noncoding RNAs from Arabidopsis thaliana have been linked to the epigenetic regulation of the FLC locus in response to cold?" It specifies the exact molecule type -long noncoding RNAs-, it names the precise species -Arabidopsis thaliana-, and it identifies the specific environmental condition, i.e. cold.

### Guidelines for crafting the answer and distractors

**Correct Answer:** You will create a single correct answer (option in the multiple choice). Ensure the correct answer is accurate, unambiguous, and based on current scientific consensus or widely accepted theories.

**Incorrect answers:** You will create two incorrect answers (options in the multiple choice) for every question. These should be plausible and scientifically grounded. Avoid using trivial or obviously incorrect options. Incorrect answers can include common misconceptions or closely related but incorrect concepts.

*A good tip for writing the incorrect answers is to first write the correct answer and then make slight changes from it that don't change the overall structure of the response but change the semantic and meaning, giving rise to a plausible incorrect answer.*

**Option Length:** Keep all options (correct answer + incorrect answers) similar in length and complexity to avoid hinting at the correct answer. This is intended to reduce biases related to length in model answering. This length can vary from question to question but should not vary along the same question.

**Number of Options:** There should be a total of 3 options; 1 correct option and 2 incorrect.

### Example 1:

#### Question:

What is the impact of inorganic phosphate (Pi) in Arabidopsis thaliana auxin homeostasis during root hair development and what are the key genes involved in this process?

#### Correct answer:

In Arabidopsis roots, low Pi in the soil triggers auxin synthesis and transport, enhancing root hair elongation to promote Pi uptake. Under low soil Pi, auxin synthesis is enhanced specifically in the root cap and transported (mostly by AUX1, PIN2, and PGP4) from the apex to the differentiation zone, specifically leading to an increase of auxin levels in trichoblasts. In response to the high auxin microenvironment, root hairs protrude from the root epidermis controlled by RHD6 and RSL1.

#### Incorrect answer 1:

In Arabidopsis roots, high Pi in the soil triggers auxin synthesis and transport, enhancing root hair elongation to promote Pi uptake. Under high soil Pi, auxin synthesis is enhanced specifically in the root cap and transported (mostly by AUX1, PIN2, and PGP4) from the apex to the differentiation zone, specifically leading to a decrease of auxin levels in trichoblasts. In response to the low auxin microenvironment, root hairs protrude from the root epidermis controlled by RHD6 and RSL1.

#### Incorrect answer 2:

In Arabidopsis roots, low Pi in the soil blocks auxin synthesis and transport, enhancing root hair elongation to promote Pi uptake. Under low soil Pi, auxin synthesis is blocked specifically in the root quiescent center and no longer transported (mostly by LAX3, PIN1, and AZG2) from the apex to the differentiation zone, specifically leading to a decrease of auxin levels in atrichoblasts. In response to the low auxin microenvironment, root hairs protrude from the root epidermis controlled by RHD6 and RSL1.

## Example 2:

### Question:

How is the transcription factor TCP15 involved in the thermomorphogenesis response in *Arabidopsis thaliana*?

### Correct answer:

TCP15 is required for optimal petiole and hypocotyl elongation under high ambient temperature. This TF influences the levels of growth-related genes, which are induced in response to an increase in temperature. TCP15 directly targets the gibberellin biosynthesis gene GA20ox1 and the growth regulatory genes HBI1 and PRE6. Several of the genes regulated by TCP15 are also targets of the growth regulator PIF4. PIF4 binding to GA20ox1 and HBI1 is enhanced in the presence of the TCPs proteins, indicating that TCP15 directly participates in the induction of genes involved in gibberellin biosynthesis and cell expansion by high temperature functionally interacting with PIF4.

### Incorrect answer 1:

TCP15 is required for optimal root elongation under high ambient temperature. This TF influences the levels of growth-related genes, which are induced in response to an increase in temperature. TCP15 directly targets the auxin biosynthesis gene YUC9. Several of the genes regulated by TCP15 are also targets of the growth regulator IAA6, indicating that TCP15 directly participates in the induction of genes involved in auxin biosynthesis and cell expansion by high temperature functionally interacting with IAA6.

### Incorrect answer 2:

TCP15 is required for optimal flowering under high ambient temperature. This TF influences the levels of flowering-related genes, which are repressed in response to an increase in temperature. TCP15 directly targets the gibberellin biosynthesis gene GA20ox1 and the flowering genes SOC1 and SPL3. Several of the genes regulated by TCP15 are also targets of the flowering regulator BRC1. BRC1 binding to GA20ox1 and SOC1 is enhanced in the presence of the TCPs proteins, indicating that TCP15 directly participates in the repression of genes involved in gibberellin biosynthesis and flowering by high temperature functionally interacting with BRC1.

## G Prompts

### MCQ answering prompt template

The following is a multiple-choice question. Think step by step and then provide your FINAL answer between the tags <ANSWER> X </ANSWER> where X is ONLY the correct letter of your choice. Do not write additional text between the tags.  
Question:\n{question}\n\nOptions:\n{options}\n\nAnswer:

### Open-ended answering prompt template

You are given a plant molecular biology question to answer. Respond concisely in one paragraph and provide a source (document title and link) for your answer.  
{question}

The following is the prompt instruction template used to generate Synthetic MoBiPlant.

### Prompt template for synthetic MCQ generation

**\*\*Objective\*\***  
Generate 5 independent multiple-choice questions (3 options each) from a published paper. Format them as a JSON array of objects with fields: question, correct\_answer, incorrect\_answer1, incorrect\_answer2, area (optional), and plant\_species (default: "non-specific").

**\*\*Step-by-Step Instructions\*\***  
- Identify Key Facts:  
Read the whole text to understand the work and carefully extract 5 distinct facts (e.g., conclusions, unique claims).  
If a fact involves a plant species, note its scientific name.  
Assign an area (must be one from: gene regulation, genome and genomics, cell biology and signaling, growth and development, hormones, physiology and metabolism, evolution, biotechnology or environment) if applicable; otherwise, use null.

# - Craft Unique Questions:

From the extracted facts, formulate questions that are clear, concise, and unambiguous:

- The five resulting questions should be different from each other, resulting in a unique set of questions.
- The questions should NOT be pointing to the study (examples not to follow: What happened in this study?, In the work done by XX, ... , According to the study/text ...). Instead, they should be phrased WITHOUT mentioning the document (e.g. What proteins have been identified as molecular partners of the Arabidopsis lncRNA ASCO?).
- The questions should not represent methods/techniques or specific experimental details (e.g. What is the relative timing of cell expansion versus cell division in ...). Rather, they should focus on the MAIN CONCEPTS or CONCLUSIONS (e.g. What is the role of the lncRNA ASCO in Arabidopsis?, At which level does HYL1 regulate gene expression?).

Note: For plant-related facts: Explicitly mention the species in the question (e.g., "What adaptation does Zea mays use to...?").

# - Generate Options:

Correct Answer: Directly derive from the text.

Distractors: Create two incorrect options by altering key details from the correct answer (e.g., species names, quantities, causal relationships, gene names).

# - Assign Metadata:

Tag area only if the question aligns with a clear discipline from the provided list, else use "null".

For plant species: Use exact scientific names (e.g., "Oryza sativa").

# - Validate:

Once chosen, carefully ensure questions are asked about specific conclusions or key concepts derived from the work. They should not be about methods or experimental details. They should be not referring to the text itself. Instead, they should be direct.

If any generated question does not meet the criteria, discard it and create a new one that does.

Confirm plant\_species is "non-specific" unless explicitly tied to a species.

# \*\*Output Format\*\*

Return a JSON array adhering to this schema:

```
{
  "$defs": {
    "Area": {
      "enum": [
        "GENE REGULATION",
        "GENOME AND GENOMICS",
        "CELL BIOLOGY AND CELL SIGNALING",
        "GROWTH AND DEVELOPMENT",
        "HORMONES",
        "PHYSIOLOGY AND METABOLISM",
        "EVOLUTION",
        "BIOTECHNOLOGY",
        "ENVIRONMENT"
      ],
      "title": "Area",
      "type": "string"
    }
  },
}
```

```

"properties": {
  "question": {
    "title": "Question",
    "type": "string"
  },
  "correct_answer": {
    "title": "Correct Answer",
    "type": "string"
  },
  "incorrect_answer1": {
    "title": "Incorrect Answer1",
    "type": "string"
  },
  "incorrect_answer2": {
    "title": "Incorrect Answer2",
    "type": "string"
  },
  "area": {
    "anyOf": [
      {
        "$ref": "#/$defs/Area"
      },
      {
        "type": "null"
      }
    ],
    "default": null
  },
  "plant_species": {
    "title": "Plant Species",
    "type": "string"
  }
},
"required": [
  "question",
  "correct_answer",
  "incorrect_answer1",
  "incorrect_answer2",
  "plant_species"
],
"title": "MCQSchema",
"type": "object"
}

```

## H Journal Distribution

We report the journal distribution of the Expert MoBiPlant set in Supplementary Table 5 and the models' performance on the journal performance experiment in Supplementary Table 6, along the journal distribution.

**Supplementary Table 5. Journal distribution of the Expert MoBiPlant QA set.**

Journal	Count
The Plant Cell	51
Nature Communications	35
Plant Physiology	30
New Phytologist	29
The Plant Journal	25
Proceedings of the National Academy of Sciences	24
Journal of Experimental Botany	21
Nucleic Acids Research	18
Nature Plants	16
Science	15
Current Biology	14
Annual Review of Plant Biology	13
Current Opinion in Plant Biology	12
Molecular Plant	12
The EMBO Journal	12
Trends in Plant Science	11
Nature	10
Cell Reports	10
Frontiers in Plant Science	8
eLife	8
Cell	8
PLOS Genetics	7
Plants	7
International Journal of Molecular Sciences	5
Plant Cell Reports	5
Molecular Cell	5
Genome Biology	5
BMC Genomics	4
Annals of Botany	4
Genes	4
Plant and Cell Physiology	4
Plant, Cell & Environment	3
Transgenic Research	3
PLOS ONE	3
Plant Molecular Biology	3
Development	3
Developmental Cell	3
BMC Plant Biology	3
Scientific Reports	3
Methods in Molecular Biology	2
Planta	2
Science Advances	2
Biochemical Society Transactions	2
Plant Science	2
PLoS ONE	2
Biochimica et Biophysica Acta (BBA) - Bioenergetics	2



Journal	Count
Annual Review of Genetics	2
Environmental and Experimental Botany	2
Nature Genetics	2
Molecular Plant Pathology	2
Current Opinion in Genetics & Development	2
Physiologia Plantarum	2
Plant Communications	2
Viruses	2
Horticulturae	2
Plant Biotechnology Journal	2
Symbiosis	2
Annual Review of Cell and Developmental Biology	2
Biological Reviews	2
Journal of Integrative Plant Biology	1
Nature Reviews Molecular Cell Biology	1
BMC Bioinformatics	1
Photochemical & Photobiological Sciences	1
Food Chemistry: X	1
Genome Research	1
Stress Biology	1
Virus Research	1
Journal of Virological Methods	1
GM Crops & Food	1
The Arabidopsis Book	1
PLOS Pathogens	1
Fungal Biology	1
FEMS Microbiology Ecology	1
Molecular Plant-Microbe Interactions®	1
Molecular Biology of the Cell	1
Frontiers in Microbiology	1
Journal of Fungi	1
FEMS Microbiology Reviews	1
Nature Chemical Biology	1
Nature Structural & Molecular Biology	1
Communications Biology	1
Plant and Soil	1
Microbiological Research	1
Environmental Microbiology	1
Annual Review of Cancer Biology	1
Journal of Cell Biology	1
Biochemical Journal	1
Functional Plant Biology	1
Journal of Biological Chemistry	1
Frontiers in Environmental Science	1
Cell Host & Microbe	1
Plant Transcription Factors	1
iScience	1
Genetics	1
Peptides	1
Nature Machine Intelligence	1
Biology	1
Evolutionary Applications	1
PLOS Biology	1
Science Signaling	1

Journal	Count
Cells	1
EMBO reports	1
Biomedical Reports	1
Cellular and Molecular Life Sciences	1
Journal of General Virology	1
Gene	1
Archives of Virology	1
Current Opinion in Biotechnology	1
Nature Biotechnology	1
EMBO Reports	1
Trends in Biochemical Sciences	1

**Supplementary Table 6. Journal distribution and model performance on the Synthetic MoBiPlant set.** Answer accuracy is reported for every model.

Journal	Count	LLaMA 3.1 405B	Claude 3.5 Sonnet	DeepSeek V3	DeepSeek R1	GPT-4o	o1-mini	Gemini 1.5 Pro
<i>COPB</i>	125	96,7	99,2	99,2	97,6	98,4	96,8	91,2
<i>TIPS</i>	135	96,3	97,8	97,8	97,8	97	96,3	95,6
<i>JExBot</i>	125	82,1	83,2	82,4	79,2	76,8	78,4	80
<i>Molecular Plant</i>	90	89,2	88,9	90	86,5	86,7	84,4	85,6
<i>Nature Plants</i>	55	75,9	72,7	80	75,9	80	72,7	80
<i>New Phytol</i>	145	84,8	76,1	79,6	78,2	82,4	82,4	80,3
<i>Plant Cell</i>	145	86,4	89	92,4	86,5	82,1	84,1	84,8
<i>Plant Journal</i>	130	90,4	89,2	90	87,3	86,2	78,5	81,5
<i>Plant Phys</i>	125	86,7	84,8	86,4	82,5	81,6	80,8	74,4
<b>Overall</b>	1075	87,3	88,4	89,6	87,3	85,7	84	83,9