RESEARCH ARTICLE

Keeping Cultivar Values in Their Ecological Type Zones: A Multimedia Knowledge Pipeline for Cultivar Selection and Daily Crop Management

Yusen Yang[1,2,3,4], **Youlin Cui**[1,2,3,4], **Peng Hao**[1,2,3,4] **and Huaji Zhu**[1,2,3,4,*]

[1] National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China
[2] Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China
[3] Key Laboratory of Digital Village Technology, Ministry of Agriculture and Rural Affairs, Beijing 100097, China
[4] Beijing Research Institute of Agricultural Artificial Intelligence and Robotics, Beijing 100097, China

Abstract

In China's national variety approvals, each cultivar's growth duration and disease grades are reported under the ecological type zone of its trials. A cultivar approved for several zones carries one set of values for each zone, and these values often disagree between zones. We describe a multimedia knowledge pipeline that keeps the type zone, assessment method and source with every value. Using the official approval descriptions of 5,885 rice and maize registrations and the 2025 announcement, we kept every grade and duration under the type zone that reports it. We then counted cultivar–disease pairs whose grades differ between zones, compared pooled and conservative shortlists with the answer for the intended zone, compared the pipeline's determinacy test with published retrieval policies that decide when to read the source passage, and predicted withheld grades from the cultivar's other zones. For more than a third of the cultivar–disease pairs documented in several zones, the official grades differ between zones, and the documented durations of one cultivar can differ by weeks from zone to zone. Pooling across zones adds unsupported cultivars to the 2025 shortlists, whereas requiring every profile to pass leaves out supported ones. The pipeline answers from stored profiles when the determinacy bounds agree and reads the source passage otherwise, and this rule reproduced every documentary answer in the 2025 workload. On sampled requests with a shared language-model reader, the rule made no errors and read the source passage for only 1.2% of requests. It also used far fewer generation tokens than the adapted retrieval policies we evaluated. In predicting grades across zones, a random forest beat the zone–disease majority grade only slightly, and the cultivar-level sign test did not settle the difference. Keeping the type zone with each value avoids the errors of pooled summaries and ties every answer to the record that supports it.

Academic Editor:
Muzammil Khan

Submitted: 09 July 2026
Revised: 19 October 2026
Accepted: 05 November 2026
Published: 17 November 2026

Vol. 3, **No.** 4, 2026.
10.62762/TIS.2026.382002

***Corresponding author:**
✉ Huaji Zhu
zhuhuaji007@nercita.org.cn

Citation
Yang, Y., Cui, Y., Hao, P., & Zhu, H. (2026). Keeping Cultivar Values in Their Ecological Type Zones: A Multimedia Knowledge Pipeline for Cultivar Selection and Daily Crop Management. *ICCK Transactions on Intelligent Systematics*, 3(4), 261–311.

© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

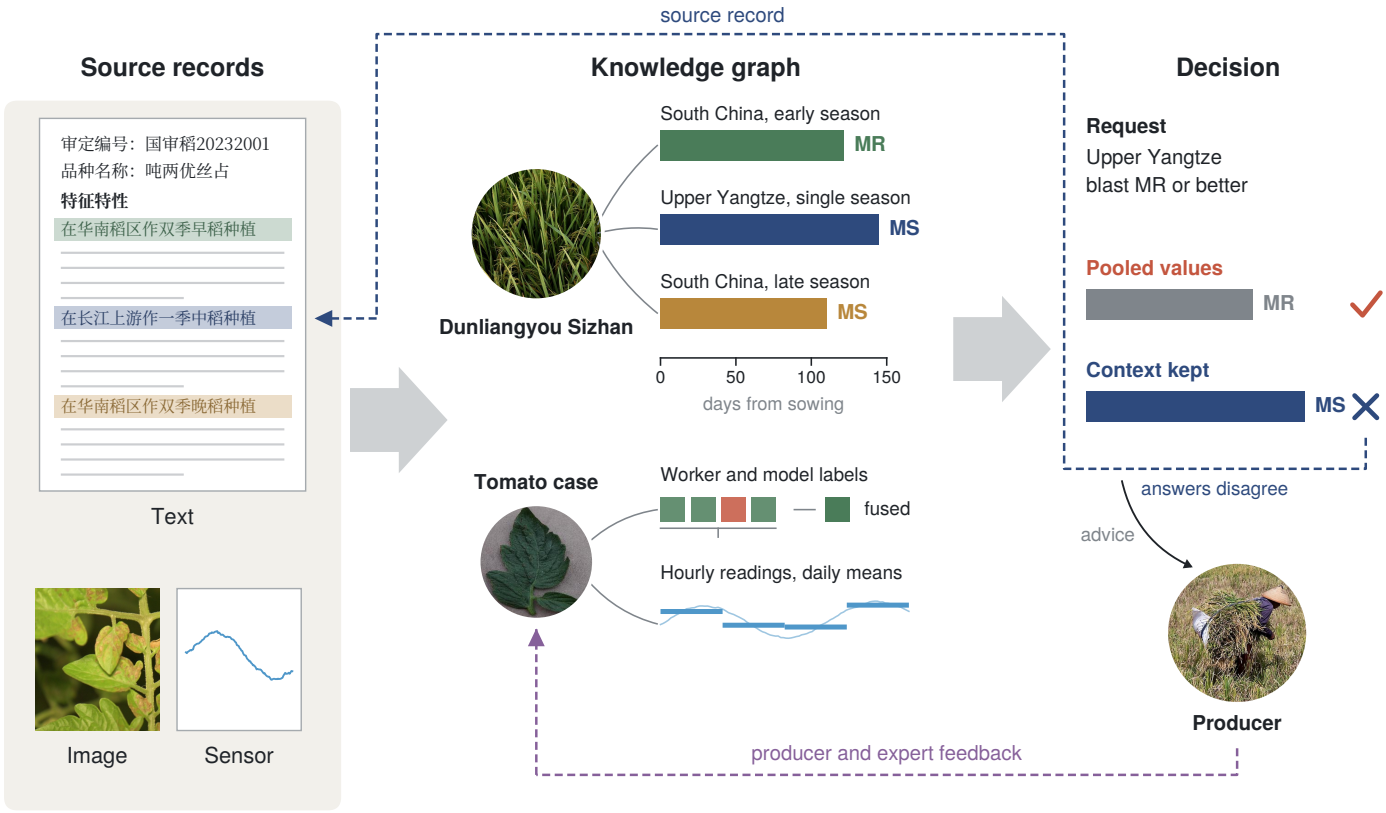


Figure 1. The type zone travels from the source record to the decision. The national approval of the rice cultivar Dunliangyou Sizhan reports one profile under each type zone, highlighted on the record page in the color of its type zone. The knowledge graph stores each profile as a growth-duration bar from sowing with its rice-blast grade at maturity (MR for moderately resistant and MS for moderately susceptible), together with a tomato case whose worker and model labels and hourly sensor readings keep their origin. An Upper Yangtze request for blast resistance of MR or better is answered twice. Pooled values combine the shortest duration and the best grade from different South China seasons and pass the request, whereas the Upper Yangtze profile fails it. The disagreement sends the query back to the governing type zone (dashed blue), and producer and expert feedback returns to the case (dashed purple). The tomato labels and sensor traces are schematic. Photographs are public examples from Shixart1985 (rice, CC BY 2.0), Crisco 1492 (producer, CC BY-SA 3.0), Wolan268 (field leaf, CC0) and PlantVillage [1] (laboratory leaf, CC BY-SA 3.0).

Keywords: smart agriculture, knowledge pipeline, cultivar selection, multimedia knowledge graph.

1 Introduction

Pathogens and pests are estimated to destroy 30.0% of the world's rice yield and 22.5% of its maize yield [2]. Choosing a cultivar fixes the genetics planted for the season, whereas its disease reactions and time to maturity vary with the growing environment. In China, the National Crop Variety Approval Committee approves rice, maize, wheat, soybean and cotton, and the Ministry of Agriculture and Rural Affairs publishes an announcement for each approval round [3]. The national regional trials behind an approval are organized by the National Agricultural Technology Extension Service Center, run for at least two production cycles, and compare each candidate with an approved cultivar that is being promoted in the same ecological type zone in the same period [3]. Every value in an approval description therefore belongs to the type zone named in the announcement [3], for example single-season middle rice of the Upper Yangtze or, for maize, the Huang-Huai-Hai summer maize zone [4]. The 2025 announcement lists 1,267 registrations, 1,189 of them rice or maize [5]. Extension centers, seed companies and large-scale growers read the characteristics section of these announcements to decide what to plant, and an approved cultivar may be introduced into another province only within the same suitable ecological zone and only after filing [3]. Some cultivars are tested in more than one type zone, and their descriptions then report one profile for each zone, where a profile is the group of values reported under one type zone [6, 7].

The approval record provides the evidence for this study. We read a registry of national approval descriptions with fixed extraction grammars that keep every duration and disease grade under the type zone that reports it, which leaves 5,885 rice and maize registrations with approval years from 2018 to 2024, and we checked the registry text against the 2023 and 2024 announcement documents and parsed the 2025 announcement from its own document (Section 5).

Dunliangyou Sizhan, a rice hybrid approved in 2023, shows what happens when those values are aggregated per cultivar. Its registration reports three profiles. As early rice in South China it reports 121.5 days and moderate resistance to blast, as late rice in South China 110.2 days and moderate susceptibility, and as single-season rice of the Upper Yangtze 144.7 days and moderate susceptibility (Figure 1). An extension officer seeking the earliest cultivar with at least moderate blast resistance for the Upper Yangtze must evaluate the third profile alone. A summary that combines the cultivar's shortest duration and its best blast grade across the three profiles returns Dunliangyou Sizhan at 110.2 days. The earliest documented Upper Yangtze cultivar meeting this resistance requirement is Jingyuanyou 5368, approved in 2024, with a duration of 146.1 days (Figure 2).

Differences between environments are the reason regional trials exist, and genotype-by-environment interaction is among the best studied problems in crop science [8–10]. Most recent work models the interaction to predict performance. The Genomes-to-Fields initiative has evaluated more than 4,000 maize hybrids in over 130 year–location combinations, and genomic prediction with environmental covariates predicted grain yield across environments with moderate accuracy [11]. Graph neural networks trained on 10,000 maize variety records from 14 test sites classified whether a variety should continue testing with 74.8% accuracy [12], and graph collaborative filtering ranked planting locations for 1,102 hybrids [13].

Cultivar information reaches growers through several channels, and most of them carry the published description rather than the trials behind it. Provincial extension centers recommend cultivars for their own zone, seed companies describe the cultivars they sell, and cultivar databases and mobile applications aggregate the approval descriptions into one record per cultivar. Agricultural knowledge graphs extract entities and relations from documents for question answering [14, 15], and language-model advisers answer growers from them [16]. Recent systems add time-aware retrieval [17] and clarifying questions for ambiguous requests [18], and advice adapted to the field raised rice yield and profit in a randomized trial in Nigeria [19]. Growers also describe their problems with photographs and incomplete questions, and consultations with extension experts combine text, images and follow-up questions [20]. A cultivar-selection tool therefore has to join documents, images, sensor series and expert judgments, and it has to keep for each of them the conditions under which it was obtained. The systems above decide which document, which time or which question a grower means. We examine the evidence inside one retrieved approval description, where the same cultivar carries a different grade and a different duration under each type zone.

We propose a multimedia knowledge pipeline for cultivar selection and daily crop management, and we evaluate its decision rule on published approval records. The pipeline stores each assertion with the qualifiers, method and source that fix where it applies (Section 3). A request is answered from the profile of the type zone it names. When a database holds the values of several profiles but not which type zone each belongs to, a determinacy test compares a conservative and an optimistic answer over the candidate profiles and reads the answer from a stored summary when the two agree. This is the certain-answer condition of incomplete-information databases [21–23], and Section 4 gives the rule and the test. We compare the shortlist rule with aggregation across zones, the common alternative, and with a conservative rule that requires every profile to pass. We also ask whether a model could replace the profile of the intended zone by predicting a withheld grade from the cultivar's other zones, using logistic regression, a random forest and gradient boosting, the standard learners of the multi-environment trial literature.

Field trials measure genotype-by-environment interaction directly, and that literature reports its size for yield. The present study examines the published disease grades and durations available to extension centers and growers. We count how often the official disease grades and durations of one cultivar differ between the zones that report them, how the differences change a shortlist, which requests they affect, and whether a grade in one zone predicts the

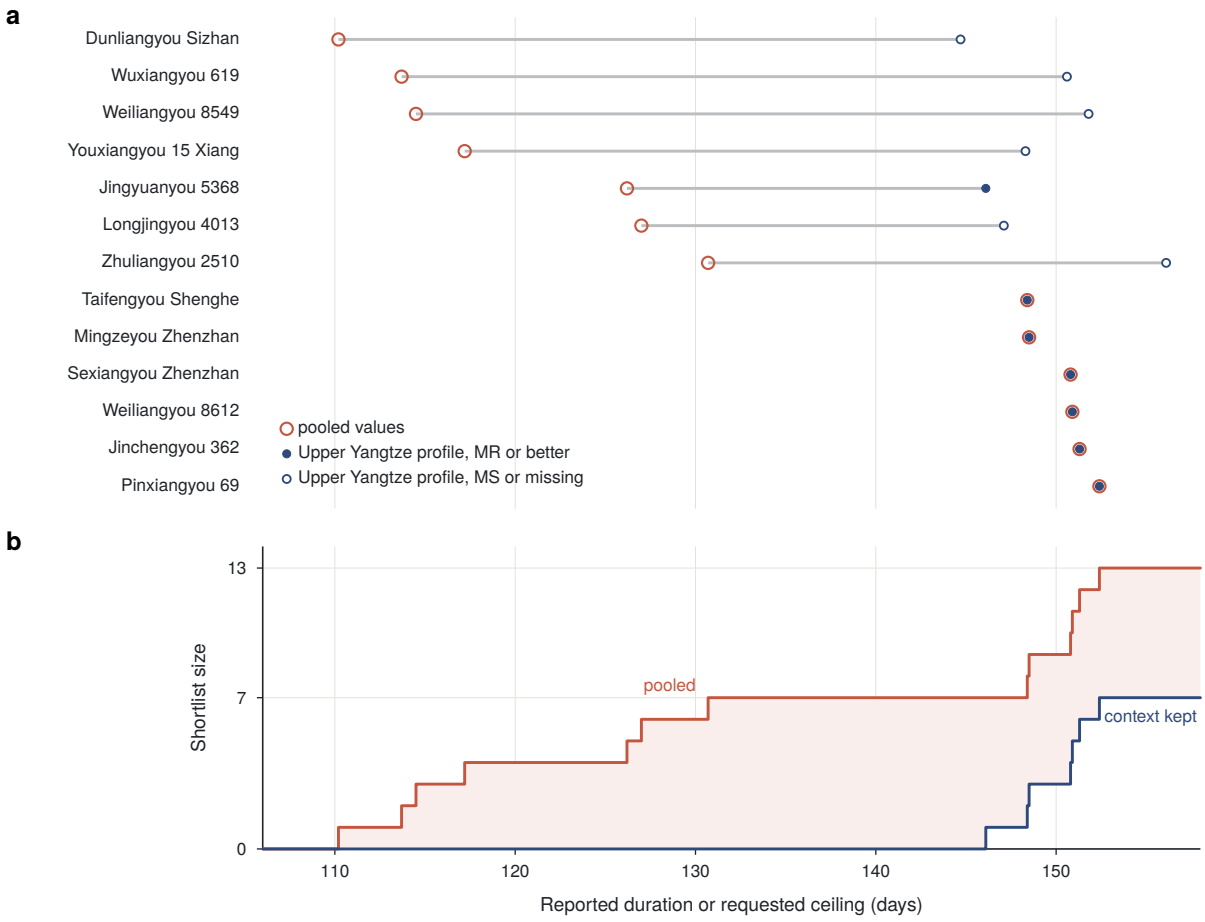


Figure 2. Pooling shortens the apparent Upper Yangtze shortlist. **a**, The thirteen rice cultivars that satisfy an Upper Yangtze single-season request for blast MR or better under pooling. Each row joins the pooled duration to the duration of the cultivar's Upper Yangtze profile, and the filled marker shows the seven profiles that document the requested resistance. **b**, Shortlist size as the requested maturity ceiling rises. The shaded area contains options that exist only under pooling. The first documented option appears 146.1 days after sowing, 35.9 days after the first pooled option.

grade in another. An approval description combines local pathogen populations, trial seasons, disease nurseries and inoculation protocols, so these counts describe the variability that published advice itself carries.

When a registration has several profiles without zone labels, the determinacy test answers a request directly whenever the conservative and optimistic bounds agree. Disagreement between the coarse bounds triggers further checking of the grouped profiles or the source record, whereas disagreement between the bounds over complete anonymous profiles establishes that the documentary answer depends on the intended type zone. Using common language-model readers, we compare this deterministic rule with the Sufficient Context autorater, an MBA-RAG adaptation and an A2RAG-style controller across documentary error rate, source-access frequency and recorded generation tokens, with the inputs and training budget of each router specified [24–26].

2 Related Work

Agricultural knowledge graphs link textual expressions to domain concepts [27, 28], and VEG-MMKG extends the construction to vegetable images and text [14]. Wheat variety approval texts have been parsed into entities and relations by in-context learning with a language model and by joint entity-relation extraction [29, 30]. Crop GraphRAG retrieves entity neighborhoods and summary passages to answer questions about crop pests and diseases [15, 31], and language-model systems built on domain knowledge, retrieval and tools cover several production scenarios [16, 32, 33]. TARAG annotates time and retrieves by season for pest and disease advice [17], while KGQA-CQG asks a clarifying question when a maize query is ambiguous [18]. A human-in-the-loop approach keeps a domain expert in the review of language-model farm-management insight [34], TRACE-KG builds context-enriched knowledge

graphs from complex documents [35], and a survey of data provenance covers how the production process of data is described and what that description is used for [36]. While those systems resolve which document, time or question a user intends, we examine how much of the evidence inside one retrieved registration belongs to a specific type zone. Adaptive retrieval methods decide when to consult a source from model confidence, context sufficiency, bandit routing or answer verification [24–26, 37, 38]. Because a cultivar request states its maturity and disease conditions explicitly, the stored profiles and the request suffice to test whether the missing zone label can change the answer.

Cultivar adaptation is studied in multi-environment trials, where genotype-by-environment interaction and climatic covariates explain why the ranking of cultivars changes between sites and seasons [8–10]. Genomic prediction and pedigree models rank untested maize hybrids across environments [11, 39, 40], and learned adaptability models rank planting sites from trait and climate features [12, 13]. Whereas those studies model trial performance, we examine the published approval ratings that extension officers and growers read, measuring how strongly they depend on the type zone and testing the same standard learners on them.

In agricultural knowledge representation, AGROVOC supplies multilingual concepts, the Semantic Sensor Network ontology describes observation procedures and observed subjects, and PROV-O records derivation and attribution [41–43]. Provenance semirings formalize how annotations pass through relational transformations [44], and certain answers from incomplete-information databases [21, 22] extend to certain predictions over incomplete feature records [23]. We apply this certain-answer condition to published variety approvals to count how often unlinked profiles leave a cultivar request undetermined.

Crowdsourcing models jointly estimate worker reliability and latent labels from noisy observations [45, 46]. Generated explanations can raise acceptance of algorithmic recommendations without improving decision accuracy [47]. Our simulated copying experiment examines how unmeasured dependence between crowd annotators and a model affects consensus in plant disease diagnosis. Case-based reasoning organizes irregular agronomic experience into retrievable episodes [48]. PlantVillage and PlantDoc benchmark disease identification in laboratory and field photographs [1, 49], and a survey reviews deep-learning recognition of leaf diseases [50]. AgMMU benchmarks multimodal agricultural reasoning, HiTab tests question answering over hierarchical tables, and AgriEval, AgReason, and AgriCoT evaluate reasoning in specialized language models [51–55]. AgroMind and MIRAGE add remote-sensing reasoning and multimodal consultations with extension experts [20, 56].

Mechanistic crop models link sensor observations to biological processes. FAO irrigation standards and the WOFOST engine model evapotranspiration, soil moisture and phenological development [57, 58]. Knowledge-guided machine learning combines process formulations with remote sensing for agroecosystem carbon accounting [59], and ground-level photography supports in-season rice yield forecasting [60]. Autonomous greenhouse trials have controlled climate, irrigation and crop production in cucumber, cherry tomato and dwarf tomato [61–63], reinforcement learning has been compared with model predictive control for greenhouse climate [64], and assimilating UAV observations into a crop model estimates crop water stress [65]. Our daily-management design states the evidence associations that connect these simulators to multimodal knowledge and the feedback of extension officers and growers.

3 The Multimedia Knowledge Pipeline

The Dunliangyou Sizhan registration holds three durations and three blast grades, and only the type zone says which of them an Upper Yangtze extension officer may use. The pipeline in Figure 1 keeps that type zone with every value. A request enters on the left of the figure, evidence is stored in the middle, and advice and feedback leave on the right. The unit of storage is a case anchored to a field plot or a plant group. A case binds text passages, diagnostic images and sensor intervals to the conditions under which they were obtained, and a variety approval is handled in the same way as an in-season management episode [3]. When a request depends on a condition that the stored summary lacks, the answer comes from the source record.

3.1 What a Request Must State

A request starts with a decision, and extracting the key factors of the user's need makes the production

unit, the intended action and the constraints explicit. A cultivar shortlist names the type zone, a maturity ceiling and the disease requirements. A tomato irrigation request names the rooting medium, recent watering, estimated crop demand and the water available. A language model parses the grower's wording into these conditions and flags what is missing before retrieval [18]. A daily irrigation calculation needs the crop stage or the sowing date used to estimate it. A missing type zone matters only for requests whose answer changes between the candidate profiles (Section 4). Personalized advice requires matching evidence to the grower's conditions [19].

3.2 Observations, Cases and Assertions with Qualifiers

An assertion is a statement extracted from a source. We represent assertion a by

$$a = (e, p, V, u, c, m, s, \ell). \tag{1}$$

Every assertion has one subject e, one property p, and a value or admissible value set V with a unit or category scale u. Every assertion has qualifiers c that state where and when it holds, for example the type zone of an approval profile or the greenhouse compartment of a tomato case, together with the spatial and temporal support of the observation, and a method m that states how the property was assessed. Every assertion has a source s and a location ℓ, and the location is a text span, an image region or a recording interval. Assertions from one reported assessment share an assessment identifier. The normalized value is stored together with the original source expression. Attaching qualifiers to a statement follows the data model of Wikidata, in which qualifiers refine the value of a statement [66].

Context says where an assertion applies, and provenance says how it was obtained. A derived assertion links to its parent assertions and to the activity or measurement that produced it [43]. A disease label copied from an automated suggestion links to that prediction, and a calculated soil-water state links to the sensor readings, recorded water applications and evapotranspiration model behind it. These links separate independent field observations from repeated transformations of one measurement. A graph edge, a table row or a linked case record can hold the same assertion with its qualifiers, as long as it keeps the type zone. In a tomato case, a visible lesion, an expert's diagnosis and a recorded chemical treatment stay separate assertions with their own sources.

Multimedia semantic recognition turns a photograph into assertions of the form in Equation 1. A model identifies the crop, the plant organ, the lesion and any visible event, and case-knowledge extraction groups the observations of one episode into a case. A case description keeps the raw image identifiers and separates what is visible from what a producer reports. Foundation and large language models have been reviewed for smart agriculture, crop production and agricultural extension [16, 67–69]. They have been built into agricultural question-answering systems [32, 33], a multimodal model for crop pest and disease diagnosis [70] and a vision-language benchmark [71]. We refer to them as complex agricultural large models when they carry out the language, vision and table tasks of the pipeline. Here Qwen3.8-27B extracts agricultural statistical tables [72], and Qwen3.8-27B and Qwen3.5-4B describe and answer multimedia cases [73]. Each output enters the record as an attributed interpretation linked to its source span. The source images and sensor series stay available for requests that need details a description omits.

Producers and experts label the same cases through crowdsourced multimedia annotation, and clustering groups the annotated cases for review. Crowdsourcing models estimate each annotator's reliability and the latent label jointly [45, 46]. Each contribution records the annotator, the camera settings, the lighting and whether a model suggestion was shown. A cluster of tomato leaves with similar foliar symptoms proposes a grouping, and a disease label still needs diagnostic annotation or clinical evidence. The tomato experiment tests how visual representations and simulated annotations change this grouping under a domain shift.

An expert's account of a difficult season does not reduce to a class label. Extracting this irregular experiential knowledge records the cue that prompted attention, the microclimate, the intervention and its timing, the outcome and any stated exception. An expert may report that canopy wilting persisted after irrigation in one tomato compartment because root disease impaired water uptake. The extracted case keeps the sequence of irrigation and wilting, which prevents a diagnostic rule from attributing the wilting to a simple water deficit. Assertions with

qualifiers and provenance links give later algorithms an auditable chain of evidence [35, 42–44].

3.3 Linking Evidence Across Media

Cross-media association mining links the photographs, text passages, tables and sensor series that describe the same subject or episode. Registration identifiers, plot codes and acquisition timestamps are strong association keys. When they are absent, name similarity, visual prompt captioning and semantic similarity propose candidate links [74–78], and the source evidence decides whether a link is accepted. Multi-semantic alignment then maps one expression to concepts at several levels. A tomato photograph carries crop, organ, symptom and disease concepts, and a management log adds cultural actions and growth stages. AGROVOC supplies the shared terminology [41]. Alignment keeps the distinctions that change agronomic meaning. Generic maize ear rot and pathogen-specific ear rot stay separate entities, and artificial inoculation and natural field screening stay separate methods even when they share a resistance vocabulary.

The agricultural cross-media knowledge graph that these links build connects subjects, observations, cases, concepts, actions and sources, and supports entity-grounded question answering [15, 16, 31]. A request is matched first to the production unit and conditions of a record, and then to compatible meanings, units and assessment scope. Missing conditions stay explicit. Resistance measured in an artificial nursery cannot silently satisfy a requirement for natural field resistance, and a sensor summary built from several readings carries its observation interval and depth-aggregation rule. The graph evolves as registrations, seasons and cases arrive. New assertions and links are added, and a revision of an existing assertion reaches the answers that depended on it (Section 3.5).

3.4 Summaries an Adviser Can Navigate

Big-data mining of agricultural topic knowledge organizes a large corpus around production problems, such as blast-resistant rice within a maturity window or tomato symptoms after irrigation. Topic indexes and clusters nominate candidate evidence, and its applicability is checked against the request before it supports an answer. A compact summary records which questions it can answer. Per-disease best grades answer separate-disease questions, and a request for one jointly qualifying assessment returns to the grouped records.

A multimedia knowledge tree gives an extension officer a hierarchy to navigate, from crop species and growth stage to symptoms, comparable cases and candidate actions. Cross-links between assessments and episodes stay in the graph, which a tree alone would hide. The visual expression of agricultural multimedia knowledge shows the observation, its production conditions and the source behind each recommendation, and marks an unresolved condition where it limits certainty.

3.5 Rules, Recommendation and Feedback

Construction of agricultural multimedia knowledge rules turns reviewed cases and published guidance into conditional statements. A rule records its crop, developmental stage, soil conditions, proposed action and supporting evidence. A stage-specific fertilizer prescription is a normative recommendation, and direct observations of plant uptake are the physical evidence for demand estimation.

Human–machine hybrid knowledge recommendation combines retrieval and explicit condition checking with expert interpretation [34]. A recommendation assembled this way lists the supported options and the evidence that separates them, and experts resolve ambiguous observations, judge environmental exceptions and confirm that an action suits the place. Crowd decision-making from multi-source knowledge weighs each contribution by its source and by its dependence on the others. Several labels copied from one model prediction count as one suggestion.

Knowledge learning and feedback record the corrections, the performed actions and the field responses of agricultural experts and producers. An accepted piece of advice is stored with the field parameters, and the grower's later crop response enters the case history. A corrected observation triggers re-evaluation of the states, recommendations and rules that depended on it, while the earlier assertion stays with the source and the reason for its revision. Each recommendation remains attached to the evidence available when it was issued.

Self-learning iterative evolution of agricultural multimedia knowledge proposes updated labels, case clusters, rules or model parameters from the linked records. Experts set the scope of a rule change,

and evaluation uses episodes outside those that built it. A producer's agreement with a displayed suggestion shows interface adoption, and diagnostic accuracy needs an observed biological outcome or an independently verified label. The tomato experiment isolates the dependence introduced by copying a suggestion. Expert verification and outcome feedback limit the drift that unreviewed automated updates could introduce. Concept drift is a change over time in the relation between a model's inputs and its target, and adapting to it requires detecting the change [79].

The tomato, image-and-text and table experiments of Section 6 test recognition, annotation and evidence reuse. The next section defines the stage that answers a cultivar request from the stored profiles, which the approval record tests directly.

3.6 Extraction and Revision Procedure

In the evaluated recognition stages, tomato images pass through a frozen ResNet-18 encoder and a fitted linear classifier whose calibrated probabilities [80] enter the worker-pooling and clustering procedures in Appendix B. Multimodal case descriptions and table extractions use the prompts in Appendix E under the model and decoding settings in Appendices B.4 and F. Generated case descriptions enter the graph as attributed interpretations, whereas the approval analysis extracts literal text spans with fixed heading grammars.

Algorithm 1 joins records by exact identifiers, routes uncertain links to expert review and propagates corrections along derivation links. ExtractAssertions returns the fields in Equation 1, including the literal source expression. NormalizeTerms adds vocabulary terms without replacing that expression. FindCandidateLinks proposes links from identifiers or similarity. A link is accepted automatically only when registration or episode identifiers agree and its scope is compatible; sensor links require compatible production units, areas, depths and recording intervals. AppendVersion preserves the earlier assertion and records the update's source, reviewer and reason. Rule or model updates require expert approval after evaluation on separate episodes.

In this workflow, sensor aggregation requires a common production unit and recording interval, while duration and disease conditions require a shared type-zone profile and a compatible assessment method.

4 Shortlist Rule and Determinacy Test

The pipeline of Section 3 stores every value with the type zone that produced it, so a request can be answered from the profile it names. Such a request names a crop, a type zone, a duration ceiling and the diseases that a cultivar must resist. A cultivar is on the shortlist of the request if it has a profile for the type zone of the request, if the duration in that profile does not exceed the ceiling, and if the profile reports an acceptable grade for every requested disease. A cultivar whose profile lacks a requested grade is not on the shortlist. Figure 2 shows what happens when a cultivar database skips the first condition. It combines the shortest duration and the best blast grade that Dunliangyou Sizhan reports under different type zones [6], although the Upper Yangtze request needs the duration and the grade that the cultivar reports together for the Upper Yangtze.

4.1 Support Within a Production Profile

Every registration r has a nonempty set $\mathcal{K}_r$ of profiles, and every profile belongs to exactly one type zone. A profile k reports a duration $v_{r,k}$ with an endpoint $b_{r,k}$, for example emergence to maturity or emergence to fresh-ear harvest. For every disease d, the profile reports a set $V_{r,k,d}$ of admissible grades, and the set is empty if the registration gives no grade for d. The grades run from highly resistant to highly susceptible. A request q names an intended zone c_q, an endpoint b_q, a ceiling T_q, a set $\mathcal{D}_q$ of diseases and, for every disease d in $\mathcal{D}_q$, a set $A_{q,d}$ of acceptable grades.

A profile meets the duration condition if its endpoint equals b_q and its duration is at most T_q. A profile meets the condition of a disease if the grade set of the profile is nonempty and lies inside the acceptable set. With the index 0 for the duration condition,

$$B_{r,k,0}(q) = \mathbf{1}[b_{r,k} = b_q \text{ and } v_{r,k} \le T_q],$$
$$B_{r,k,d}(q) = \mathbf{1}[V_{r,k,d} \neq \varnothing \text{ and } V_{r,k,d} \subseteq A_{q,d}], \quad d \in \mathcal{D}_q. \tag{2}$$

A missing duration gives $B_{r,k,0}(q) = 0$ in Equation 2. A profile supports the request if it meets every condition. With $\mathcal{J}_q = \{0\} \cup \mathcal{D}_q$, the support of profile k is

$$f_{r,k}(q) = \prod_{j \in \mathcal{J}_q} B_{r,k,j}(q). \tag{3}$$

Joint support in Equation 3 means that the duration and all disease grades come from the same profile,

Algorithm 1 Source extraction, rule checking and knowledge revision

Input: `sources`, `proposedRules`, `updates` (corrections or post-action outcomes), knowledge `graph`, `vocabulary`, development and evaluation episodes
Output: Updated `graph`, `ruleResults` and `reviewQueue`

```
ruleResults, reviewQueue ← empty lists
Extract assertions and check their links
for each source in sources do
    provenance ← source identifier, version, acquisition time and production unit
    assertions ← ExtractAssertions(source)
    for each assertion in assertions do
        assertion ← NormalizeTerms(assertion, vocabulary)
        acceptedLinks ← empty list
        for each link in FindCandidateLinks(graph, assertion) do
            if the linked records have matching identifiers and compatible scope then
                Append link to acceptedLinks
            else
                Append (assertion, link) to reviewQueue
            end if
        end for
        Store assertion, provenance, acceptedLinks and derivation links in graph
    end for
end for
Check proposed rules against the available evidence
for each rule in proposedRules do
    Store the rule's conditions, action, production system and supporting evidence
    if the rule's agronomic scope lacks expert approval then
        Append rule to reviewQueue; continue
    end if
    result, missingInputs ← CheckConditions(rule, graph)
    Append (rule, result, missingInputs) to ruleResults
end for
Propagate corrections and evaluate proposed changes
for each update in updates do
    revised ← AppendVersion(graph, update)
    affected ← states and advice derived from update.assertionId
    Recompute affected in dependency order using revised
    candidate ← propose a rule or model update using development episodes
    evaluation ← evaluate candidate on separate evaluation episodes
    if ExpertApproves(candidate, evaluation) then
        Replace the active rule or model with candidate
    end if
end for
return (graph, ruleResults, reviewQueue)
```

which is how the national maize approval standard evaluates a regional trial [4]. If the profile for the intended type zone is known, denote it by $k^{\star}$. The value $s_r(q) = f_{r,k^{\star}}(q)$ is the documentary answer to the request, and the evaluation treats it as the reference. In the evaluated workload the registration is correct and one of its profiles belongs to the requested zone.

4.2 Requests That Depend on the Type Zone

A cultivar database sometimes holds the values of all profiles of a registration but not which type zone each

belongs to. It then cannot tell which profile belongs to the intended zone. Let $\mathcal{K}_r^{b_q}$ be the set of profiles whose endpoint equals b_q. The intended profile is one of them, and each profile in this set has a known duration. The conservative answer $L_r(q)$ is 1 if the longest duration in $\mathcal{K}_r^{b_q}$ is at most T_q and every profile of r has an acceptable grade set for every requested disease. The optimistic answer $U_r(q)$ is 1 if the shortest duration in $\mathcal{K}_r^{b_q}$ is at most T_q and, for every requested disease, some profile has a grade set that overlaps the acceptable set:

$$\begin{aligned} L_r(q) &= \mathbf{1}\left[\max_{k\in\mathcal{K}_r^{b_q}} v_{r,k} \le T_q\right] \prod_{d\in\mathcal{D}_q} \min_{k\in\mathcal{K}_r} B_{r,k,d}(q), \\ U_r(q) &= \mathbf{1}\left[\min_{k\in\mathcal{K}_r^{b_q}} v_{r,k} \le T_q\right] \\ &\quad \times \prod_{d\in\mathcal{D}_q} \mathbf{1}[\exists k \in \mathcal{K}_r : V_{r,k,d} \cap A_{q,d} \neq \varnothing]\,. \end{aligned} \tag{4}$$

The optimistic answer lets each condition find a favorable value in a different profile. Because $k^\star$ belongs to $\mathcal{K}_r^{b_q}$,

$$L_r(q) \le s_r(q) \le U_r(q). \tag{5}$$

The conservative conditions include those of the intended profile. An intended profile that passes has a passing duration and a grade in every acceptable set, so the optimistic conditions hold as well. Partial grade sets obey the same ordering.

If $L_r(q) = U_r(q)$ in Equation 5, every endpoint-compatible intended profile has the same documentary answer. We call that request zone invariant for r, and its answer follows directly from the stored values [21–23]. If $L_r(q) < U_r(q)$, the coarse summaries leave the answer undetermined. This disagreement is a sufficient reason to inspect the grouped evidence or recover the source association. It does not establish that two complete profiles give different answers, because the optimistic summary can combine passing values from several failing profiles.

Anonymous profiles keep the values of one profile together without the type zone. They give tighter bounds,

$$\begin{aligned} L_r^{\text{profile}}(q) &= \min_{k\in\mathcal{K}_r^{b_q}} f_{r,k}(q), \\ U_r^{\text{profile}}(q) &= \max_{k\in\mathcal{K}_r^{b_q}} f_{r,k}(q), \end{aligned} \tag{6}$$

and $L_r \le L_r^{\text{profile}} \le s_r \le U_r^{\text{profile}} \le U_r$. The coarse upper bound asks whether each condition is met by some profile. The profile upper bound asks whether one profile meets all conditions together. The two differ when two profiles each meet part of the request. Appendix A.5 gives an example and the treatment of partial grades. The coarse bounds can disagree although every profile fails the request, so profile bounds identify the requests that depend on the type zone more exactly. Section 6 reports the number of affected requests.

4.3 Computable Workflow

Algorithms 2 and 3 evaluate a request directly from its maturity ceiling and disease thresholds using the six-level grade scale and regional disease panels in Appendix A, without fitted parameters. The request specifies the crop, type zone, duration endpoint, maximum duration and allowed grades for each disease. `PASS`, `FAIL` and `UNKNOWN` denote a satisfied condition, a contradicted condition and missing or ambiguous evidence, respectively. A failed condition excludes the cultivar; otherwise, an unknown condition leaves it unresolved. Both outcomes yield zero documentary support $s_r(q) = 0$.

MatchProfiles selects profiles with the requested crop, type zone and duration endpoint. MetadataMatch checks the subject, disease, assessment method and requested qualifiers against the source. A grade set containing both allowed and disallowed grades leaves the disease check `UNKNOWN`. Passes returns `TRUE` only when the duration and all disease conditions in one profile pass Equation 2; missing or unresolved evidence returns `FALSE`. RecoverProfile returns the unique matching source profile with a known duration and literal source spans, or `MISSING` if recovery fails or remains ambiguous. In Algorithm 3, `lower` and `upper` are the conservative and optimistic Boolean bounds in Equations 4 and 6.

Coarse and profile modes are evaluated separately using Equations 4 and 6, examining at most $|\mathcal{K}_r|(|\mathcal{D}_q|+1)$ conditions per registration. Scope checks use the registration's zone inventory without revealing which anonymous profile belongs to each zone.

For source recovery, the template baseline splits the passage at recognized rice or maize zone headings and matches each anonymous profile to the unique segment containing its duration and every supplied resistance quotation. Ambiguous matches remain

Algorithm 2 Shortlisting with known type zones

Input: Planting `request`; `registrations` with source-linked profiles
Output: Lists of `supported`, `excluded` and `unresolved` registrations

```
supported, excluded, unresolved ← empty lists
for each record in registrations do
  Identify the requested production profile
    matches ← MatchProfiles(record, request)
    if count(matches) ≠ 1 then
        Append record to unresolved; continue
    end if
    profile ← the single profile in matches
  Check duration and disease requirements
    checks ← UNKNOWN for duration and each requested disease
    if profile.duration is known then
        checks[duration] ← PASS if profile.duration ≤ request.maxDays, else FAIL
    end if
    for each disease in request.diseases do
        grades ← profile.grades[disease]
        allowed ← request.allowedGrades[disease]
        if MetadataMatch(profile, disease, request) and grades is nonempty then
            if every grade in grades belongs to allowed then
                checks[disease] ← PASS
            else if no grade in grades belongs to allowed then
                checks[disease] ← FAIL
            end if
        end if
    end for
  Assign the registration to one output list
    if any entry in checks is FAIL then
        Append record to excluded
    else if any entry in checks is UNKNOWN then
        Append record to unresolved
    else
        Append (record, profile, profile.sourceSpans) to supported
    end if
end for
return (supported, excluded, unresolved)
```

unresolved. The common-reader experiment instead asks the fixed language model for a decision from that same passage.

5 Data and Analyses

Sections 3 and 4 describe the record and the decision rule, and Section 7 extends both to daily management. This section describes the data on which we test the record and the rule. The official rice and maize approvals show how strongly cultivar grades and durations depend on the type zone, what a shortlist loses when a summary pools the zones, and whether grades from other zones can replace the profile of the intended zone. Tomato annotation, multimodal case descriptions, agricultural tables and two field datasets test the other stages of the pipeline. The subsections follow the order of Section 6, after the description of the approvals and the cohorts used in the approval analyses.

5.1 Official Approvals

The Ministry of Agriculture and Rural Affairs announces the national rice and maize approvals each year, and every announcement carries an

Algorithm 3 Determinacy testing and selective recovery

Input: `request`; anonymous `profiles`; `zoneInventory`; `recordId`; `mode` (`COARSE` or `PROFILE`)
Output: Boolean answer or `UNKNOWN`, supporting evidence, source-access status

Check that the request is within the recorded scope
`eligible` ← profiles whose duration endpoint matches `request.endpoint`
if `request.zone` is absent from `zoneInventory`, or `eligible` is empty, or any eligible duration is missing **then**
  return `(UNKNOWN, empty, UNRESOLVED)`
end if
Compute bounds without assigning profiles to type zones
if `mode = COARSE` **then**
  `lower` ← every profile in `eligible` has duration ≤ `request.maxDays`
  `upper` ← some profile in `eligible` has duration ≤ `request.maxDays`
  for each `disease` in `request.diseases` **do**
    `allowed` ← `request.allowedGrades[disease]`
    `allPass` ← every profile in `profiles` has nonempty `grades[disease]` contained in `allowed`
    `anyPass` ← some profile in `profiles` has a grade in both `grades[disease]` and `allowed`
    `lower` ← `lower AND allPass;  upper` ← `upper AND anyPass`
  end for
else
  `lower, upper` ← `TRUE, FALSE`
  for each `profile` in `eligible` **do**
    `answer` ← Passes(`profile, request`)
    `lower` ← `lower AND answer;  upper` ← `upper OR answer`
  end for
end if
Return a determined answer or recover the source profile
if `lower = upper` **then**
  return `(lower, profiles, NO_SOURCE_ACCESS)`
end if
`passage` ← ReadSource(`recordId`)
`(profile, spans)` ← RecoverProfile(`passage, request`)
if `profile` is `MISSING` **then**
  return `(UNKNOWN, spans, UNRESOLVED)`
end if
Cache(`recordId, request.zone, passage.version, profile, spans`)
return `(`Passes(`profile, request`)`, spans, SOURCE_RECOVERED)`

approval description for each cultivar [5–7]. A description lists the cultivar, its parentage, its yield performance, its cultivation techniques and its characteristics. The characteristics give one profile for each type zone in which the cultivar was tested. A profile is the group of values reported under one type zone. A type zone is the regional trial group of the approval standard, which fixes the region together with the cropping season and pattern, for example single-season middle rice of the Upper Yangtze or Huang-Huai-Hai summer maize. Durations cover the whole growth period for rice and the period from emergence to maturity for grain maize. Disease grades follow a scale of six steps from highly resistant to highly susceptible.

Two sources supply the records of this study. The first is a registry of national approval descriptions collected from a public variety-approval query service. It contains 9,960 registrations across maize, rice, wheat, soybean and cotton, of which 6,326 rows and 6,280 registration numbers are rice or maize, with approval years from 2018 to 2024. We read its characteristics fields with fixed grammars for the type zones, the duration inside each type zone and the disease grades, and every extracted value keeps the character span it came from. The extraction leaves 25,486 disease-grade records in 5,885 registrations of 5,575 cultivars under 64 type zones, after setting aside records with no recognized type zone and type zones whose duration is missing or ambiguous. The second source is the set of official announcement documents. We compared the 2,683 characteristics fields of the 2023 and 2024 announcements with the registry text, and 2,680 match exactly (Appendix A).

We parsed the 2025 announcement from its own document, because the registry does not contain it, and no registration number appears in both sources.

The analysis of differences between zones uses the whole extracted registry, so it covers the development cohort and the earlier approval years but not the test cohort. Every record carries the type zone under which it is reported. The shortlist analyses use two cohorts. The announcements of 2023 and 2024 form the development cohort. Of the registrations whose registry text matches those two documents exactly, 1,249 rice and maize registrations have a profile in the six zones evaluated here, which gives 1,280 profiles. The disease vocabulary, the grade grammar and the type-zone and duration parsers were developed using these documents. The 2025 announcement forms the test cohort. It lists 1,267 registrations, 1,189 of them rice or maize. Fourteen of these lack a recognized type zone, and 24 profiles have a missing or ambiguous duration. The remaining 1,200 profiles include 567 profiles of 558 registrations in the six zones evaluated here, and the other profiles of these registrations are available to the bounds of Section 4.

Rice requests cover Upper Yangtze single-season rice, Middle and Lower Yangtze single-season rice and Middle and Lower Yangtze double-season late rice, with panels of blast alone, blast with bacterial blight, or blast with brown planthopper. Maize requests cover Eastern and Northern middle-late spring maize, Huang-Huai-Hai summer maize and Southwest middle- and high-altitude spring maize, with the regional diseases named in the national approval standard [4]. Appendix A lists every panel. Two criteria are evaluated, at least moderate resistance (MR or better) and no highly susceptible grade (HS). For maize these are the approval standard's own thresholds. Its basic condition requires that natural field infection and artificial inoculation identification do not reach highly susceptible for the diseases of the zone, and its condition for a disease-resistant cultivar requires at least moderate resistance to every disease identified in the zone [4]. The rice panels use the same two criteria. Every maturity duration observed in a zone sets one ceiling, and every eligible cultivar is evaluated at every ceiling for its zone and panel. A missing grade withholds the cultivar from the supported shortlist. The documentary answer to a request is the answer computed from the profile of the intended zone (Section 4). The test cohort yields 157,522 candidate–request pairs.

5.2 Differences Between Zones

A pair is a cultivar and a disease with grades in two or more zones, and 1,772 pairs of 567 cultivars qualify. The grades of a pair differ if they are not identical. The pair crosses the resistance boundary if one zone grades the cultivar MR or better and another grades it MS or worse. For durations, a cultivar with two or more zones under the same endpoint contributes the range from its shortest to its longest duration. We report counts and shares for the whole registry, by crop and by disease, and we show a disease only if it has at least 25 pairs. Cultivar families contribute several pairs, and we therefore also report the number of cultivars.

5.3 Requests and Associations

The transgenic-maize approvals of 2023 and 2024 [81, 82] serve to compare requests of different kinds. All 64 official characteristics passages match the registry after normalization. Four records of two cultivar families were used to develop the parser, and 59 of the remaining 60 registrations report a grade from artificial inoculation identification and a grade from natural field infection, the two assessments that the maize approval standard requires separately [4], for two or three diseases within one zone. Their 385 distinct grade matrices give 1,155 cases at the cutoffs highly resistant, resistant and moderately resistant.

Five requests are posed over the same grades in every case. Does any complete assessment qualify? Do all assessments qualify? Does each disease have an acceptable grade somewhere? Does the named inoculation assessment qualify? Does the named field assessment qualify? We remove the link between a grade and its assessment, which leaves an unordered pair of grades for each disease, and we enumerate every assignment of these grades to the two assessments. A request is supported if every assignment satisfies it, excluded if no assignment does, and unresolved otherwise. Values and cutoffs stay fixed, so the unresolved cases measure the information that the link carries. A second enumeration lets the cutoff differ between diseases and compares the complete assessments with a summary that keeps the best grade of each disease (Appendix A.6).

The association ablation changes one factor at a time on the test cohort, namely whether the duration stays linked to the named zone and whether the disease grades stay linked to it. Without the duration

link, the rule takes the shortest duration among the profiles with the requested endpoint. Without the grade link, each disease may be satisfied by a different profile with that endpoint. A fifth variant keeps complete anonymous profiles and accepts a cultivar if any one profile meets the request. The shortlist computed with both links serves as the reference across all six type zones. We report coarse-bound disagreements as flagged requests and complete-profile disagreements as zone-dependent requests, summarized by registration and profile (Section 4).

5.4 Comparison with Retrieval Policies

A database that holds the profiles of a registration without their zone labels has to decide, for each request, whether to answer from the stored values or to read the source passage. Published retrieval methods make this decision from model signals. We compare four of them with the determinacy test of Section 4 on the same evidence. The Sufficient Context autorater [24] asks a model whether the stored values suffice. A sufficiency-and-confidence router combines that judgment with the confidence of the reader and a target share of requests that read the source. An adaptation of MBA-RAG [25] learns from answer accuracy and access cost which requests go to the source. An A2RAG-style controller [26] verifies the answer and recovers the source text when the verification fails. Two baseline readers bracket these policies, one reading only the anonymous profiles and the other reading the complete source passage for every request. All reader arms use Qwen3.8-27B with a 2,048-token output limit on identical profile or source evidence. For routing, the Sufficient Context autorater inspects the request and the anonymous profiles, whereas the MBA-RAG router receives the cultivar identity and the request under the information and training budgets in Appendix F. The determinacy test routes by either coarse-bound or complete-profile agreement, executing the shortlist rule directly when the bounds agree and passing the source passage to the reader when they differ.

The sample has 1,000 requests from 229 registrations. It contains 400 requests whose coarse bounds disagree, 300 other requests on registrations with several profiles and 300 requests on registrations with a single profile, and results are weighted to the 157,522 requests of the test cohort. We report the error rate, the share of requests that read the source passage, and the input and output tokens per request. Learned routers are fitted, calibrated and evaluated on disjoint sets of registrations. Invalid outputs and abstentions count as errors. Token counts cover model calls and exclude feature preparation and router training. A paired comparison crosses two ways of choosing which requests read the source, the multi-profile rule and profile agreement, with two ways of answering, a model reader and the shortlist rule itself, so that the origin of a saving can be identified (Appendix A.3).

5.5 Prediction Across Zones

A pooled summary and a model of genotype-by-environment interaction both assume that a cultivar's grade in one zone informs its grade in another. We test this assumption against standard approaches. Each of 3,751 instances is a cultivar, a disease and a zone with an official grade, taken from the pairs documented in two or more zones. The grade is withheld. The inputs are the cultivar's grades for that disease in its other zones (minimum, maximum, mean, median, range and count), the disease, crop and zone, and the approval year.

Five rules serve as baselines. They are the global majority grade, the majority grade of the zone and disease among the training cultivars, and the median, best and worst grade of the other zones. The last two are the summaries that pooling applies. Three standard learners follow the multi-environment trial literature, namely logistic regression, a random forest [83] and gradient boosting. Cultivars are assigned to five folds at random, so that no cultivar appears in both training and test data, and the assignment is repeated with five seeds. We report exact accuracy, accuracy within one grade, accuracy on the resistant or susceptible side and mean absolute grade error, as means and standard deviations over the seeds. The best learner is compared with a baseline by an exact sign test over cultivars, using each cultivar's accuracy averaged over the seeds. No learner is tuned on the test folds (Appendix A.4).

5.6 Tomato Annotation and Clustering

The tomato experiment extracts fixed ResNet-18 image features and fits a linear classifier on them [84]. Exact pixel duplicates are removed before the PlantVillage images are divided into training, validation and test sets of 10,875, 1,919 and 3,203 images. PlantDoc supplies 657 external field photographs [1, 49]. The classes are ten disease and

healthy categories, and nine of them occur in the external photographs.

Thirty-two simulated annotators differ in class accuracy, domain specialty and item difficulty [45, 46]. Ten random seeds vary the annotator responses, the nominal error rates and the annotation budgets. In the copying condition, an annotator replaces the response with the model prediction with probability 0.3. Within a condition, every fusion method receives the same responses. The weights for the external photographs are calibrated on a separate external validation subset. The main comparison uses five responses per image and a nominal error rate of 0.3. Case clustering concatenates the image features with model, worker or pooled probabilities and uses five seeds (Appendix B). Results are means and sample standard deviations over seeds, and paired comparisons use exact sign-flip permutation tests.

5.7 Case Descriptions and Tables

AgMMU tests whether a saved description of a multimodal case keeps the information that later agricultural questions need [51]. It belongs to a group of diagnostic vision-question suites and domain evaluations [53–55, 71]. After data cleaning, 768 cases of the public release remain. Qwen3.8-27B and Qwen3.5-4B write a description of each case without seeing the question or the answer options. Later calls answer from the background text, the raw images, the generated description, or a combination of these [72, 73]. A reasoning follow-up and a crossed design, in which one model writes the description and another answers, use the same saved descriptions. We report answer counts and paired net corrections.

The HiTab analysis uses 205 questions about 65 agricultural tables from 12 official reports [52]. It compares the raw table hierarchy, an exact expansion into coordinates, direct model extraction and source-checked extraction, each under greedy and reasoning settings. Every table is extracted once before the questions are shown (Appendix B.4). The description, answering and extraction prompts appear verbatim in Appendix E.

5.8 Field Allocation in Maize and Potato

The two field datasets check the temporal and spatial alignment rules that Section 7 states for a daily recommendation. Both allocate a fixed budget over a candidate pool that the record defines, and both restrict a forecast to the observations that existed when the decision was made.

The maize reanalysis uses Genomes-to-Fields records [11, 40, 85–88]. It selects five hybrids in held-out trials. Parental links, which extend single-cross and pedigree methods [89–91], enlarge the set of hybrids that can be ranked. A second study gives all models the same pool of hybrids with markers and compares genomic ridge regression, GBLUP and machine-learning models [39, 83, 92–96]. In the potato reanalysis, aerial multispectral indices and physiological measurements are linked to individual plants and dates, and a fixed scouting budget is allocated against later disease severity [97–99]. Feature contributions are assessed with Shapley values for tree models [100]. We report selected yields and their spread across locations, and Appendix C gives the candidate definitions, feature sets and protocols.

6 Results

The results follow the questions of the Introduction. We first measure how far official grades and durations differ between type zones, then show what a summary that pools the zones does to a shortlist and which requests depend on the zone at all. We then compare the test that finds those requests with published retrieval policies. The query families and the association ablation identify the links that carry an answer, and the prediction experiment asks whether a model can supply a profile that the record lacks. The last three subsections report the annotation, evidence-reuse and field-allocation stages.

6.1 Official Grades and Durations Depend on the Type Zone

The registry holds 25,486 disease-grade records from 5,885 registrations. Of these registrations, 321 report profiles for two or more type zones, and further cultivars are documented in several zones through separate registrations. 1,772 pairs of a cultivar and a disease, belonging to 567 cultivars, have grades in at least two zones. In 678 of these pairs (38.3%) the grades differ, and in 360 pairs (20.3%) one zone grades the cultivar MR or better and another grades it MS or worse (Figure 3(a)). Rice and maize have similar shares of differing pairs, 287 of 756 in rice and 391 of 1,016 in maize. In maize, 261 pairs cross the resistance boundary and 269 differ by two or more grade steps. In rice, 99 pairs cross it and 58 differ by two or more steps. Pairs documented across separate

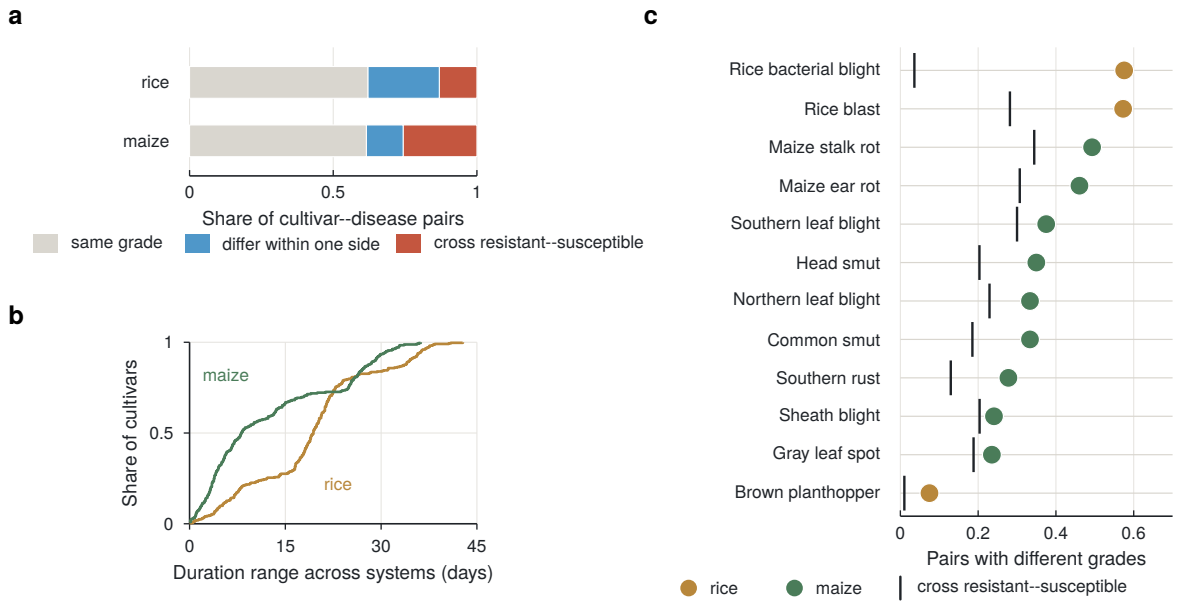


Figure 3. Official grades and durations differ between type zones. **a**, Cultivar–disease pairs documented in two or more zones, divided by whether the grades are identical, differ within the resistant or the susceptible side, or cross between MR or better and MS or worse. **b**, Cumulative share of cultivars by the range between their longest and shortest documented duration across zones. **c**, Share of pairs with different grades for each disease with at least 25 pairs. Black ticks give the share that cross the resistant–susceptible boundary. Exact counts are in Table A1.

registrations differ somewhat more often (327 of 803, 40.7%) than pairs documented within a single registration (351 of 969, 36.2%), but the differences are already common within one registration, so merging separate approval documents does not produce them.

The share of differing pairs varies by disease (Figure 3(c)). Rice blast and bacterial blight differ in 57% of their pairs, maize stalk rot in 49% and ear rot in 46%, and gray leaf spot, sheath blight and southern rust in 24% to 28%. Brown planthopper grades differ in 7.5% of their pairs. Bacterial blight grades seldom cross the resistance boundary (5 of 80 differing pairs), whereas 91 of the 185 differing blast pairs do.

Documented durations differ as well. A cultivar with two or more zones under the same endpoint has a median duration range of 19.4 days in rice (323 cultivars) and 8.1 days in maize (245 cultivars), and the 90th percentiles are 34.5 and 29.0 days (Figure 3(b)). Rice zones differ in cropping season as well as in location, which may explain much of the longer rice ranges. A pooled minimum duration therefore need not describe the intended zone, as the Dunliangyou Sizhan example shows.

6.2 Pooling Changes the Shortlist, and the Approved Pool Limits Supply

Pooling values from different zones can make a cultivar appear to meet a maturity and disease requirement that its intended zone does not support. For blast resistance of MR or better, the earliest option documented for the intended zone is 146.1 days in Upper Yangtze single-season rice, and the earliest pooled option is 110.2 days, which is 35.9 days earlier. The gaps are 6.8 days in Middle and Lower Yangtze single-season rice and 3.5 days in double-season late rice, and the maize panels that require MR or better for every regional disease show none. Under the Not HS criterion the maize gaps are 24.2 days in Eastern and Northern maize and 8.7 days in Southwest maize. Table A2 lists every zone, panel and threshold.

The approved pool supplies few options for some requests. In Upper Yangtze single-season rice, 7 of 115 candidates are documented as MR or better against blast, in Middle and Lower Yangtze single-season rice 76 of 344, and in double-season late rice 22 of 131. In maize, 2 of 175 candidates in Eastern and Northern middle-late spring maize, 3 of 95 in Southwest maize and none of 420 in Huang-Huai-Hai summer maize meet MR or better for all regional diseases. The panels differ between crops, blast alone for rice and all regional diseases for maize, so these counts do not compare rice with maize. No

Middle and Lower Yangtze single-season candidate meets moderate resistance to both blast and bacterial blight. In Upper Yangtze rice, seven blast-resistant candidates lack local bacterial blight grades, so the joint request stays unresolved for them. These gaps show which assessments a breeder or an extension center lacks before it can shortlist.

On the test cohort, pooling produces 644 unsupported additions among the 157,522 candidate–request pairs, and a conservative rule that requires every profile to pass withholds 1,212 supported ones (Table A3 gives the unsupported additions by type zone).

6.3 Zone Dependence Is Concentrated in Multi-Profile Cultivars

The two failures above concern a small part of the workload. Coarse bounds agree for 155,666 of the 157,522 requests (98.8%), and every agreed answer matches the documentary answer. The remaining 1,856 requests concern 23 profiles in 16 registrations, and profile bounds reduce them to 1,813 requests in the same profiles. Middle and Lower Yangtze rice holds 1,107 of the coarse-flagged requests and Upper Yangtze rice 600, together 92%, and Huang-Huai-Hai summer maize holds none. Most of the 157,522 requests concern registrations with a single profile, which are zone invariant by definition. Among the 8,134 requests on registrations with several profiles, the 1,856 coarse-flagged requests are 22.8%, and they concentrate in 23 profiles. Knowing the intended type zone removes the 644 unsupported additions of the optimistic rule and the 1,212 withheld options of the conservative rule.

6.4 The Determinacy Test Against Published Retrieval Policies

On the sample of 1,000 requests, the reader of anonymous profiles is wrong on 0.563% of the weighted requests, and on 47.75% of the sampled requests whose bounds disagree. The reader of the complete source passage makes no error, and it reads the source for every request at 685 tokens per request. The determinacy test makes no error either. It reads the source for 1.18% of the requests with coarse bounds and for 1.15% with profile agreement, and it uses about 11 tokens per request. The Sufficient Context autorater makes no error, reads the source for 6.55% of the requests and uses about 3,110 tokens per request. The A2RAG-style controller errs on 0.003% of the requests, reads the source for 1.22% and uses about 1,880 tokens. The MBA-RAG adaptation errs on 0.13% to 0.56% of the requests, depending on its cost weight. The sufficiency-and-confidence router errs on 0.23% of the requests at its lowest read target and on none at targets of 30% and above (Table A4). Among the policies that make no error on this sample, direct condition execution with the determinacy test uses the fewest generation tokens and reads the source less often than the autorater.

The paired comparison shows where the saving comes from. When model readers answer every request, choosing the requests that read the source by profile agreement instead of the multi-profile rule changes the tokens per request by 0.17%. When the shortlist rule answers the requests whose bounds agree, profile agreement reduces the tokens per request by 77.7% against the multi-profile rule (Table A5). The saving therefore comes mainly from answering with the shortlist rule instead of a model call. The published policies are designed to decide when to read the source, and each of them leaves the answer to a model reader.

6.5 Which Associations a Request Needs

Figure 4(a) removes the link between a disease grade and the assessment that produced it, on 1,155 transgenic-maize cases built from 385 grade matrices. Five requests are posed over the same grades. The request that every assessment qualifies and the request that each disease has an acceptable grade somewhere stay fully determined. The request for one jointly qualifying assessment leaves 155 cases unresolved. A request that names the inoculation or the field assessment leaves 328 cases unresolved, including cases in which a qualifying assessment certainly exists but cannot be identified.

Taking the best grade of each disease is a common summary. Under a cutoff shared by all diseases it reproduces every existential answer in 352 of the 385 matrices, but 41 of these fail once the cutoff differs between diseases (Appendix A.6). A request for a joint assessment therefore needs the complete groupings, a request for a named assessment also needs the protocol labels, and a request about single diseases needs only per-disease summaries. A multimedia knowledge tree can present the compact summary while the graph keeps the groupings.

The ablation on the test cohort removes the corresponding links between a value and its type zone. Figure 4(b) drops one link at a time. When the named zone keeps neither the duration nor

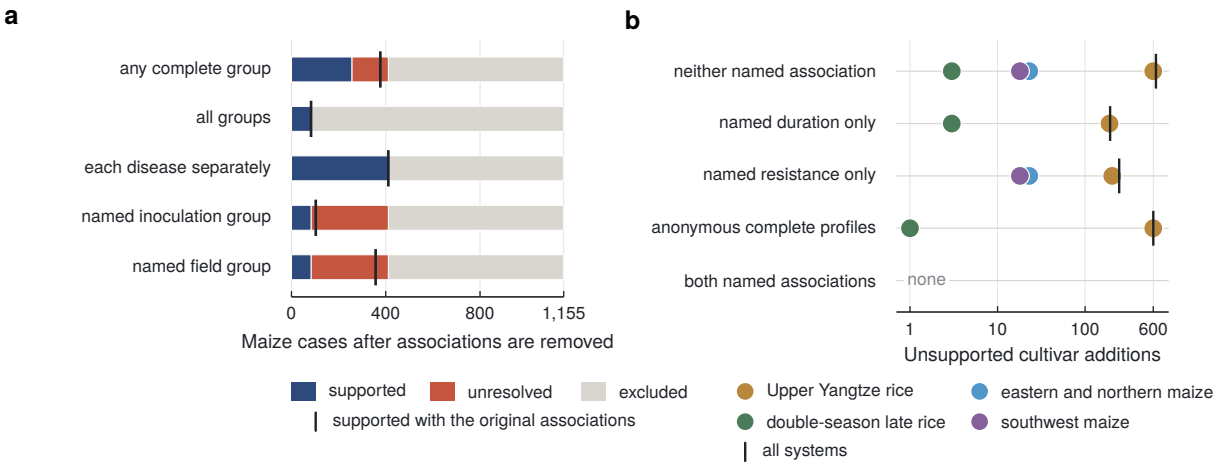


Figure 4. Which associations a cultivar request needs. **a**, The 1,155 transgenic-maize cases under five requests posed over the same disease grades. Bars divide the cases after grade-to-assessment associations are removed into supported, unresolved and excluded, and the black tick marks the number supported with the original associations. **b**, Unsupported cultivar additions among the 157,522 test-cohort requests when the named type zone keeps neither, one or both of its maturity and resistance associations. Each dot gives the additions for one type zone on a logarithmic axis, and the black tick the total over all zones. Anonymous complete profiles keep the grouping without the type zone. Both named associations reproduce every documentary answer. Exact counts appear in Table A3.

the grades, 644 candidate–request pairs receive unsupported positive answers. Keeping the duration of the intended zone leaves 193 of these errors, and keeping its grades leaves 245. Each association prevents errors that the other leaves, and keeping both reproduces every documentary answer. Complete profiles without zone labels still give 601 unsupported positive answers, because a profile that qualifies can belong to another zone.

Upper Yangtze rice accounts for 600 of the 644 additions, and either association alone removes about two thirds of them. The 41 maize additions in the Eastern and Northern and the Southwest zones disappear when the duration association returns, and the three additions in double-season late rice disappear when the grade association returns. Two zones have no additions in any variant, and no variant withholds a supported answer (Table A3).

6.6 A Withheld Grade Is Hard to Predict from Other Zones

A database that lacks the type zone might predict the missing grade instead of retrieving it. A pooled summary already assumes that a cultivar's grade in one zone informs its grade in another. Figure 5 tests that assumption on 3,751 instances of 567 cultivars. The most common grade alone gives 37.7% exact accuracy. Imputing the median, best or worst grade of the other zones gives 62.4%, 63.0% and 62.9%. The base rate of the zone and disease, which ignores the other ratings of the cultivar, gives 66.4%. Logistic regression, the random forest and gradient boosting give 69.1%, 69.5% and 67.6%.

The random forest has the highest exact accuracy among the learners, but its lead over logistic regression (0.4 points) equals its standard deviation over seeds, and logistic regression is slightly higher on the resistant or susceptible side (84.3% against 84.1%). The base rate of the zone and disease is the strongest simple baseline. Against it the random forest is more accurate for 214 cultivars and less accurate for 183 (exact sign test, $p = 0.13$), which leaves the comparison open. Against the pooled median rule it is more accurate for 311 cultivars and less accurate for 133 ($p = 1.6 \times 10^{-17}$). On the resistant or susceptible side, the random forest reaches 84.1%, the base rate 81.6% and the median rule 80.4%. The other-zone grades and metadata yield a small, statistically inconclusive improvement over the zone–disease base rate. Rules that carry those grades into another zone are less accurate than the base rate of that zone.

6.7 Simulated Annotation of Tomato Disease Images

The approval analyses test the pipeline where its record is a document. The remaining subsections test the stages that handle images, derived descriptions and dated field measurements. In the tomato experiment, the visual classifier reaches 0.926 accuracy and 0.914 ten-class macro F1 on the PlantVillage test images and 0.244 and 0.118 on

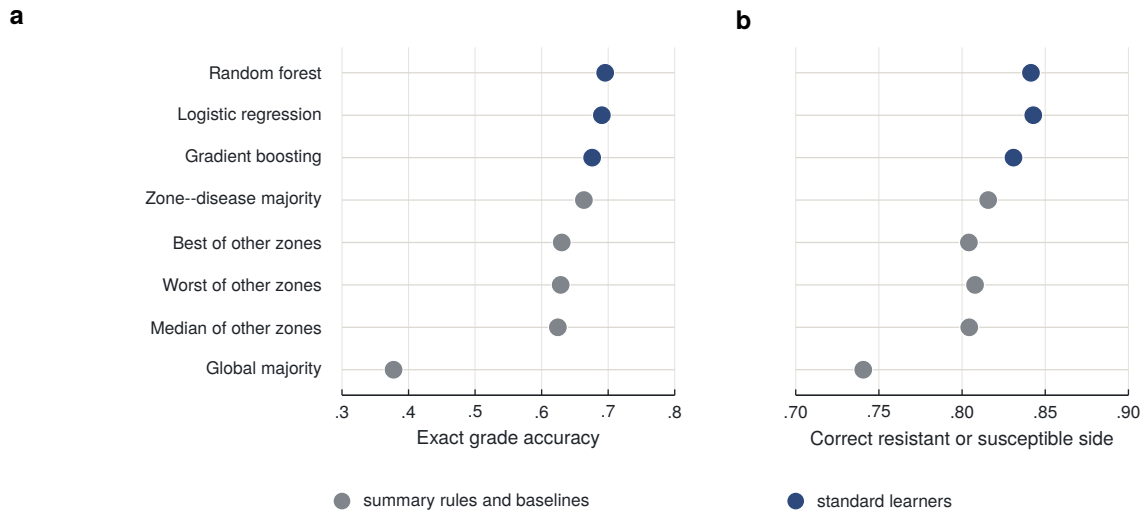


Figure 5. Predicting a withheld grade from a cultivar's other zones. **a**, Exact agreement with the official grade. **b**, Agreement on the resistant (MR or better) or susceptible side. Gray dots are baselines and summary rules, and blue dots are standard learners. Values are means over five repeats of grouped cross-validation. Standard deviations and the remaining metrics are in Table A6.

the PlantDoc field photographs (Appendix B). Figure 6(a) shows how that shift changes which fusion rule works best. With independent simulated contributions, the fixed product and the calibrated logarithmic pool both outperform majority voting on source images. On external photographs the ranking reverses, and majority voting exceeds the calibrated logarithmic pool by 0.044 macro F1 in mean paired differences over ten seeds (exact sign-flip $p = 0.0039$). Calibrating the logarithmic weights on external images gains 0.392 over the fixed product ($p = 0.0020$) and still loses to majority voting.

Worker reliance on a model suggestion lowers consensus quality, and storing each annotator's exposure is what makes the loss visible. Copying model suggestions decreases external macro F1 by 0.246 for majority voting and 0.200 for the calibrated logarithmic pool (both $p = 0.0020$). The paired comparison between the two fusion methods is inconclusive under copying ($p = 0.86$). Complete fusion results appear in Table A7, clustering results in Appendix B and exact paired tests in Table A8.

Clustering responds to the two probability blocks in opposite ways (Figure 6(b)). Adding the image model's probabilities raises the adjusted Rand index on source images in every seed and gives no benefit on external photographs. Adding the simulated worker probabilities raises it on external photographs in every seed. In this simulation, calibrated weights did not compensate for unrecorded model exposure [45, 46].

6.8 A Saved Description and a Saved Table

AgMMU and HiTab ask whether a derived summary can replace the source it was made from. On the 768 AgMMU cases, Qwen3.8-27B answers 673 correctly from the raw images, 658 from a description written before the model saw the question, 664 from both and 580 from the released background text. Qwen3.5-4B answers 629, 603, 603 and 528 (Appendix B.4). The description carries most of what the images carry, and adding it to the images corrects some answers and breaks others. The reasoning configuration and the crossed design, in which one model writes the description and another answers, show the same two-way shifts.

HiTab separates agreement with the source table from answer accuracy. Source checking raises exact agreement on header paths and body values from 4,620 to 5,149 of 5,779 fields, and it corrects six errors while introducing two new errors under greedy answering, and corrects six while introducing nine under reasoning. Reasoning answers fall from 161 of the 205 questions on the source hierarchy to 151 on model extraction and 148 on source-checked extraction, and greedy answering falls from 136 to 107 and 111. A table that answers its first question correctly often fails a later one, in every representation including the source hierarchy. Over all 65 tables, model extraction answers every question

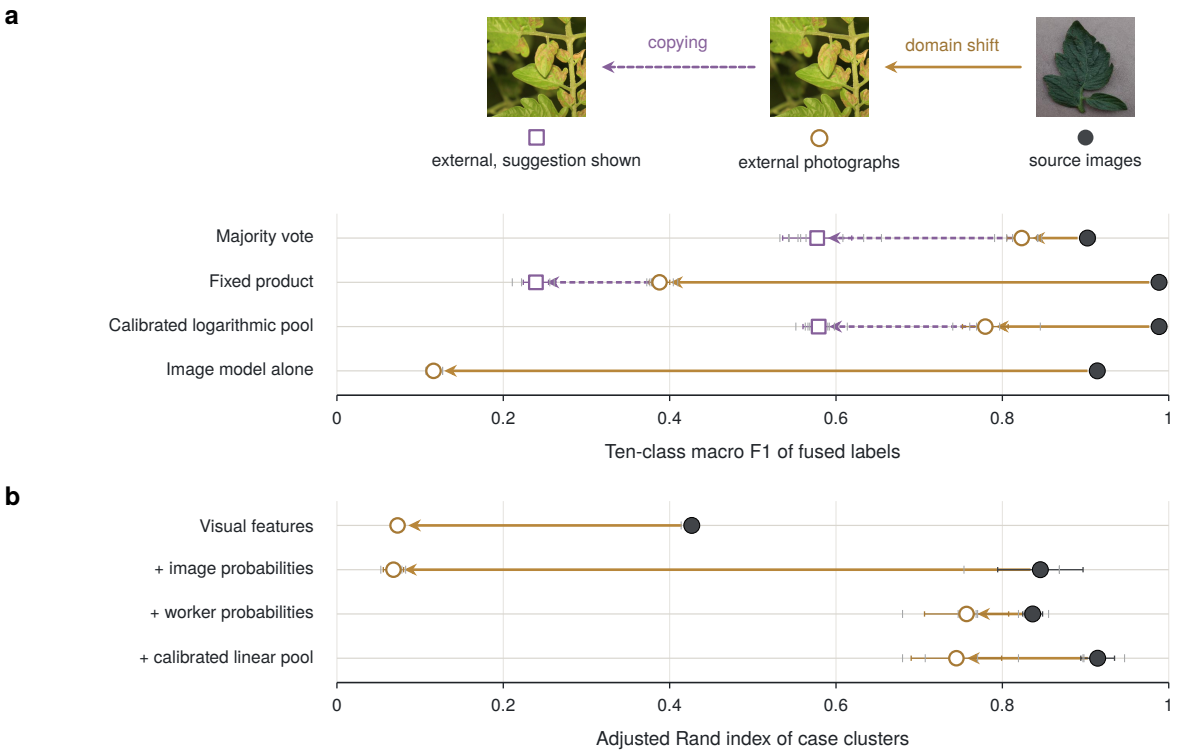


Figure 6. Tomato annotation under domain shift and copying. **a**, Macro F1 of fused disease labels for source test images and external photographs with independent simulated workers, and for external photographs whose workers see the model's suggestion. Solid arrows show the loss from domain shift and dashed arrows the further loss from copying. **b**, Adjusted Rand index of tomato case clusters built from cumulative representations, with independent workers. Markers give means, whiskers the sample standard deviation and gray ticks individual seeds. The key above the axis pairs each marker with an example image of its condition. External photographs cover nine classes, and F1 uses the fixed ten-class average. The external example is from Wikimedia Commons (Wolan268, CC0), and the source example is from PlantVillage [1], CC BY-SA 3.0.

of 33 correctly, the exact expansion 35 and the source hierarchy 39 (Appendix B.4).

6.9 Field Allocation in Maize and Potato

The two field datasets rank candidates from records that already exist, so the candidate pool and the decision date are fixed before any model runs. In the maize reanalysis, parental links raise the share of hybrid–trial candidates with a usable own or parental history from 35% to 86%, and the both-parent rule selects 9.766 Mg ha^{-1} against 9.462 for uniform allocation over 81 trials. In the later season, fitted parental effects select 10.00 Mg ha^{-1} and unpenalized parental means 10.03, so a simple observed baseline is competitive once the pool changes. Given the same 955 marker-covered cross–trial pairs, every genomic model selects a higher yield than parental effects, and genomic ridge regression is highest at 10.84 ± 2.45 Mg ha^{-1}, positive in 12 of the 14 locations (Table A11).

In potato scouting, the value of an aerial measurement depends on the hold-out design. At a quarter of the budget under variety transfer, combined flight history and the latest aerial survey both capture 26.9% of assessed early-blight severity, and their paired difference across varieties ranges from about -16 to $+3$ percentage points. The pooled parity hides that spread. The latest physiological measurements capture 35.3% under the same design, more than any aerial family. Under block exclusion, where the varieties are familiar, field and variety alone capture 75.6% against 76.7% for combined history (Table A12). Both studies allocate a fixed budget over a pool that the record defines, which is the situation the daily-management design of Section 7 treats on a one-day horizon.

7 Extension to Daily Crop Management

Cultivar selection is one decision of the season, and the same records support the daily decisions that follow. Estimating daily precision water and nutrient demand over the full crop cycle needs dated weather drivers, crop development stages, soil supply and verified management interventions on

one time axis. A recommendation for the next day must distinguish completed irrigation from a planned schedule, and accumulated daily precipitation from an instantaneous soil-moisture reading.

7.1 Coupling Observations Across Space and Time

The coupling of heterogeneous agricultural data across spatial and temporal scales brings sensor streams, drone imagery and laboratory assays onto the day and the production unit of the crop model. Multi-environment trial analyses model cultivar performance as a function of the environment of each trial, so environmental drivers have to match the plot where the response was measured [8–10]. Every measurement carries its acquisition time, spatial footprint, sampling depth and unit. Precipitation is accumulated over the modeled day, vertical soil-sensor readings are combined by an explicit depth-weighting rule, and aerial imagery is mapped to the plot boundary and the interval it represents. Interpolation models record their assumptions, and the raw observations stay in the graph.

Temporal alignment fixes the information available when an action was chosen. A later disease assessment can evaluate an earlier scouting decision and cannot enter the features of its forecast. The potato reanalysis follows this rule and builds forecasts only from observations made before the outcome assessment. Spatial alignment applies the same rule across space, because a sensor in one greenhouse compartment does not measure an adjacent compartment with independent controls. With these boundaries, predictive models use only evidence that existed at the decision time.

7.2 Crop-State Estimation and Daily Balances

Data- and knowledge-driven interpretable inversion of crop conditions estimates latent states, such as root-zone water availability and canopy stress, from physical observations and process relationships. Let $\boldsymbol{x}_t$ denote the state of one production unit on day t, $\boldsymbol{z}_t$ its measured observations and $\boldsymbol{a}_t$ its management inputs. A state-space formulation is

$$\boldsymbol{x}_{t+1} = \mathcal{F}_\theta(\boldsymbol{x}_t, \boldsymbol{a}_t, \boldsymbol{w}_t) + \boldsymbol{\eta}_t, \qquad \boldsymbol{z}_t = H_\theta(\boldsymbol{x}_t) + \boldsymbol{\epsilon}_t. \quad (7)$$

The weather drivers are $\boldsymbol{w}_t$, the model parameters are θ, and $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$ represent process and observation discrepancies. The transition function $\mathcal{F}_\theta$ describes biophysical crop and soil dynamics, while the observation operator H_θ maps latent states to observable quantities. Data assimilation updates the state trajectory with both sources of information [101], and it has been used to follow crop water stress through a season [65]. Interpretable inversion exposes the specific observations, mass balances and parameter assumptions that drive each state update.

For a fixed modeled soil volume and a one-day interval, a water balance is

$$W_{t+1} = W_t + P_t + I_t + C_t - ET_{c,t} - Q_t - D_t. \quad (8)$$

The stored water is W_t. The remaining terms are precipitation reaching the soil P_t, irrigation reaching the soil surface I_t, capillary input C_t, crop evapotranspiration $ET_{c,t}$, surface runoff Q_t and drainage D_t. All are expressed as water depth over the same area, with fluxes accumulated over the day. Runoff is subtracted once from the surface inputs. Conveyance losses upstream of the soil surface belong to the delivery calculation. Crop development, atmospheric demand and stress determine evapotranspiration in established procedures [57]. Changing rooting depth requires an explicit mapping between the soil layers and the crop-accessible store.

The cultivation system determines the stores and their boundaries. Flooded rice requires a surface-water store and its exchanges with the soil. Maize and soil-grown tomato require the relevant root-zone layers. A substrate-grown tomato system instead uses its substrate and drainage compartments. The knowledge record identifies the selected model and its supported cultivation system before observations enter the balance.

For nutrient h, an available-pool balance on the same area is

$$N_{h,t+1} = N_{h,t} + F_{h,t} + M_{h,t} - U_{h,t} - L_{h,t} - B_{h,t}. \quad (9)$$

Here $N_{h,t}$ is available nutrient mass and $F_{h,t}$ is external supply entering that pool. Release from other soil pools is $M_{h,t}$, crop uptake is $U_{h,t}$, losses leaving the modeled soil volume are $L_{h,t}$, and transfer into unavailable pools is $B_{h,t}$. The selected nutrient model resolves processes such as mineralization, immobilization, sorption and release within these terms. All nutrient fluxes share a common mass-per-area unit and elemental chemical basis. Uptake demand, realized uptake and applied fertilizer mass are three different

quantities. Soil nutrient supply and the rooting architecture limit how much of the demand the crop meets, and fertilizer formulation and soil chemistry determine how much of an applied dose reaches the plant-available pool.

A full-cycle calculation repeats these balances through crop development. Crop models supply stage-dependent growth and uptake relationships [58, 102]. A balance accounts for what enters and leaves the modeled volume. It bounds the store without fixing demand, which follows from the crop's growth stage and physiological condition, so the chosen model has to represent the nutrient dynamics that the request needs. WOFOST 8.1 provides dynamic nitrogen limitation, while its documented QUEFTS coupling supplies a different, non-dynamic treatment of nutrient limitation [58]. Source prescriptions constrain application events, and measured soil and crop responses update states or parameters. With the balances closed, the daily inversion of Equation 7 cannot move water or nutrient demand away from the root-zone processes that the crop model specifies [57, 58].

7.3 Choosing an Application

A producer needs an application decision after crop demand has been estimated. Let $\widehat{W}^0_{t+1}$ and $\widehat{N}^0_{h,t+1}$ be next-day stores predicted from Equations 8 and 9 with the candidate day's irrigation and external nutrient input set to zero. Let $W^\dagger_{t+1}$ and $N^\dagger_{h,t+1}$ be agronomically specified lower target stores. For fixed forecast fluxes, the minimum additional supplies needed to meet those targets are

$$\begin{aligned} I^{\text{need}}_t &= [W^\dagger_{t+1} - \widehat{W}^0_{t+1}]_+, \\ F^{\text{need}}_{h,t} &= [N^\dagger_{h,t+1} - \widehat{N}^0_{h,t+1}]_+, \end{aligned} \tag{10}$$

where $[x]_+ = \max(0, x)$. These are water depth reaching the soil and nutrient mass entering the available pool. Converting them into withdrawn water or fertilizer product uses the delivery pathway, product composition and estimated availability.

The targets depend on crop stage and the modeled soil layers. Upper storage limits, application windows and equipment capacity further restrict a feasible dose. Soil-water scheduling uses the crop's allowable depletion and available storage [57]. When uptake, runoff or drainage changes with the proposed dose, Equation 10 becomes implicit in the dose, and the calculation evaluates the coupled balance for each candidate dose. Repeated state estimation connects daily applications over the crop cycle.

The official cultivation guidance for the maize cultivar Huaxingdan 88DT specifies split nitrogen applications at sowing, the five-leaf stage and the late whorl stage [82]. Such transgenic cultivars are approved under a national transgenic maize standard [103]. These guidelines fix the seasonal timing of fertilizer but not the daily uptake trajectory. A phenological model maps crop development stages to calendar dates, while physical field observations constrain dynamic crop and soil states (Appendix D).

Consider a tomato case with a moisture reading collected before irrigation and a foliar wilting photograph collected afterward. The irrigation event separates the physiological states that produced those observations (Figure 7). The calculation first updates the root-zone water balance with the performed irrigation, then evaluates whether further application remains indicated. A post-irrigation reading constrains the updated state. An expert correction to the recorded irrigation volume recomputes the water balance and its dependent recommendation. Similarly, a correction to the photograph's symptom label revises the diagnostic evidence. These updates show where expert and producer feedback enters the same management episode. We formulate daily application doses as bounded inverse optimization problems conditioned on stage-specific guidelines and soil-water targets, which keeps proposed interventions within those limits [82, 102].

7.4 A Numerical Day and Calibration

Consider a synthetic one-day scenario for soil-grown tomato in a greenhouse compartment, where all meteorological inputs, calibration pairs and target stores are illustrative. Two 0.20-m root-zone layers with volumetric water contents of 0.24 and 0.22 give an initial soil-water store of $W_t = 1{,}000(0.20 \times 0.24 + 0.20 \times 0.22) = 92.0$ mm. The roof excludes rainfall, the compartment has no lateral inflow, and groundwater is assumed too deep for capillary rise, giving $P_t = C_t = 0$.

At an external reference weather station, suppose daily maximum and minimum temperatures are 25 and 15 °C, maximum and minimum relative humidities are 80% and 45%, wind speed at 2 m is 2.0 m s^{-1}, and pressure is 101.3 kPa. A net radiometer gives a daily integral of 10.2 MJ

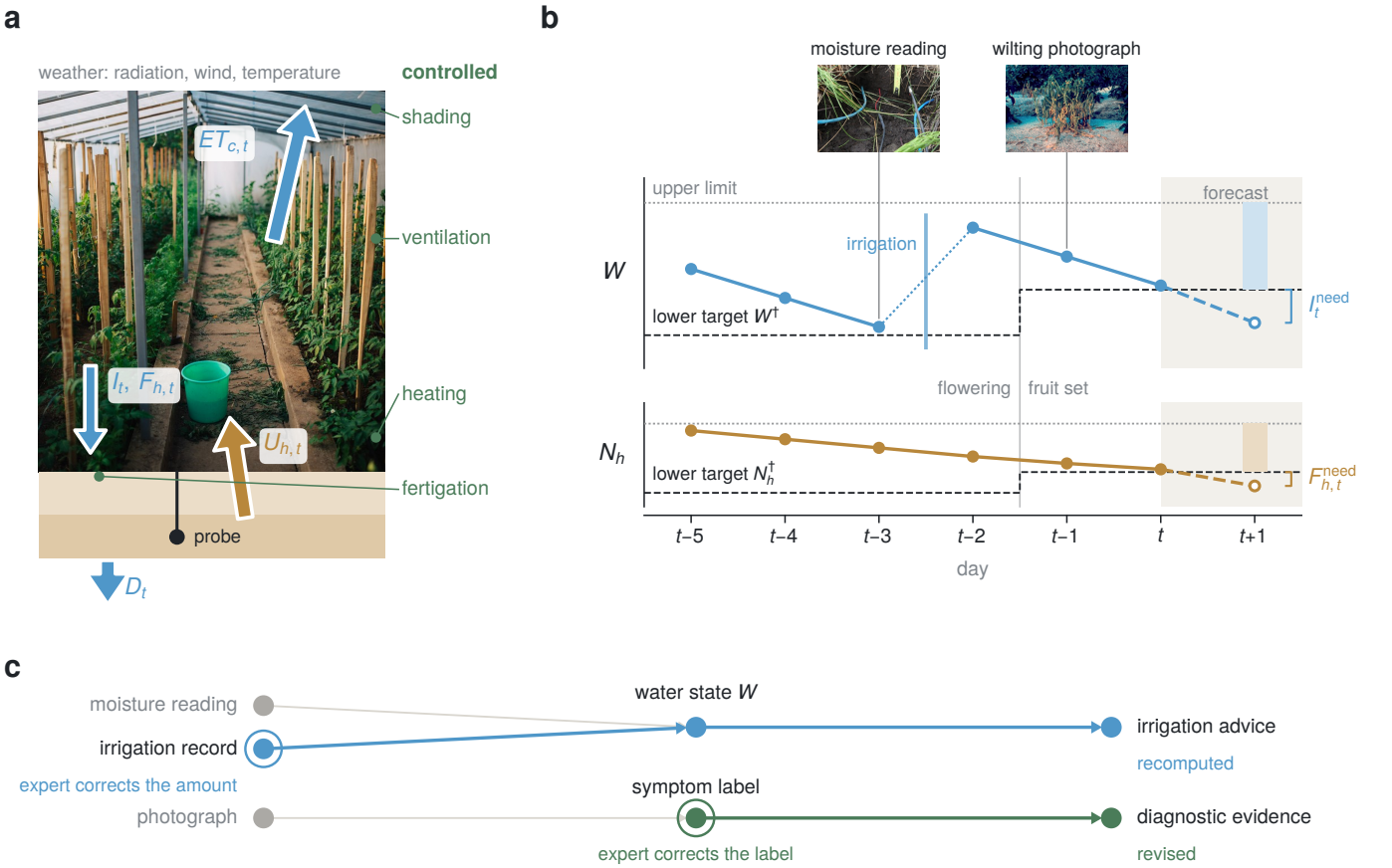


Figure 7. Context in a proposed daily tomato-management calculation. **a**, One greenhouse compartment with the fluxes of Equations 8 and 9 and the controlled inputs that act on them. **b**, Root-zone water and nutrient stores by day. The gap between the next-day forecast without application (open marker) and the lower target gives the minimum additional supply of Equation 10, and the shaded bar marks doses up to the upper limit. A recorded irrigation separates the moisture reading from the later wilting photograph. Store levels are schematic. **c**, A corrected irrigation amount recomputes the water state and irrigation advice, whereas a corrected symptom label revises only the diagnostic evidence. Photographs are public examples from Wikimedia Commons by Shixart1985 (CC BY 2.0), UBC Micrometeorology (CC BY 2.0) and Scot Nelson (CC0).

m^{-2}, with daily soil heat flux approximated as zero. The FAO Penman–Monteith calculation uses mean temperature 20 °C, mean saturation vapor pressure 2.43656 kPa, actual vapor pressure 1.39489 kPa, saturation-curve slope 0.14474 kPa $^{\circ}\mathrm{C}^{-1}$ and psychrometric constant 0.0673645 kPa $^{\circ}\mathrm{C}^{-1}$ [57]. Actual vapor pressure is calculated by pairing saturation pressure at the minimum temperature with maximum relative humidity and vice versa. These inputs give reference evapotranspiration of 4.00694 mm, rounded to 4.0 mm for the daily balance. To account for the greenhouse canopy and cover, local calibration pairs reference evapotranspiration with lysimeter evapotranspiration calculated as irrigation minus drainage minus the increase in stored water. Paired values of $(2.0, 1.8)$, $(3.0, 2.7)$ and $(4.0, 3.6)$ mm give a least-squares coefficient through the origin of $(2 \times 1.8 + 3 \times 2.7 + 4 \times 3.6)/(2^2 + 3^2 + 4^2) = 0.90$, yielding $ET_{c,t} = 3.6$ mm.

A simple daily bucket sets drainage to the excess above a 104-mm field-capacity store and runoff to zero while the applied drip rate remains below infiltration capacity. The predicted stores, both without irrigation and with the selected dose, remain below field capacity, so $D_t = Q_t = 0$. Equation 8 then gives $\widehat{W}_{t+1}^{0} = 92.0 - 3.6 = 88.4$ mm. A specified lower target of 94.0 mm requires $I_t^{\mathrm{need}} = 5.6$ mm by Equation 10. With delivery efficiency 0.90, withdrawal is $5.6/0.90 = 6.22$ mm, or 62.2 m^3 ha^{-1}. An integrated delivery-meter volume divided by the irrigated area checks the applied depth.

For nitrogen, let a mineral-N assay give 4.17 mg N kg^{-1} of dry soil over the same 0.40-m profile with bulk density 1,200 kg m^{-3}. Its area conversion gives 20.016 kg N ha^{-1}, rounded to 20.0 kg N ha^{-1} for the following calculation. A modeled daily dry-matter increment of 120 kg ha^{-1}, with an assumed marginal

N concentration of 0.020 kg N per kg dry matter, gives an uptake demand of 2.4 kg N ha^{-1}. The available pool meets this demand. A first-order mineralization calculation releases $0.002 \times 300 = 0.6$ kg N ha^{-1} from a specified mineralizable pool. An immobilization calculation transfers $0.0005 \times 200 = 0.1$ kg N ha^{-1} into an unavailable pool. A prescribed daily gaseous-loss fraction of 0.01 gives $L_{N,t} = 0.2$ kg N ha^{-1}, and drainage loss is zero for this day.

Equation 9 yields $\widehat{N}^{0}_{N,t+1} = 20.0+0.6-2.4-0.2-0.1 = 17.9$ kg N ha^{-1}. A lower target of 20.0 kg N ha^{-1} requires 2.1 kg N ha^{-1} entering the available pool. For a product containing 15.5% N with a specified 0.90 fraction reaching that pool, the product dose is $2.1/(0.155 \times 0.90) = 15.05$ kg ha^{-1}. Re-evaluating the water and N balances with the selected supplies gives stores of 94.0 mm and 20.0 kg N ha^{-1}.

In field deployment, calibration pairs moisture-probe readings with gravimetric samples at each depth and checks delivery meters against collected volumes. Monitored wetting episodes determine field capacity and drainage, while lysimeter records calibrate stage-specific evapotranspiration coefficients with later episodes reserved for validation. Destructive biomass and tissue-N samples constrain crop growth and uptake, whereas soil mineral-N series, incubation assays and drainage concentrations constrain mineralization, immobilization and losses, because soil moisture and final mineral N alone do not identify every individual flux. Daily storage and flux estimates are then validated across separate compartments and crop stages alongside water and nitrogen balance residuals.

7.5 Control and Feedback in the Production Unit

Precision control of water, light, soil, air and fertilizer during crop growth means choosing the management inputs from the estimated state and from explicit production objectives. Irrigation and fertigation act directly on the stores of Equations 8 and 9. In protected tomato production, automated shading and supplementary lighting regulate the radiation reaching the canopy, while ventilation adjusts the humidity and carbon dioxide concentration of the greenhouse air. Greenhouse experiments have tested remote AI control of climate, irrigation and crop production in cucumber and cherry tomato [61, 62], and multiple-sensor feedback algorithms in autonomous dwarf-tomato cultivation [63]. Reinforcement learning has been compared with model predictive control for greenhouse climate [64].

The crop's aerial microclimate brings light, wind, temperature, soil, air and fertilizer into one control problem. Wind speed and air temperature determine reference evapotranspiration through the FAO Penman–Monteith equation [57]. In open maize and rice fields, weather variables act as forecast boundary conditions for Equation 8, guiding the scheduling of irrigation and fertilizer events around incoming weather fronts. In greenhouse tomato production, motorized roof vents, circulation fans and hot-water heating make air movement and temperature partly controllable. A single actuator often changes several states at once. Opening vents, for example, lowers temperature and relative humidity and also vents enriched carbon dioxide. A greenhouse climate controller evaluates candidate multi-actuator schedules through the coupled crop and microclimate model, subject to stage-specific stress thresholds and equipment capacity limits, and weighs resource use against projected yield. Uncontrolled weather drivers enter the case as observed or forecast assertions, while actuator adjustments enter as explicit actions indexed by timestamp and production unit.

Every issued recommendation records its target production unit, current state estimate, parameter assumptions, selected action and expected crop response. Because ventilation alters the microclimate driving transpiration, a ventilation command directly shifts the subsequent irrigation schedule. After execution, observed microclimate dynamics and crop growth update the case history. A favorable crop response provides empirical evidence about the recorded episode, while extending the resulting control rule to other greenhouse compartments or cropping seasons requires validation under those specific conditions.

8 Discussion

8.1 Grade Differences Between Type Zones

A cultivar in the approval record has one profile for each zone in which it was tested, and the profiles often disagree. More than a third of the cultivar–disease pairs documented in several zones carry different grades, and one pair in five is graded resistant in one zone and susceptible in another. A summary that keeps the best grades across zones can describe a combination that no single tested zone supports. The Dunliangyou Sizhan example shows the effect on

one shortlist. The 2025 approvals show that the same effect recurs in hundreds of candidate–request pairs, mostly in Upper Yangtze rice.

The record mixes pathogen populations, trial years, nurseries and inoculation protocols with the type zone, and blast and bacterial blight grades differ between zones more often than brown planthopper grades. Separating these factors needs paired assessments of the same cultivars in the same year under matched nursery protocols across zones. Whatever the cause, an extension officer needs the grade of the intended zone.

8.2 What a Cultivar Database Must Keep

The shortlist needs two links to the intended zone, one for the duration and one for the disease grades. When only one link is kept, the other still lets unsupported cultivars in. Complete profiles without zone labels also admit unsupported cultivars, because a profile that qualifies can belong to another zone. The approved pool limits what a correct shortlist can contain. No Huang-Huai-Hai summer maize cultivar in the evaluated pool is documented as moderately resistant to all four regional diseases, and the count describes the documented evidence for that panel.

Many requests do not need the type zone, mostly because the registration has a single profile. Among the requests on registrations with several profiles, the coarse test flags 22.8% for checking, and they concentrate in 23 profiles of 16 registrations, few enough to check by hand against the announcement. For the rest, the database can store and serve a summary. A summary built for one kind of request can fail for another. Taking the best grade of each disease gave the same existential answers as the complete assessments under a shared cutoff in 352 of 385 matrices, but 41 of these gave different answers once the cutoff varied between diseases. A stored summary should therefore record which requests it can answer.

Against published retrieval policies, the determinacy test matched the most accurate ones. The Sufficient Context autorater, the complete-source reader and the test made no error in the sample, and among the policies without errors the test read the source least often, for 1.2% of requests, at 11 tokens per request against 685 for the complete-source reader and 3,110 for the autorater. The saving comes from answering with the shortlist rule on stored profiles, and routing alone saved almost nothing. The comparison therefore favors keeping the profiles in a structured form, which text-based policies do not need to do. The 400 sampled requests whose bounds disagree come from 15 registrations, so the sample is small for the cases that matter most.

Ratings from other zones carry a small amount of information about the intended zone. A random forest predicted a withheld grade slightly better than the usual grade of the zone and disease, but the improvement was inconclusive in the exact test. Rules that copy the best, worst or median grade from other zones were less accurate than the usual grade. Models of genotype-by-environment interaction use richer inputs such as pedigree, weather and soil. A database can leave a zone's grade unspecified until a supporting source is available.

8.3 The Same Dependence Appears in Other Evidence

The tomato experiment simulates how dependence between annotators and a model enters a consensus label. The move from laboratory to field photographs reversed the ranking of majority voting and calibrated pooling, and copying model suggestions lowered the macro F1 of both. The simulation supports recording annotator identity and model exposure with each label. Generated case descriptions lowered multimodal answering accuracy relative to the original images, which supports keeping the raw observation next to its description. In the field analyses, pedigree links increased the number of maize hybrids that could be ranked and genomic markers improved selection within a matched pool, and the potato results depended on the candidate pool and the sampling date. Each of these results applies to a stated candidate pool or decision horizon.

8.4 From Records to Daily Decisions

The daily-management design of Section 7 uses the same records. Estimated daily water and nutrient demands enter mass-conservation balances against soil moisture and nutrient pools. Stage-specific guidelines and equipment capacities turn the balances into irrigation and fertigation schedules, and the case record keeps the sensor streams, guidelines and simulation parameters behind each calculation. In the tomato scenario, a corrected irrigation volume updates the water balance, and a revised symptom label updates only the diagnostic evidence [35, 44]. The design can be tested at three levels. A record-level test checks that observations

and trial documents are linked to the right zone and conditions. A decision-level test compares recommendations with agronomic baselines over the same candidates and information horizon. An outcome-level test measures input efficiency and yield response after the intervention, in later seasons and on independent production units. Greenhouse challenges that compare remotely controlled compartments by crop yield and net profit are an example [62].

8.5 Scope

The approval analyses are retrospective. The registry was collected from a public variety-approval query service and read with extraction rules, and the documentary answers come from the same grammar; 2,680 of the 2,683 characteristics fields of the 2023 and 2024 announcements match the registry text. The cross-zone counts describe a complete population and carry no sampling error, and repeated cultivar families make pairs dependent, which is why we report the number of cultivars as well. The evaluated requests enumerate disease panels and maturity breakpoints, and the requests of growers may follow another distribution. The request-kind analysis uses published matrices.

In the multimedia experiments the annotators are simulated, and the five clustering seeds limit the smallest attainable paired p-value to 0.0625. AgMMU relies on its reference labels, and HiTab is a subset of published tables. The later-season maize evaluation covers one season at 14 locations, and the potato reanalysis covers eight cultivars in two fields, where management and field effects are confounded. Each stage was tested on its own dataset.

9 Conclusion

The pipeline keeps the type zone, the method and the source with each assertion, and it answers a cultivar request from the profile of the intended zone. On the national rice and maize approvals, pooling produced 644 unsupported candidate–request decisions and a conservative rule withheld 1,212 supported decisions. Among the requests on registrations with several profiles, the coarse test flagged 22.8% for further checking. Against retrieval-policy adaptations with common readers, the determinacy test made no error and read the source for 1.2% of the requests, at about 11 tokens per request against 1,880 to 3,110 for the published policies of comparable accuracy. Official grades differ between zones for 38% of the cultivar–disease pairs that have more than one zone, and standard learners predicted a withheld grade only slightly better than the usual grade of the zone and disease.

The other stages of the pipeline were tested on public benchmarks and simulated annotators. The daily-management design extends the records to crop demand, soil supply and application limits.

Data Availability Statement

The public datasets and official approval records used in this study are cited in the bibliography. Supporting data and analysis code are available on reasonable request. Appendix F describes these materials and the experimental settings.

Funding

This work was supported by the National Major Agricultural Science and Technology Project under Grant 0304.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors used generative AI tools to assist with reference formatting, language polishing, and translation. Specifically, GLM 5.3 and Kimi K3 were used for language polishing, while Kimi K3 was also used for LaTeX formatting. DeepSeek V4.1 Flash was used to assist with translation. Separately, the large language models Qwen3.8-27B and Qwen3.5-4B were evaluated as experimental subjects, as described in Section 5, Section 6, and Appendix F. The authors reviewed and verified all AI-assisted content and take full responsibility for the content of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in plant science, 7*, 215232. [CrossRef]

[2] Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global

burden of pathogens and pests on major food crops. *Nature ecology & evolution, 3*(3), 430-439. [CrossRef]

[3] Ministry of Agriculture and Rural Affairs of the People's Republic of China (2022). *Measures for the approval of major crop varieties*. Retrieved from https://www.moa.gov.cn/gk/nyncbgzk/gzk/202210/P020221012660321003677.pdf. Amended 21 January 2022. Articles 16-21. In Chinese. Accessed 21 June 2026.

[4] National Crop Variety Approval Committee (2021). *National maize cultivar approval standard* (2021 revision). Retrieved from https://zys.moa.gov.cn/flfg/202109/P020210930591894754607.docx. Issued 28 September 2021. Effective 1 October 2021. In Chinese. Accessed 21 June 2026.

[5] Ministry of Agriculture and Rural Affairs of the People's Republic of China (2025). *Announcement no. 985: 2025 national crop variety approvals*. Retrieved from https://www.agri.cn/zx/hxgg/202601/t20260123_8805730.htm. Official approval descriptions for 1,267 varieties, issued 31 December 2025. In Chinese. Accessed 21 June 2026.

[6] Ministry of Agriculture and Rural Affairs of the People's Republic of China (2023). *Announcement no. 726 and attached cultivar approval descriptions*. Retrieved from https://www.moa.gov.cn/govpublic/nybzzj1/202311/t20231122_6441029.htm. In Chinese. Jinxiang 5201, p. 299. Wannong Tiannuo 158, p. 639. Accessed 21 June 2026.

[7] Ministry of Agriculture and Rural Affairs of the People's Republic of China (2024). *Announcement no. 867 and attached cultivar approval descriptions*. Retrieved from https://www.moa.gov.cn/govpublic/nybzzj1/202412/t20241231_6468720.htm. In Chinese. Registrations 20242088 and 20243276. Accessed 21 June 2026.

[8] Smith, A. B., Cullis, B. R., & Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *The Journal of Agricultural Science, 143*(6), 449-462. [CrossRef]

[9] Jarquin, D., De Leon, N., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I., ... & Lorenz, A. (2021). Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. *Frontiers in genetics, 11*, 592769. [CrossRef]

[10] Barreto, C. A. V., das Graças Dias, K. O., de Sousa, I. C., Azevedo, C. F., Nascimento, A. C. C., Guimarães, L. J. M., ... & Nascimento, M. (2024). Genomic prediction in multi-environment trials in maize using statistical and machine learning methods. *Scientific Reports, 14*(1), 1062. [CrossRef]

[11] Lopez-Cruz, M., Aguate, F. M., Washburn, J. D., De Leon, N., Kaeppler, S. M., Lima, D. C., ... & de Los Campos, G. (2023). Leveraging data from the Genomes-to-Fields Initiative to investigate genotype-by-environment interactions in maize in North America. *Nature communications, 14*(1), 6904. [CrossRef]

[12] Zhang, Q., Li, B., Zhang, Y., & Wang, S. (2022). Suitability evaluation of crop variety via graph neural network. *Computational intelligence and neuroscience, 2022*(1), 5614974. [CrossRef]

[13] Han, Y., Liu, Z., Wang, A., Zhang, Y., Pan, S., Zhao, X., ... & Wang, K. (2026). Recommendation of suitable planting locations for hybrid maize varieties based on graph collaborative filtering. *in silico Plants, 8*(2), diag026. [CrossRef]

[14] Lv, B., Wu, H., Chen, W., Chen, C., Miao, Y., & Zhao, C. (2024). VEG-MMKG: Multimodal knowledge graph construction for vegetables based on pre-trained model extraction. *Computers and electronics in agriculture, 226*, 109398. [CrossRef]

[15] Wu, H., Xie, N., Wang, X., Fan, J., Li, Y., & Meng, Z. (2025). Crop GraphRAG: pest and disease knowledge base Q&A system for sustainable crop protection. *Frontiers in Plant Science, 16*, 1696872. [CrossRef]

[16] Gong, R., & Li, X. (2025). The application progress and research trends of knowledge graphs and large language models in agriculture. *Computers and electronics in agriculture, 235*, 110396. [CrossRef]

[17] Liu, L., Li, S., Qi, J., Yuan, Z., & Yang, P. (2026). TARAG: A time-aware retrieval-augmented generation framework for supporting precision crop pest and disease management through large language models. *Computers and Electronics in Agriculture, 248*, 111786. [CrossRef]

[18] Zou, J., Yuan, Y., & Chen, L. (2026). Clarification question generation for ambiguous agricultural knowledge graph question answering. *Computers and Electronics in Agriculture, 246*, 111617. [CrossRef]

[19] Arouna, A., Michler, J. D., Yergo, W. G., & Saito, K. (2021). One size fits all? Experimental evidence on the digital delivery of personalized extension advice in Nigeria. *American Journal of Agricultural Economics, 103*(2), 596-619. [CrossRef]

[20] Dongre, V., Gui, C., Garg, S., Nayyeri, H., Tur, G., Hakkani-Tür, D., & Adve, V. S. (2025). MIRAGE: A benchmark for multimodal information-seeking and reasoning in agricultural expert-guided conversations. *arXiv preprint arXiv:2506.20100*. [CrossRef]

[21] Imieliński, T., & Lipski Jr, W. (1984). Incomplete information in relational databases. *Journal of the ACM (JACM), 31*(4), 761-791. [CrossRef]

[22] Nash, A., Segoufin, L., & Vianu, V. (2010). Views and queries: Determinacy and rewriting. *ACM Transactions on Database Systems (TODS), 35*(3), 1-41. [CrossRef]

[23] Karlaš, B., Li, P., Wu, R., Gürel, N. M., Chu, X., Wu, W., & Zhang, C. (2020). Nearest neighbor classifiers over incomplete information: From certain answers to certain predictions. *Proceedings of the VLDB*

Endowment, *14*(3), 255-267. [CrossRef]

[24] Joren, H., Zhang, J., Ferng, C. S., Juan, D. C., Taly, A., & Rashtchian, C. (2025, May). Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations* (Vol. 2025, pp. 20310-20334). https://proceedings.iclr.cc/paper_files/paper/2025/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html

[25] Tang, X., Gao, Q., Li, J., Du, N., Li, Q., & Xie, S. (2025, January). Mba-rag: a bandit approach for adaptive retrieval-augmented generation through question complexity. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 3248-3254). https://aclanthology.org/2025.coling-main.218/

[26] Liu, J., Chen, Z., Qiao, S., Ju, M., Zhang, D., Han, B., ... & Yang, Z. (2026, May). A2rag: Adaptive agentic graph retrieval for cost-aware and reliable reasoning. In *2026 IEEE 42nd International Conference on Data Engineering Workshops (ICDEW)* (pp. 187-196). IEEE. [CrossRef]

[27] Ayadi, N. Y., Bernard, S., Bossy, R., Courtin, M., Happi, B. G. H., Larmande, P., ... & Faron, C. (2024). A unified approach to publish semantic annotations of agricultural documents as knowledge graphs. *Smart Agricultural Technology, 8*, 100484. [CrossRef]

[28] Chen, Y., Kuang, J., Cheng, D., Zheng, J., Gao, M., & Zhou, A. (2019, April). AgriKG: An agricultural knowledge graph and its applications. In *International conference on database systems for advanced applications* (pp. 533-537). Cham: Springer International Publishing. [CrossRef]

[29] Xiong, M., Wang, S., Che, Y., Ma, X., Xie, C., Jia, X., ... & Xi, L. (2026). Structured Knowledge Extraction from Wheat Variety Approval Texts via Relation-Aware In-Context Learning and Progressive Knowledge Calibration. *Information, 17*(10), 1041. [CrossRef]

[30] Wang, S., Che, Y., Jia, X., Li, Y., Xi, L., Ma, X., & Xiong, S. (2026). Joint entity-relation extraction from wheat variety certification texts for knowledge graph construction and variety relationship analysis. *Electronics, 15*(12), 2684. [CrossRef]

[31] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... & Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*. [CrossRef]

[32] Yang, S., Liu, Z., Mayer, W., Ding, N., Wang, Y., Huang, Y., ... & Feng, Z. (2024, November). Shizishangpt: An agricultural large language model integrating tools and resources. In *International Conference on Web Information Systems Engineering* (pp. 284-298). Singapore: Springer Nature Singapore. [CrossRef]

[33] Ravindran, D. J. S., Skarga-Bandurova, I., Sivakumar, V., Awais, M., & Mithra, S. (2026). AgroLLM: Connecting farmers and agricultural practices through large language models for enhanced knowledge transfer and practical application. *AgriEngineering, 8*(1), 38. [CrossRef]

[34] Mourtzinis, S., Silva, T. S., Lo, J. C. M., Smith, D. L., & Conley, S. P. (2026). A human in the loop approach to applying large language models for farm management insight. *Scientific Reports, 16*(1) 1273. [CrossRef]

[35] Abolhasani, M. S., Ba, Y., He, Y., & Pan, R. (2026). Beyond predefined schemas: TRACE-KG for context-enriched knowledge graphs from complex documents. *arXiv preprint arXiv: 2604.03496*. [CrossRef]

[36] Herschel, M., Diestelkämper, R., & Ben Lahmar, H. (2017). A survey on provenance: What for? What form? What from? *The VLDB Journal, 26*(6), 881-906. [CrossRef]

[37] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024, May). Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations* (Vol. 2024, pp. 9112-9141). https://proceedings.iclr.cc/paper_files/paper/2024/hash/25f7be9694d7b32d5cc670927b8091e1-Abstract-Conference.html

[38] Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. C. (2024, June). Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 7036-7050). [CrossRef]

[39] Meuwissen, T. H., Hayes, B. J., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *genetics, 157*(4), 1819-1829. [CrossRef]

[40] Washburn, J. D., Varela, J. I., Xavier, A., Chen, Q., Ertl, D., Gage, J. L., ... & de Leon, N. (2025). Global genotype by environment prediction competition reveals that diverse modeling strategies can deliver satisfactory maize yield estimates. *Genetics, 229*(2), iyae195. [CrossRef]

[41] Food and Agriculture Organization of the United Nations (2026). AGROVOC multilingual thesaurus. Retrieved from https://www.fao.org/agrovoc/. AGROVOC 2026 release. Accessed 21 June 2026.

[42] Haller, A., Janowicz, K., Cox, S., Le Phuoc, D., Taylor, K., & Lefrançois, M. (2017). *Semantic sensor network ontology. W3c recommendation, World Wide Web Consortium*. Retrieved from https://www.w3.org/TR/2017/REC-vocab-ssn-20171019/. Issued 19 October 2017. Accessed 19 October 2026.

[43] Lebo, T., Sahoo, S., & McGuinness, D. (2013). PROV-O: The PROV ontology. W3c recommendation, World Wide Web Consortium. Available online: https://www.w3.org/TR/2013/REC-p

rov-o-20130430/. Issued 30 April 2013. Accessed 19 October 2026.

[44] Green, T. J., Karvounarakis, G., & Tannen, V. (2007, June). Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 31-40). [CrossRef]

[45] Whitehill, J., Wu, T. F., Bergsma, J., Movellan, J., & Ruvolo, P. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems, 22*. https://proceedings.neurips.cc/paper_files/paper/2009/hash/f899139df5e1059396431415e770c6dd-Abstract.html

[46] Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of machine learning research, 11*(4). https://www.jmlr.org/papers/volume11/raykar10a/raykar10a.pdf

[47] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., ... & Weld, D. (2021, May). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-16). [CrossRef]

[48] Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications, 7*(1), 39-59. [CrossRef]

[49] Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., & Batra, N. (2020). PlantDoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD* (pp. 249-253). [CrossRef]

[50] George, R., Thuseethan, S., Ragel, R. G., Mahendrakumaran, K., Nimishan, S., Wimalasooriya, C., & Alazab, M. (2025). Past, present and future of deep plant leaf disease recognition: A survey. *Computers and Electronics in Agriculture, 234*, 110128. [CrossRef]

[51] Gauba, A., Pi, I., Man, Y., Pang, Z., Adve, V., & Wang, Y. X. (2026). Agmmu: A comprehensive agricultural multimodal understanding benchmark. *Advances in Neural Information Processing Systems, 38*. [CrossRef]

[52] Cheng, Z., Dong, H., Wang, Z., Jia, R., Guo, J., Gao, Y., ... & Zhang, D. (2022, May). Hitab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1094-1110). [CrossRef]

[53] Yan, L., Wang, H., Tang, C., Liu, H., Sun, T., Liu, L., ... & Jiang, J. (2026, March). Agrieval: A comprehensive chinese agricultural benchmark for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 40, pp. 34205-34213). [CrossRef]

[54] Zaremehrjerdi, H., Ganguly, S., Rairdin, A., Tranel, E., Feuer, B., Di Salvo, J. I., ... & Sarkar, S. (2025). Towards large reasoning models for agriculture. *arXiv preprint arXiv:2505.19259*. [CrossRef]

[55] Wen, Y., Li, Q., Ye, Z., Zhang, J., Fan, X., Mai, Z., ... & Zheng, J. (2025). AgroCoT: A Chain-of-Thought Benchmark for Evaluating Reasoning in Vision-Language Models for Agriculture. *arXiv preprint arXiv:2511.23253*. [CrossRef]

[56] Li, Q., Zhang, Y., Mai, Z., Chen, Y., Huang, H., Zhang, J., ... & Zheng, J. (2026). Can large multimodal models understand agricultural scenes? benchmarking with agromind. *Advances in Neural Information Processing Systems, 38*. [CrossRef]

[57] Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. *Fao, rome, 300*(9), D05109. https://www.avwatermaster.org/filingdocs/195/70653/172618e_5xAGWAx8.pdf

[58] Wageningen University & Research (2026). WOFOST: *World food studies*. Retrieved from https://www.wur.nl/en/research/products-services/wofost-world-food-studies. WOFOST 8.1 documentation. Accessed 21 June 2026.

[59] Liu, L., Zhou, W., Guan, K., Peng, B., Xu, S., Tang, J., ... & Jin, Z. (2024). Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. *Nature communications, 15*(1), 357. [CrossRef]

[60] Tanaka, Y., Watanabe, T., Katsura, K., Tsujimoto, Y., Takai, T., Tanaka, T. S. T., ... & Saito, K. (2023). Deep learning enables instant and versatile estimation of rice yield using ground-based RGB images. *Plant Phenomics, 5*, 0073. [CrossRef]

[61] Hemming, S., de Zwart, F., Elings, A., Righini, I., & Petropoulou, A. (2019). Remote control of greenhouse vegetable production with artificial intelligence—greenhouse climate, irrigation, and crop production. *Sensors, 19*(8), 1807. [CrossRef]

[62] Hemming, S., de Zwart, F., Elings, A., Petropoulou, A., & Righini, I. (2020). Cherry tomato production in intelligent greenhouses—sensors and AI for control of climate, irrigation, crop yield, and quality. *Sensors, 20*(22), 6430. [CrossRef]

[63] Maree, S. C., Zhang, P., van Marrewijk, B. M., de Zwart, F., Bijlaard, M., & Hemming, S. (2025). Autonomous greenhouse cultivation of dwarf tomato: Performance evaluation of intelligent algorithms for multiple-sensor feedback. *Sensors, 25*(14), 4321. [CrossRef]

[64] Morcego, B., Yin, W., Boersma, S., Van Henten, E., Puig, V., & Sun, C. (2023). Reinforcement learning versus model predictive control on greenhouse climate control. *Computers and Electronics in Agriculture, 215*, 108372. [CrossRef]

[65] Liu, Q., Hu, X., Zhang, Y., Shi, L., Wang, L., Yang, Y., ... & Qu, Z. (2025). Assimilating UAV observations

and crop model simulations for dynamic estimation of crop water stress. *Agricultural Water Management, 318*, 109688. [CrossRef]

[66] Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM, 57*(10), 78-85. [CrossRef]

[67] Li, J., Xu, M., Xiang, L., Chen, D., Zhuang, W., Yin, X., & Li, Z. (2024). Foundation models in smart agriculture: Basics, opportunities, and challenges. *Computers and Electronics in Agriculture, 222*, 109032. [CrossRef]

[68] Kuska, M. T., Wahabzada, M., & Paulus, S. (2024). AI for crop production – Where can large language models (LLMs) provide substantial value? *Computers and electronics in agriculture, 221*, 108924. [CrossRef]

[69] Tzachor, A., Devare, M., Richards, C., Pypers, P., Ghosh, A., Koo, J., ... & King, B. (2023). Large language models and agricultural extension services. *Nature food, 4*(11), 941-948. [CrossRef]

[70] Zhang, Y., Han, Z., Chen, J., Xu, C., Qin, Y., Wang, B., ... & Zhang, L. (2026). CropGPT: A large multimodal model for precise and explainable diagnosis of crop pests and diseases. *Journal of Industrial Information Integration*, 101107. [CrossRef]

[71] Shinoda, R., Inoue, N., Kataoka, H., Onishi, M., & Ushiku, Y. (2025, October). Agrobench: Vision-language model benchmark in agriculture. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 7634-7644). IEEE. [CrossRef]

[72] Qwen Team (2026). Qwen3.8-27B model card. Retrieved from https://huggingface.co/Qwen/Qwen3.8-27B. Accessed 12 September 2026.

[73] Qwen Team (2026). Qwen3.5-4B model card. Retrieved from https://huggingface.co/Qwen/Qwen3.5-4B. Hugging Face model card, Qwen3.5-4B release. Accessed 12 September 2026.

[74] Fellegi, I. P. & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183-1210. [CrossRef]

[75] Guo, J., Li, J., Li, D., Tiong, A. M. H., Li, B., Tao, D., & Hoi, S. (2023, June). From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10867-10877). IEEE. [CrossRef]

[76] Hu, Y., Hua, H., Yang, Z., Shi, W., Smith, N. A., & Luo, J. (2023, October). Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 2951-2963). IEEE. [CrossRef]

[77] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR. [CrossRef]

[78] Reimers, N., & Gurevych, I. (2019, November). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982-3992). [CrossRef]

[79] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), 1-37. [CrossRef]

[80] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321-1330). PMLR. [CrossRef]

[81] Ministry of Agriculture and Rural Affairs of the People's Republic of China (2023). Announcement no. 732 and attached transgenic cultivar approval descriptions. Retrieved from https://www.moa.gov.cn/govpublic/nybzzj1/202312/t20231207_6442285.htm. Issued 6 December 2023. In Chinese. Accessed 21 June 2026.

[82] Ministry of Agriculture and Rural Affairs of the People's Republic of China (2024). Announcement no. 830 and attached transgenic cultivar approval descriptions. Retrieved from https://www.moa.gov.cn/govpublic/nybzzj1/202410/t20241008_6463810.htm. Issued 30 September 2024. In Chinese. Accessed 21 June 2026.

[83] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. [CrossRef]

[84] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). [CrossRef]

[85] Genomes to Fields (2025). Genomes to Fields 2023 dataset. [CrossRef]

[86] McFarland, B. A., AlKhalifah, N., Bohn, M., Bubert, J., Buckler, E. S., Ciampitti, I., ... & De Leon, N. (2020). Maize genomes to fields (G2F): 2014–2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. *BMC research notes, 13*(1), 71. [CrossRef]

[87] AlKhalifah, N., Campbell, D. A., Falcon, C. M., Gardiner, J. M., Miller, N. D., Romay, M. C., ... & Lawrence-Dill, C. J. (2018). Maize genomes to fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets. *BMC research notes, 11*(1), 452. [CrossRef]

[88] Lima, D. C., Washburn, J. D., Varela, J. I., Chen, Q., Gage, J. L., Romay, M. C., ... & De Leon, N. (2023). Genomes to fields 2022 maize genotype by environment prediction competition. *BMC Research Notes, 16*(1), 148. [CrossRef]

[89] Bernardo, R. (1996). Best linear unbiased prediction of maize single-cross performance. *Crop Science, 36*(1), 50-56. [CrossRef]

[90] Technow, F., Schrag, T. A., Schipprack, W., Bauer, E., Simianer, H., & Melchinger, A. E. (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics, 197*(4), 1343-1355. [CrossRef]

[91] Lariepe, A., Moreau, L., Laborde, J., Bauland, C., Mezmouk, S., Décousset, L., ... & Charcosset, A. (2017). General and specific combining abilities in a maize (*Zea mays L.*) test-cross hybrid panel: relative importance of population structure and genetic divergence between parents. *Theoretical and Applied Genetics, 130*(2), 403-417. [CrossRef]

[92] Massman, J. M., Gordillo, A., Lorenzana, R. E., & Bernardo, R. (2013). Genomewide predictions from maize single-cross data. *Theoretical and applied genetics, 126*(1), 13-22. [CrossRef]

[93] Escamilla, D. M., Li, D., Negus, K. L., Kappelmann, K. L., Kusmec, A., Vanous, A. E., ... & Yu, J. (2025). Genomic selection: Essence, applications, and prospects. *The Plant Genome, 18*(2), e70053. [CrossRef]

[94] Fernandes, I. K., Vieira, C. C., Dias, K. O., & Fernandes, S. B. (2024). Using machine learning to combine genetic and environmental data for maize grain yield predictions across multi-environment trials. *Theoretical and Applied Genetics, 137*(8), 189. [CrossRef]

[95] Kusmec, A., Negus, K. L., & Yu, J. (2026). Critical evaluation of the theory and practice of feed-forward neural networks for genomic prediction. *G3: Genes, Genomes, Genetics, 16*(3), jkaf314. [CrossRef]

[96] Kick, D. R., Wallace, J. G., Schnable, J. C., Kolkman, J. M., Alaca, B., Beissinger, T. M., ... & Washburn, J. D. (2023). Yield prediction through integration of genetic, environment, and management data through deep learning. *G3: Genes, Genomes, Genetics, 13*(4), jkad006. [CrossRef]

[97] Lapajne, J., Vončina, A., Vojnović, A., Donša, D., Dolničar, P., & Žibrat, U. (2025). Field-scale UAV-based multispectral phenomics: Leveraging machine learning, explainable AI, and hybrid feature engineering for enhancements in potato phenotyping. *Computers and Electronics in Agriculture, 229*, 109746. [CrossRef]

[98] Van De Vijver, R., Mertens, K., Heungens, K., Nuyttens, D., Wieme, J., Maes, W. H., ... & Saeys, W. (2022). Ultra-high-resolution UAV-based detection of Alternaria solani infections in potato fields. *Remote Sensing, 14*(24), 6232. [CrossRef]

[99] Duarte-Carvajalino, J. M., Alzate, D. F., Ramirez, A. A., Santa-Sepulveda, J. D., Fajardo-Rojas, A. E., & Soto-Suárez, M. (2018). Evaluating late blight severity in potato crops using unmanned aerial vehicles and machine learning algorithms. *Remote Sensing, 10*(10), 1513. [CrossRef]

[100] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56-67. [CrossRef]

[101] Reichle, R. H. (2008). Data assimilation methods in the Earth sciences. *Advances in water resources, 31*(11), 1411-1418. [CrossRef]

[102] Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., ... & Ritchie, J. T. (2003). The DSSAT cropping system model. *European journal of agronomy, 18*(3-4), 235-265. [CrossRef]

[103] National Crop Variety Approval Committee (2022). *National transgenic maize cultivar approval standard (trial)*. Retrieved from https://www.moa.gov.cn/govpublic/nybzzj1/202206/P020220608572965836389.pdf. Issued 8 June 2022. In Chinese. Accessed 21 June 2026.

A Source Scope and Documentary Decisions

A.1 Source Material and Extraction

The source registry is a local collection of national approval descriptions obtained from a public variety-approval query service. It contains 9,960 registrations across five crops, namely 3,321 maize, 3,005 rice, 2,358 wheat, 778 soybean and 498 cotton, and the approval years in its registration numbers run from 2018 to 2024. The rice and maize subset contains 6,326 rows and 6,280 distinct registration identifiers. Its fields include parentage, characteristics, yield performance, cultivation guidance and approval opinions. The categorical grammar extracts 27,451 resistance statements after removing 207 duplicate literal assertions. The duration grammar keeps one duration per type zone, and records whose type zone is missing or holds several distinct durations are set aside, which leaves 25,486 disease-grade records in 5,885 registrations of 5,575 cultivars under 64 type zones. The official 2023 and 2024 approval documents provide the matched source passages for development [6, 7]. The registry text was checked against the native text of announcements 726 and 867. Of 2,683 characteristics fields, 2,680 match exactly, and the three exceptions are two substantive text differences and one reading-order problem in the native document. The 2025 announcement was parsed from its own official document, and none of its 1,267 registration numbers occurs in the registry.

Normalization preserves the original expressions and their source locations. Explicit type zones delimit profiles. Disease-grade expressions govern only the diseases within their stated list, ending at a new grade, numerical assessment or list boundary. Numerical disease indices become categorical grades only when the source explicitly supplies that category. Specific pathogen assessments remain distinct from generic disease names.

The resistance scale has six ordered categories, coded from 0 to 5 as highly resistant, resistant, moderately resistant, moderately susceptible, susceptible and highly susceptible. The moderate-resistance request accepts codes 0–2. The request excluding high susceptibility accepts codes 0–4. A duration comparison also requires an identical endpoint. Rice uses whole growth duration, and the evaluated grain-maize requests use emergence to maturity.

Rice panels are blast alone, blast with bacterial blight, and blast with brown planthopper resistance. Eastern and Northern maize uses northern leaf blight, head smut and gray leaf spot. Huang-Huai-Hai maize uses southern leaf blight, southern rust, Curvularia leaf spot and common smut. Southwest maize uses northern leaf blight, gray leaf spot, sheath blight and head smut. The regional maize panels follow the named diseases in the 2021 approval standard [4].

Table A2 reports every development-cohort panel. Unresolved denotes resistance evidence insufficient to decide the panel before a duration ceiling is imposed. None means no candidate has the required documentary support. Reported durations are source descriptors and are not predictions of maturity at a new farm.

The fixed test-cohort rules recover all 1,267 registration names and characteristics fields from the official 2025 announcement [5]. Fourteen of the 1,189 rice and maize records have no recognized type zone, and 24 profiles have missing or multiple distinct durations. The 1,200 accepted profiles supply 5,047 disease-grade assertions with source spans. The six evaluated zones contain 567 profiles in 558 registrations. Table A3 gives their request and context-dependence counts by type zone together with the association ablation of Appendix A.7.

A.2 Cross-Zone Differences

The cross-zone analysis reads every disease-grade claim of the rice and maize registry together with the type zone or test group under which the claim is reported. A pair is a cultivar and a disease with claims in two or more zones. Grades use the six-category scale coded 0 to 5, and a pair crosses the resistance boundary when its lowest code is at most 2 (MR or better) and its highest at least 3 (MS or worse). Durations are compared only within one endpoint, whole growth period for rice and emergence to maturity for maize, and a cultivar's range is its longest minus its shortest documented duration across zones. Table A1 gives the exact counts behind Figure 3. The pairs split into 969 documented within one registration and 803 documented across

Table A1. Cultivar–disease pairs documented in two or more type zones. Different grades counts pairs whose grades are not identical, and Cross boundary counts pairs with MR or better in one zone and MS or worse in another. Shares divide by the pairs in the row.

Group	Pairs	Different grades	Cross boundary
Rice	756	287 (38.0%)	99 (13.1%)
Maize	1,016	391 (38.5%)	261 (25.7%)
All pairs	1,772	678 (38.3%)	360 (20.3%)
One registration	969	351 (36.2%)	187 (19.3%)
Several registrations	803	327 (40.7%)	173 (21.5%)
Rice bacterial blight	139	80 (57.6%)	5 (3.6%)
Rice blast	323	185 (57.3%)	91 (28.2%)
Maize stalk rot	215	106 (49.3%)	74 (34.4%)
Maize ear rot	215	99 (46.0%)	66 (30.7%)
Southern leaf blight	80	30 (37.5%)	24 (30.0%)
Head smut	123	43 (35.0%)	25 (20.3%)
Common smut	27	9 (33.3%)	5 (18.5%)
Northern leaf blight	144	48 (33.3%)	33 (22.9%)
Southern rust	54	15 (27.8%)	7 (13.0%)
Sheath blight	54	13 (24.1%)	11 (20.4%)
Gray leaf spot	85	20 (23.5%)	16 (18.8%)
Brown planthopper	294	22 (7.5%)	3 (1.0%)

registrations. Disease shares are reported for diseases with at least 25 pairs. The brown planthopper is the only insect pest in this set.

A.3 Retrieval Policy Comparison

Table A4 lists every policy setting evaluated on the sample of 1,000 requests. Error is the workload-weighted share of wrong answers, and Read is the share of requests that read the source passage. Input and Output are the tokens per request. Invalid and abstain count outputs that could not be parsed and refusals to answer. The sufficiency-and-confidence rows name the target read share that sets the threshold, and the MBA-RAG rows name the weight on access cost. The coarse-trigger and profile-trigger rows use the bounds only to choose which requests go to the common readers. Table A5 reports the paired comparison. Routing changes whether the multi-profile rule or profile agreement chooses the requests that read the source, and execution changes whether a model reader or the shortlist rule answers. Workload is the change in tokens per request over the weighted workload, Registration is the mean and sample standard deviation of the change over the 229 registrations, and Token saving is the percentage change in the weighted tokens.

A.4 Cross-Zone Prediction

Instances are the 3,751 combinations of cultivar, disease and zone with an official point grade, taken from cultivar–disease pairs that have grades in two or more zones. Where a cultivar has several registrations in one zone, the median grade is used. Grades are coded 0 to 5 from highly resistant to highly susceptible. The inputs are the minimum, maximum, mean, median, range and count of the cultivar's grades for the disease in its other zones, one-hot codes for crop, disease and zone, and the earliest approval year. Logistic regression uses standardized inputs, an inverse regularization strength of 0.3 and up to 5,000 iterations. The random forest has 500 trees with a minimum leaf size of two. Gradient boosting uses the histogram implementation with 200 iterations and a learning rate of 0.05. Folds assign whole cultivars, and each of five seeds draws a new assignment. Table A6 gives means and standard deviations over the seeds. The exact sign tests count cultivars for which the best learner has a higher or lower averaged accuracy than the comparison rule. Cultivars with equal accuracy are excluded from the test. The inputs are limited to grades and approval metadata.

Table A2. All development-cohort zones and panels. MR requires at least moderate resistance, and Not HS excludes high susceptibility. Context and Pooled are the earliest reported durations in days.

Protection requirement	Threshold	Candidates	Supported	Unresolved	Context	Pooled
Upper Yangtze, single-season rice						
Blast	MR	115	7	0	146.1	110.2
Blast	Not HS	115	115	0	144.7	110.2
Blast + bacterial blight	MR	115	0	7	none	none
Blast + bacterial blight	Not HS	115	0	115	none	110.2
Blast + brown planthopper	MR	115	0	0	none	none
Blast + brown planthopper	Not HS	115	4	0	147.7	147.7
Middle/lower Yangtze, single-season rice						
Blast	MR	344	76	0	121.3	114.5
Blast	Not HS	344	344	0	121.3	111.8
Blast + bacterial blight	MR	344	0	0	none	116.7
Blast + bacterial blight	Not HS	344	314	0	121.3	111.8
Blast + brown planthopper	MR	344	0	0	none	none
Blast + brown planthopper	Not HS	344	13	0	126.0	126.0
Middle/lower Yangtze, double-season late rice						
Blast	MR	131	22	0	113.7	110.2
Blast	Not HS	131	131	0	107.6	107.6
Blast + bacterial blight	MR	131	4	0	116.0	110.9
Blast + bacterial blight	Not HS	131	115	0	110.6	110.2
Blast + brown planthopper	MR	131	0	0	none	none
Blast + brown planthopper	Not HS	131	6	0	109.0	109.0
Eastern/northern, middle-late spring maize						
Regional disease panel	MR	175	2	0	127.8	127.8
Regional disease panel	Not HS	175	168	0	125.8	101.6
Huang-Huai-Hai, summer maize						
Regional disease panel	MR	420	0	0	none	none
Regional disease panel	Not HS	420	26	0	99.6	99.6
Southwest, middle/high-altitude spring maize						
Regional disease panel	MR	95	3	0	130.1	130.1
Regional disease panel	Not HS	95	77	1	125.5	116.8

Table A3. Test-cohort workload and association ablation by type zone. Flagged counts requests whose coarse bounds disagree and the second Profiles column counts the distinct evaluated profiles with at least one such request. The four ablation columns count unsupported positive answers to candidate–request pairs using neither zone association, duration alone, resistance alone, or complete anonymous profiles.

	Test-cohort workload				Unsupported additions			
Type zone	Profiles	Requests	Flagged	Profiles	Neither	Duration	Resistance	Joint
Upper Yangtze rice	47	13,536	600	8	600	190	204	600
Middle/lower Yangtze rice	178	102,528	1,107	7	0	0	0	0
Double-season late rice	65	18,330	7	1	3	3	0	1
Eastern/northern maize	74	5,476	112	6	23	0	23	0
Huang-Huai-Hai maize	171	15,732	0	0	0	0	0	0
Southwest maize	32	1,920	30	1	18	0	18	0
Total	567	157,522	1,856	23	644	193	245	601

Table A4. Policies on the common-evidence sample. Error and Read are weighted percentages, and Input and Output are tokens per request. Invalid and abstain give the counts of unparsable outputs and refusals.

Method	Error (%)	Read (%)	Input	Output	Invalid/abstain
Anonymous profiles, common reader	0.563	0.00	367.6	323.0	112/0
Complete source, common reader	0.000	100.00	442.7	242.2	0/0
Coarse bounds + source reader	0.000	1.18	8.0	3.1	0/0
Profile agreement + source reader	0.000	1.15	7.8	3.0	0/0
Coarse trigger, common readers	0.000	1.18	370.5	311.3	0/0
Profile trigger, common readers	0.000	1.15	370.5	311.4	0/0
Sufficient Context autorater	0.000	6.55	1376.4	1734.3	0/0
Sufficiency + confidence, 0% target	0.233	0.33	1721.0	1761.2	0/0
Sufficiency + confidence, 1.2% target	0.174	0.85	1723.7	1762.4	0/0
Sufficiency + confidence, 5.2% target	0.009	4.28	1740.1	1770.3	0/0
Sufficiency + confidence, 10% target	0.009	5.36	1744.9	1772.9	0/0
Sufficiency + confidence, 30% target	0.000	15.84	1790.1	1797.7	0/0
Sufficiency + confidence, 50% target	0.000	15.84	1790.1	1797.7	0/0
Sufficiency + confidence, 100% target	0.000	100.00	2161.4	2002.6	0/0
A2RAG-style verified recovery	0.003	1.22	1238.9	643.6	0/0
Triple-check routing control	0.003	1.22	1225.4	640.0	0/0
MBA-RAG adaptation, cost 0.01	0.127	56.02	413.1	276.1	27/0
MBA-RAG adaptation, cost 0.1	0.465	10.68	377.3	313.5	96/0
MBA-RAG adaptation, cost 0.3	0.563	0.00	367.6	323.0	112/0

Table A5. Paired comparison of routing and execution. Each row changes one factor and holds the other fixed.

Changed factor	Workload Δ	Registration Δ	Token saving (%)
Routing, model readers	-1.19	3.52 ± 33.18	0.17
Routing, direct execution	-37.86	-87.00 ± 254.30	77.67
Execution, multi-profile rule	-634.38	-601.33 ± 222.02	92.87
Execution, profile agreement	-671.04	-691.84 ± 106.31	98.40

Table A6. Prediction of a withheld grade from a cultivar's other zones. Entries are percentages, except mean error, which is in grade steps, with standard deviations over five seeds in parentheses. Within one step counts predictions within one grade of the official grade, and resistant side counts agreement on MR or better against MS or worse.

Method	Exact	Within one step	Resistant side	Mean error
Global majority	37.7 (0.0)	74.1 (0.0)	74.1 (0.0)	0.93 (0.00)
Zone–disease majority	66.4 (0.4)	83.8 (0.4)	81.6 (0.4)	0.52 (0.01)
Median of other zones	62.4 (0.0)	83.6 (0.0)	80.4 (0.0)	0.57 (0.00)
Best of other zones	63.0 (0.0)	82.7 (0.0)	80.4 (0.0)	0.58 (0.00)
Worst of other zones	62.9 (0.0)	82.6 (0.0)	80.8 (0.0)	0.59 (0.00)
Logistic regression	69.1 (0.2)	86.5 (0.1)	84.3 (0.2)	0.46 (0.00)
Random forest	69.5 (0.4)	86.5 (0.1)	84.1 (0.1)	0.46 (0.00)
Gradient boosting	67.6 (0.4)	85.4 (0.2)	83.1 (0.2)	0.50 (0.01)

A.5 Partial Grades and Shared Assessments

For a nonempty admissible grade set V and an acceptable set A, the assertion supports the requirement when $V \subseteq A$, excludes it when $V \cap A = \varnothing$, and leaves it unresolved otherwise. Missing evidence is unresolved. Before these comparisons, the subject, method and requested contextual conditions must match. A conjunction is supported when every requirement is supported, excluded when at least one is excluded, and unresolved otherwise. The binary shortlist includes only supported candidates.

For example, a statement excluding only high susceptibility gives $V = \{0, 1, 2, 3, 4\}$. It supports exclusion of high susceptibility but leaves moderate resistance unresolved. An optimistic summary accepts the overlap with $\{0, 1, 2\}$, whereas documentary support requires the full admissible set to meet the request. Equations 4 and 6 preserve that distinction.

Consider two assessments and two disease requirements. Suppose one assessment meets only the first requirement and the other meets only the second. Each disease has a favorable value somewhere, but no single assessment meets their conjunction. If $B_{k,d}$ indicates point-grade acceptance for assessment k and disease d, then

$$\exists k \, \forall d \, B_{k,d} \quad \Longrightarrow \quad \forall d \, \exists k \, B_{k,d}. \tag{A1}$$

The example shows why the converse of Equation A1 fails. Excluding every assessment means that each has a failed requirement. The failed disease may differ between assessments.

Uncertain association also distinguishes a supported existence claim from identification of a particular assessment. With grade rows $(0, 0)$ and $(0, 1)$ at threshold zero, every independent permutation of the disease columns leaves some qualifying row, but the qualifying row's identity can change. A supported existential answer therefore need not recover a fixed named assessment. These logical distinctions concern supplied records and threshold conditions.

A.6 Request Families and Later Reuse

The association cohort uses paired categorical disease grades from the 2023 and 2024 transgenic-maize approvals. The source grouping is a reported assessment under one type zone. A disease enters a matrix only when its identity is resolved and both assessment methods supply one explicit point grade. The 59 contributing registrations yield 251 two-disease and 134 three-disease matrices. Each matrix appears at all three shared resistant cutoffs, giving 1,155 cases. Repeated cultivar families and overlapping disease subsets remain dependent observations.

Association removal permits each disease's two values to exchange assessment names independently. For each compatible assignment, a named assessment qualifies only when all its grades meet the specified thresholds. An existential request accepts any qualifying assessment, an all-assessment request requires both, and a separate-disease request permits a different assessment for each disease. Exhaustive assignment gives the final supported, excluded or unresolved state. An independent calculation from the number of accepted grades in each column agrees on all 5,775 case–family comparisons. With the original associations, 377 of the 1,155 cases support the existence of a complete qualifying assessment, 83 support all assessments, 411 support each disease separately, and 103 and 357 support the named inoculation and field assessments. After removal, the existential request keeps 256 supported and 744 excluded cases with 155 unresolved, the all-assessment and separate-disease requests keep every answer, and each named request keeps 83 supported and 744 excluded cases with 328 unresolved (Figure 4(a)).

Across every compatible assignment, the largest grade for each disease determines the all-assessment answer, while the smallest determines the separate-disease answer. Complete enumeration verifies those summaries across all 37,980 disease-specific threshold vectors on the 385 matrices. For requests naming either assessment, the unordered values permit a global exchange of names. The two named requests therefore have identical uncertainty after removal even though their original answers differ. Their 328 unresolved cases include 155 with uncertain existence and 173 with certain existence but no individually identifiable qualifying assessment.

The threshold-reuse analysis compares a best-grade summary with existence of a complete qualifying assessment. The summary preserves every disease-specific threshold answer exactly when some original row attains all the per-disease minima. Such a row supplies every supported summary answer. If it is absent, setting each threshold to its corresponding minimum makes the summary pass while every original row fails at least one condition.

Over the full six-grade domain of 37,980 cutoff vectors, the summary gives at least one incorrect existential answer for 34 of the 251 two-disease matrices and 40 of the 134 three-disease matrices. A shared cutoff across diseases exposes 19 and 14 of those losses, leaving 15 and 26 that appear only under disease-specific cutoffs. Restricting cutoffs to the highly resistant, resistant and moderately resistant grades gives 5,877 vectors and the same two-disease counts, while the three-disease losses fall to 33, of which 14 are exposed by a shared cutoff and 19 remain hidden. The 2021 maize standard motivates distinguishing requirements involving both methods from requirements specific to field assessments [4]. These selected disease subsets and cutoff vectors

serve a controlled information analysis. Applying the complete regional approval rules would require their additional criteria.

A.7 Association Ablation by Type Zone

Table A3 decomposes the association ablation by type zone. Every variant uses identical disease-grade sets and applies the full-set acceptance rule. Pooling is restricted to endpoint-compatible profiles. On this workload its answers coincide with the original optimistic bound, although these definitions can differ for other records. Both named associations give zero errors in every zone. All ablated variants with a positive documentary answer remain positive, so their errors are unsupported additions. These are exact counts in the evaluated workload, with repeated requests within registrations. The 644 additions without either association affect 12 registrations, the 193 with the duration alone affect eight, the 245 with the resistance ratings alone affect 11, and the 601 under anonymous complete profiles affect nine.

B Multimedia Annotation and Evidence Reuse

B.1 Tomato Image Recognition

The ten categories are bacterial spot, early blight, late blight, leaf mold, Septoria leaf spot, target spot, mosaic virus, yellow leaf curl virus, two-spotted spider mites and healthy leaves. The frozen ResNet-18 encoder uses the torchvision ImageNet-1K V1 checkpoint. Images are converted to RGB, resized bilinearly with antialiasing to a 256-pixel shorter edge and center-cropped to 224×224 pixels. Pixel values scaled to $[0, 1]$ are standardized with channel means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$. The 512-dimensional global-average-pooled features are L2-normalized and support a refitted multinomial logistic classifier. Feature extraction runs in evaluation mode with batches of 128 images. Inverse regularization is chosen from the fixed grid $C \in \{0.1, 1, 10\}$ by source-validation macro F1 using L-BFGS (at most 1,000 iterations, seed 20260912), selecting $C = 10$. Temperature scaling minimizes source-validation negative log likelihood by bounded scalar optimization over $[0.25, 5]$, yielding a temperature of 0.62753 [80]. Exact decoded-pixel identities are disjoint across the accepted partitions.

Source-test accuracy is 0.926 and ten-class macro F1 is 0.914. On the complete external collection, the corresponding values are 0.244 and 0.118. Restricting external macro F1 to its nine represented classes gives 0.131. These are descriptive scores on fixed image sets. The fusion experiment uses separate external calibration and evaluation subsets, whose simulated-run variation appears in Table A7.

B.2 Simulated Contributions and Pooling

For nominal error rate η, a worker's initial class accuracy is $1 - \eta + \delta$, where δ is normal with mean zero and standard deviation 0.06. Accuracy is clipped to $[0.3, 0.99]$. A specialty increases accuracy by 0.06 within one of six groups covering bacterial, fungal, oomycete and viral disease, mites and healthy leaves. That adjusted accuracy is clipped to $[0.3, 0.995]$. Item difficulty subtracts 0.16ξ, with $\xi \sim \mathrm{Beta}(2, 5)$, and the final accuracy lies in $[0.2, 0.995]$. Incorrect classes from the same group receive three times the sampling weight of other incorrect classes.

The simulation uses 32 workers, seeds 11, 22, 33, 44, 55, 66, 77, 88, 99 and 110, nominal error rates from 0.1 to 0.4 and budgets of one, three or five responses per evaluation image. Worker calibration always uses five responses per source calibration item. Each worker's confusion matrix estimates response probabilities given the reference class, with a pseudocount of 0.5 per cell. A uniform prior and the product of those probabilities give the worker-derived class probabilities.

Let π_g and ρ_g denote the visual-model and worker-derived probabilities for class g. Linear and logarithmic pooling use

$$\omega_g^{\mathrm{lin}} = \alpha \pi_g + (1 - \alpha)\rho_g, \qquad \omega_g^{\mathrm{log}} = \frac{\pi_g^{\beta_{\mathrm{v}}} \rho_g^{\beta_{\mathrm{c}}}}{\sum_{g'} \pi_{g'}^{\beta_{\mathrm{v}}} \rho_{g'}^{\beta_{\mathrm{c}}}}. \tag{A2}$$

In Equation A2, the linear weight $\alpha \in [0, 1]$ minimizes calibration negative log likelihood by bounded scalar optimization, also checking both endpoints. The independently bounded logarithmic weights $\beta_{\mathrm{v}}, \beta_{\mathrm{c}} \in [0, 2]$

jointly minimize the same loss using L-BFGS-B with analytic gradients, initialized at $(1, 1)$. Probabilities are clipped below at 10^{-12} before logarithms. Source validation is split by class, with 65 percent for worker calibration and the remainder for pool calibration. External calibration reserves 25 percent of each represented class. The transferred-weight condition uses the source-fitted linear weight externally. The fixed product normalizes $\pi_g^{0.9}\rho_g$, and equal linear pooling sets $\alpha = 0.5$.

All methods share responses and evaluation images within a condition. In the dependence condition, the model's predicted class replaces a worker response with probability 0.3. This is a controlled model-copying simulation, not a measurement of producer behavior.

Table A7. All fusion methods at nominal error rate 0.3 and five responses per image. Values are ten-class macro F1, mean ± sample standard deviation over ten seeds. Nine represented external classes give a maximum of 0.9 for this fixed ten-class average.

	Independent responses		Model copying at 0.3	
Method	Source	External	Source	External
Majority vote	0.902 ± 0.004	0.823 ± 0.018	0.932 ± 0.005	0.577 ± 0.042
Worker confusion model	0.875 ± 0.007	0.763 ± 0.018	0.912 ± 0.009	0.579 ± 0.019
Image model	0.914 ± 0.000	0.116 ± 0.006	0.914 ± 0.000	0.116 ± 0.006
Fixed product	0.988 ± 0.001	0.388 ± 0.012	0.958 ± 0.002	0.239 ± 0.015
Equal linear pool	0.980 ± 0.003	0.685 ± 0.036	0.955 ± 0.002	0.428 ± 0.030
Calibrated linear pool	0.949 ± 0.014	0.763 ± 0.018	0.947 ± 0.005	0.577 ± 0.019
Calibrated logarithmic pool	0.988 ± 0.001	0.779 ± 0.027	0.957 ± 0.002	0.579 ± 0.019
Transferred linear weight	0.949 ± 0.014	0.191 ± 0.102	0.947 ± 0.005	0.313 ± 0.117

B.3 Assisted Clustering

Training-image features determine 32 whitened principal components, with PCA seed 20260912. Visual features and any added probability block are normalized separately before concatenation with equal block weights. K-means uses ten clusters, matching the ten diagnostic categories, and ten initializations across all representations. Clustering uses seeds 11, 22, 33, 44 and 55, five independent simulated responses per image and nominal error rate 0.3. Because the added class probabilities carry annotation or model information, the comparison evaluates assisted grouping, in which label information guides the clusters.

Visual features alone give an adjusted Rand index of 0.427 ± 0.008 on source images and 0.073 ± 0.007 on external photographs (mean ± sample standard deviation over seeds). Adding the image model's probabilities raises the source index to 0.846 ± 0.051 and leaves the external index at 0.068 ± 0.012. Adding worker-derived probabilities gives 0.836 ± 0.012 and 0.757 ± 0.051, and adding the calibrated linear pool gives 0.915 ± 0.020 and 0.745 ± 0.054 (Figure 6(b)). Normalized mutual information follows the same ordering, rising from 0.545 ± 0.006 to 0.891 ± 0.023 on source images and from 0.174 ± 0.008 to 0.789 ± 0.032 on external photographs, with 0.832 ± 0.016 and 0.157 ± 0.010 for image probabilities and 0.803 ± 0.010 and 0.792 ± 0.029 for worker probabilities.

Simulation seeds are the experimental units for inference. Each seed fixes the worker pool, the item difficulties and the calibration and evaluation partitions, so two methods or two copying rates evaluated under one seed form a matched pair. Table A8 reports exact two-sided sign-flip permutation tests of the mean paired difference. The test enumerates all 2^n sign assignments for n seeds and assumes only that the paired differences are symmetric about zero under the null hypothesis. With ten seeds the smallest attainable p is $2/1{,}024 \approx 0.0020$. With the five clustering seeds it is $2/32 = 0.0625$, so a clustering difference with the same sign in every seed cannot reach a smaller value.

B.4 Agricultural Images and Reusable Descriptions

The analyzed AgMMU release contains 772 entries [51]. Two exact duplicates, one missing answer and one truncated image are excluded, leaving 768 cases. The release differs from the article's reported 746-case population. Original questions, background, option order and answer keys remain unchanged. The background reflects the benchmark's question-construction process.

Table A8. Exact paired sign-flip permutation tests for the tomato comparisons at nominal error rate 0.3 and five responses per image. Differences are paired by simulation seed and summarized by their mean ± sample standard deviation. Clustering rows compare a representation with visual features alone, except the last row, whose reference adds worker probabilities.

Paired difference	Images, workers	Seeds	Mean ± s.d.	Exact p
Macro F1 of fused labels				
Calibrated logarithmic pool − majority vote	source, independent	10	$+0.086 \pm 0.005$	0.0020
Calibrated logarithmic pool − majority vote	external, independent	10	-0.044 ± 0.026	0.0039
Calibrated logarithmic pool − fixed product	external, independent	10	$+0.392 \pm 0.028$	0.0020
Calibrated logarithmic pool − majority vote	external, copying	10	$+0.002 \pm 0.029$	0.8555
Majority vote, copying − independent	external	10	-0.246 ± 0.052	0.0020
Calibrated logarithmic pool, copying − independent	external	10	-0.200 ± 0.036	0.0020
Adjusted Rand index of case clusters				
Image probabilities added	source, independent	5	$+0.419 \pm 0.051$	0.0625
Image probabilities added	external, independent	5	-0.005 ± 0.009	0.3750
Worker probabilities added	external, independent	5	$+0.684 \pm 0.047$	0.0625
Calibrated pool added to worker probabilities	external, independent	5	-0.012 ± 0.027	1.0000

Qwen3.8-27B and Qwen3.5-4B describe each case before receiving its explicit question, options or category [72, 73]. The instruction asks for at most 400 words and preserves image identifiers, visible features and uncertainty. Every listed image is available during description generation. Images are processed with native visual components, bounded between 256^2 and $1{,}024^2$ pixels. Descriptions use greedy generation with a 2,048-token ceiling and remain verbatim in subsequent calls.

Each answering condition receives the original background, question and options. Greedy answering uses a 64-token ceiling with an output schema restricted to the four option labels. Background, images, description and combined evidence each receive an independent call. The larger-model reasoning follow-up keeps the original descriptions and image processing, enables native thinking, and uses temperature 1.0, top-p 0.95, top-k 20 and a 12,288-token ceiling including reasoning. These are complete inference-configuration comparisons. With background alone, images, descriptions and both, Qwen3.8-27B under greedy decoding answers 580, 673, 658 and 664 of the 768 cases correctly, Qwen3.5-4B answers 528, 629, 603 and 603, and Qwen3.8-27B with reasoning answers 567, 681, 645 and 667. All conditions include the released textual background, and invalid and unfinished answers remain incorrect. Table A10 divides these counts by question category.

All 6,144 greedy answers finish with a valid option. One smaller-model description exhausts its allowance and remains in the downstream evidence. The larger-model reasoning comparison produces 75 unfinished answers, including 52 with background alone. They remain incorrect in the primary denominator. Assigning any final option to unfinished responses while fixing completed answers leaves the combined condition ahead of description alone by 8–28 correct answers. This is an attainable-score range on saved outputs.

Table A9 reports corrections and new errors for every primary evidence contrast, and Table A10 shows the heterogeneity across the complete category populations. Cases sharing identical decoded images are related observations.

The crossed follow-up varies the description producer independently of the answerer. Every model pairing receives new answers with and without images, including new same-model repetitions. All 6,144 responses finish normally. Fifteen final options change between the 3,072 repeated same-model answers and their earlier counterparts. Using only the newly generated answers, Qwen3.8-27B answers 658 cases correctly from its own descriptions and 668 once images are added, with 21 corrections and 11 new errors. From Qwen3.5-4B descriptions it answers 643 and 664, with 38 corrections and 17 new errors. Qwen3.5-4B answers 630 cases from Qwen3.8-27B descriptions with or without images, with 20 corrections offset by 20 new errors, and 602 and 603 from its own descriptions, with 23 corrections and 22 new errors. Adding images therefore corrects and introduces errors in every pairing.

Table A9. Paired AgMMU answer changes. The named evidence on the left of each comparison replaces the evidence on the right. Net is the percentage-point accuracy difference on 768 cases.

Configuration	Comparison	Corrections	New errors	Net (pp)
Qwen3.8-27B, greedy	images vs background	113	20	+12.11
Qwen3.8-27B, greedy	description vs images	27	42	-1.95
Qwen3.8-27B, greedy	description + images vs description	20	14	+0.78
Qwen3.8-27B, greedy	description + images vs images	28	37	-1.17
Qwen3.5-4B, greedy	images vs background	127	26	+13.15
Qwen3.5-4B, greedy	description vs images	49	75	-3.39
Qwen3.5-4B, greedy	description + images vs description	24	24	+0.00
Qwen3.5-4B, greedy	description + images vs images	39	65	-3.39
Qwen3.8-27B, reasoning	images vs background	149	35	+14.84
Qwen3.8-27B, reasoning	description vs images	33	69	-4.69
Qwen3.8-27B, reasoning	description + images vs description	47	25	+2.86
Qwen3.8-27B, reasoning	description + images vs images	31	45	-1.82

Table A10. All AgMMU categories. Entries give correct answers over the complete category population under each inference configuration.

Configuration	Category	Background	Images	Description	Both
Qwen3.8-27B, greedy	disease/issue identification	98/140	105/140	112/140	108/140
Qwen3.8-27B, greedy	insect/pest	110/143	125/143	120/143	120/143
Qwen3.8-27B, greedy	management instructions	147/161	151/161	147/161	147/161
Qwen3.8-27B, greedy	species	81/145	127/145	117/145	125/145
Qwen3.8-27B, greedy	symptom/visual description	144/179	165/179	162/179	164/179
Qwen3.5-4B, greedy	disease/issue identification	86/140	104/140	102/140	99/140
Qwen3.5-4B, greedy	insect/pest	90/143	113/143	106/143	107/143
Qwen3.5-4B, greedy	management instructions	144/161	145/161	144/161	141/161
Qwen3.5-4B, greedy	species	74/145	108/145	106/145	108/145
Qwen3.5-4B, greedy	symptom/visual description	134/179	159/179	145/179	148/179
Qwen3.8-27B, reasoning	disease/issue identification	105/140	108/140	108/140	112/140
Qwen3.8-27B, reasoning	insect/pest	107/143	125/143	116/143	118/143
Qwen3.8-27B, reasoning	management instructions	150/161	151/161	150/161	147/161
Qwen3.8-27B, reasoning	species	73/145	131/145	116/145	127/145
Qwen3.8-27B, reasoning	symptom/visual description	132/179	166/179	155/179	163/179

B.5 Reuse Across Agricultural Statistical Questions

The HiTab source supplies natural questions, table hierarchies and answer-cell annotations [52]. A lexical title screen selects agricultural land, crop, farm-operation and agricultural-workforce tables. All 65 matching tables and 205 questions remain. They come from 12 Statistics Canada reports and include 158 training, nine development and 38 corrected-test questions from the upstream partitions. The evaluated model is fixed without task fine-tuning.

The exact reference expands full row and column header paths while preserving all body values, coordinates and header context, including aggregate values outside the released body boundary. Every original cell remains recoverable. Qwen3.8-27B extracts paths and values once per table without questions or answer annotations. Source checking corrects supplied fields at unique source coordinates. Missing or ambiguous entries remain unavailable.

Extraction uses greedy decoding and a 12,288-token ceiling. Greedy answering has a 256-token ceiling. Reasoning uses the larger-model configuration specified for AgMMU. All conditions share the same saved extraction. The primary scorer is the released ordered-denotation comparison with numeric tolerance 10^{-5}. Invalid or unfinished answers count as incorrect. A sensitivity analysis permits answer-list reordering while preserving scalar normalization and multiplicity.

Of the 205 questions, greedy answering from the source hierarchy answers 136 correctly, from the exact expansion 117, from model extraction 107 and from source-checked extraction 111. The corresponding

reasoning counts are 161, 156, 151 and 148. Accuracy among the 12 source reports ranges from 0 to 100 percent for every condition except greedy answering from model extraction, which ranges from 0 to 76.9 percent, and reasoning from model extraction, which ranges from 56.5 to 100 percent. These ranges describe heterogeneity across the reports’ different question sets.

Three of 65 extractions exhaust the output allowance. All greedy responses are complete, while reasoning has one unfinished source answer, five expanded-reference answers, five extracted-evidence answers and six checked-evidence answers. Checking raises exact agreement from 4,620 to 5,149 of 5,779 paths and body fields, with 630 unavailable fields remaining. It corrects six answers and introduces two errors under greedy answering. The corresponding reasoning counts are six and nine.

Forty-five tables have multiple questions. A fixed identifier hash chooses an initial question without exposing that choice to extraction. Under reasoning, extracted and exact-reference evidence each answer 32 initial questions correctly. Twelve extracted-evidence tables and ten reference tables then have at least one later error. Across all 65 tables, extraction answers every question correctly for 33, compared with 39 for the source hierarchy and 35 for the exact expansion. Allowing answer-list reordering changes at most two correct-answer counts per condition.

C Complementary Field Allocation Studies

C.1 What Additional Records Contribute to Maize Selection

The maize reanalysis uses a public redistribution of the curated Genomes-to-Fields archive, with source identities matched to the original release [11, 86, 87]. Selection is chronological. Each target trial uses only earlier-year outcomes, and the location-exclusion design also removes all earlier observations from the target location. Yield, weather, flowering and harvest measurements from the target season are unavailable to prediction.

Historical means rank hybrids already observed in training. Trial adjustment fits an overall mean, an unpenalized trial intercept and a penalized hybrid effect. Parental models add categorical effects for the tester position or both recorded parent positions. Unknown effects contribute zero. Hybrid and parent penalties are selected from $\{1, 10, 50\}$ using the latest two eligible earlier years for validation. The criterion is prediction error after centering predictions and observations within each validation trial. The future trial intercept is omitted from within-trial ranking.

The untested-cross cohort contains hybrids absent from training but sharing at least one previously observed parent in the same recorded position. Requiring ten eligible candidates gives 81 trials at 31 locations over 2016–2020. Parent penalties are inherited from validation on historically observed hybrids. Every method selects five candidates within a trial. A tie at the boundary divides the remaining slots equally among tied candidates.

If w_i is candidate i’s inclusion probability and y_i its observed yield, selected yield is $\sum_i w_i y_i / \sum_i w_i$. Uniform allocation gives every candidate the same inclusion probability. Trial means receive equal weight in the primary summary. Parent links raise the share of hybrid–trial candidates with usable records for the hybrid itself or its parents from 35 percent to 86 percent in these trials.

With equally weighted trials, the both-parent effects rule selects 9.766 Mg ha^{-1}, a gain of 0.304 over uniform allocation. Parental means select 9.695 (gain 0.233), tester effects 9.664 (gain 0.202) and parent occurrence counts 9.634 (gain 0.173). Averaging trials within each of the 31 locations first, the location-level gain of the both-parent rule is 0.288 $\pm$ 0.499 Mg ha^{-1} (mean $\pm$ sample standard deviation) and ranges from -0.482 to $+1.346$. The corresponding location means are 0.186 $\pm$ 0.557 for parental means, 0.209 $\pm$ 0.483 for tester effects and 0.157 $\pm$ 0.439 for occurrence counts, with location ranges from -1.312 to $+1.512$, -0.644 to $+1.534$ and -0.895 to $+1.082$. The gain from the both-parent rule over uniform allocation remains positive in each of the five evaluated seasons. Omitting one season at a time gives mean gains from 0.259 to 0.374 Mg ha^{-1}.

C.2 A Later Season and a Matched Marker Comparison

The official 2023 release supplies a later-season evaluation [85]. Eligibility requires the main hybrid experiment, explicit non-irrigated status, a two-parent pedigree and positive observed yield. Discarded plots and fillers are excluded. Unknown irrigation status remains excluded. Replicate yields are averaged within cross and field location. The documented grain-yield unit is converted to Mg ha^{-1} on the stated 15.5-percent moisture basis.

Training uses 2014–2021 records from other location codes. Eligible crosses have no records of their own in that training set and share at least one position-matched parent. Fixed penalties of 10 for hybrid effects and 50 for parental effects follow the historical validation choices. The quota remains five. Unpenalized parental means are the primary later-season comparator.

The sequential source exclusions leave 7,762 plots. Of these, 498 represent previously observed crosses and 5,887 lack a position-matched historical parent. The remaining 1,377 plots yield 989 cross–trial pairs, comprising 78 crosses in 14 eligible trials. Fitted parental effects select a mean of 10.00 Mg ha^{-1}, compared with 10.03 for unpenalized parental means and 9.48 for uniform allocation. The simple parental mean is a competitive observed baseline in the new season.

A subsequent competition marker release covers 75 of those 78 crosses [88]. Exact cross matching excludes three identifiers and 34 cross–trial pairs. Every genomic and parental comparison uses the remaining 955 pairs at all 14 locations. Reverse-order crosses are not substituted. Marker profiles preserve upstream parental imputation and hybrid-construction assumptions.

An outcome-independent panel selects the first passing biallelic single-nucleotide variant in each 100-kb chromosome window, giving 14,982 sites. Training hybrids determine marker filtering, mean imputation and centering. Filters require at least 95 percent called genotypes, minor allele frequency of at least 0.01 and positive dosage variance. Final fits exclude target-location phenotypes.

Genomic ridge predicts from the linear marker relationship matrix. Training targets are hybrid means of trial-centered yield, weighted by their plot counts. The penalty is selected from $\{0.1, 1, 10, 100\}$ on chronological 2018–2019 development trials and is fixed before target prediction. The chosen penalty is 100. Gaussian-kernel regression, marker gradient boosting, a two-layer feed-forward network and an average of ridge and boosting use the same training records, candidate pool and allocation budget. Boosting and the network use 200 training-derived marker principal components. Their settings are selected on the same development criterion.

Table A11 reports all evaluated families. Selected-yield means weight the 14 locations equally. The location range concerns paired differences from parental effects, while Positive counts locations with a positive difference. The marker-model ordering describes this common candidate pool. The tester-effects and uniform-allocation rows are identical because tester effects give every candidate of a trial the same score in this season, and the tie rule then divides the five slots equally, which is uniform allocation.

C.3 Dated Measurements for Potato Scouting

The potato dataset supplies georeferenced five-band aerial images, chlorophyll readings, gas-exchange measurements and expert early-blight assessments in organic and conventional fields [97]. Of 256 measured plants, 254 have an explicit final assessment. Two missing original assessments remain excluded even where derived fields label them healthy.

Each plant is matched to the ten nearest pixel centers within each reflectance band's spatial reference. Their average supplies the band value. The largest plant-to-pixel-center distance is 5.34 cm. The bands yield normalized difference vegetation, normalized difference red-edge, green normalized difference vegetation and plant senescence reflectance indices. Physiological predictors include the recorded chlorophyll reading, stomatal conductance, apparent transpiration, photosystem-II quantum yield and electron transport rate, averaged by plant and occasion.

Table A11. Later-season maize selection on 955 marker-covered cross–trial pairs. Yield is mean ± sample standard deviation across 14 locations, in Mg ha^{-1}. Difference and its location range compare with parental effects. Positive counts positive location differences.

Rule	Selected yield	Difference	Location range	Positive
Genomic ridge	10.84 ± 2.45	+0.845	-0.119 to +1.935	12
Ridge–boosting ensemble	10.67 ± 2.33	+0.672	-0.875 to +1.548	12
Marker gradient boosting	10.65 ± 2.47	+0.660	-0.002 to +1.643	13
Gaussian-kernel regression	10.35 ± 2.38	+0.354	-1.802 to +1.195	13
Marker neural network	10.37 ± 2.13	+0.379	-0.576 to +1.154	10
Parental means	10.03 ± 2.62	+0.039	-1.036 to +1.869	7
Parental effects	10.00 ± 2.22	+0.000	+0.000 to +0.000	0
Parent occurrence counts	9.62 ± 2.11	-0.377	-2.126 to +0.771	5
Tester effects	9.44 ± 2.07	-0.551	-1.669 to -0.015	0
Uniform allocation	9.44 ± 2.07	-0.551	-1.669 to -0.015	0

Instrument timestamps extend the nominal June and July surveys through 16 June and 14 July. Those dates determine temporal eligibility. Every predictor precedes the 20 July outcome by at least six days. Final-date measurements are excluded. Negative apparent gas-exchange readings remain recorded, and chlorophyll readings remain on their source instrument scale.

Six feature families use field and variety alone, latest aerial measurements, aerial history, latest physiology, physiological history or combined history. History includes both early occasions and their difference. Training-only preprocessing performs median imputation, numerical scaling and categorical encoding. Inner validation selects ridge regression with penalty 1, 10 or 100, or an extremely randomized forest with 128 trees, minimum leaf size 4 or 12 and feature fraction 0.8.

Outer validation excludes a variety, a block across both fields, a field, or both the target variety and block. The designs have eight, four, two and 32 folds. Inner folds group by variety for variety transfer and by block for block and field transfer. Combined history under variety exclusion is the primary comparison, with latest aerial measurements and physiological history as its specified comparators.

Each held-out fold and field receives its own inspection budget. Let β be the inspected fraction, ζ_i the expert-assessed severity of plant i, and $w_i(\beta)$ its inclusion probability under the ranking. Severity capture is

$$C(\beta) = \frac{\sum_i w_i(\beta)\zeta_i}{\sum_i \zeta_i}. \tag{A3}$$

Fractional allocation in Equation A3 handles noninteger budgets and boundary ties. The primary fraction is 25 percent, with 12.5 and 50 percent as sensitivities. Uniform allocation has expected capture β. Hindsight ranks observed severity within the same pools.

Table A12 reports all feature families and budgets under the four hold-out designs. Holding out a variety restricts each allocation pool to that variety. Holding out a block permits ranking familiar varieties against one another. The different candidate pools explain why the observed capture levels should be interpreted within each design.

At the quarter-budget, combined history and latest aerial measurements both capture 26.9 percent of severity under variety exclusion. Their paired capture difference ranges from approximately -16 to $+3$ percentage points across varieties. Pooled parity conceals this variation. Under block exclusion, field and variety alone capture 75.6 percent, compared with 76.7 percent for combined history. Forty-one of 64 microplots assign the same disease score to every assessed plant, limiting the within-microplot variation available for ranking.

The model inputs and folds were fixed before fitting. The allocation score ranks predictions within each fold and field. Plants within a microplot are subsamples, and the two fields are not replicated management treatments. These outcomes describe retrospective allocation against the available severity assessments.

Table A12. Potato scouting under four hold-out designs. The first two blocks hold out unfamiliar varieties, excluding the variety alone or the variety and its block, and the last two rank familiar varieties while excluding a block or a whole field. Capture is the percentage of summed severity inspected at each budget. MAE and RMSE are percentage points of affected area. Each management system occupies one field.

Evidence	12.5%	25%	50%	MAE	RMSE
Held-out variety					
Field and variety	12.5	25.0	50.0	5.09	9.14
Latest UAV	15.0	26.9	53.0	3.72	8.64
UAV history	16.6	27.7	50.3	3.51	8.47
Latest physiology	20.2	35.3	60.9	3.68	8.67
Physiology history	17.2	32.9	55.6	3.66	8.72
Combined history	14.6	26.9	55.8	3.53	8.54
Held-out variety and block					
Field and variety	12.5	25.0	50.0	5.15	9.19
Latest UAV	12.4	24.8	48.5	3.82	8.70
UAV history	12.6	25.2	48.0	3.67	8.59
Latest physiology	13.6	27.3	51.0	4.14	8.94
Physiology history	12.7	25.4	52.7	4.06	8.91
Combined history	12.8	25.6	48.4	3.96	8.84
Held-out block					
Field and variety	62.3	75.6	89.4	3.07	6.53
Latest UAV	60.2	72.1	89.8	3.00	6.50
UAV history	62.1	76.1	90.2	2.74	6.29
Latest physiology	62.4	76.0	88.8	3.31	7.26
Physiology history	60.2	75.6	89.4	3.28	7.33
Combined history	62.1	76.7	90.3	2.76	6.33
Held-out field					
Field and variety	62.3	68.5	86.6	3.96	8.20
Latest UAV	58.4	67.9	84.9	3.92	7.96
UAV history	61.7	68.9	86.8	4.37	8.22
Latest physiology	59.3	68.1	86.1	3.96	7.87
Physiology history	53.8	68.4	87.8	3.97	7.99
Combined history	62.7	71.4	87.4	3.86	7.33

D Daily Crop-Demand Interpretation

The water and nutrient balances in Section 7 express conservation over a defined area and time interval. Parameter estimation additionally needs soil properties, crop development, observations and management history. A seasonal prescription constrains application and leaves several daily uptake trajectories possible.

The official cultivation guidance for Huaxingdan 88DT specifies nitrogen applications of 6–7.5 kg per mu at sowing, 6–7.5 kg per mu at the five-leaf stage and 10–12 kg per mu at the late whorl stage [82]. With 15 mu per hectare, these become 90–112.5, 90–112.5 and 150–180 kg N ha^{-1}. Their combined range is 330–405 kg N ha^{-1}. The source expresses phosphorus as P_2O_5 and potassium as K_2O, requiring conversion before an elemental nutrient-pool calculation.

A phenology model supplies the relationship between stage and date. Soil supply, crop uptake and loss determine the daily available pool. Several daily trajectories can satisfy the same stage totals. Interpretable inversion needs observations and an observation model that distinguish the relevant states. WOFOST and DSSAT supply established process structures, with the selected model's nutrient representation stated explicitly [58, 102]. The equations here specify a proposed coupling, not fitted water- or nutrient-demand performance.

E Experimental Generation Prompts

The fixed instruction wording below is reproduced from the executed experiments. Angle-bracketed slots stand for record-specific inputs. They are replaced by the original source passage, serialized profiles, question, options or saved description before generation. Chinese instructions are preserved in their original language. The public benchmark conditions and output limits are specified in Sections 5 and B.4.

Every evaluated record uses the same fixed instruction for its experimental condition.

E.1 Agricultural Case Description

System message.

```
You describe agricultural case evidence for later use.
```

User message.

```
Describe this agricultural case in at most 400 words so that a later reader can use its evidence. You have the reported
↪  background and numbered images, but no downstream question.

Preserve what is visibly observable. Include distinctive plant, pest, organ, symptom, shape, color, distribution and
↪  environmental features when they can be seen. Keep observations linked to image identifiers. Explain relationships
↪  between views when the images support them. Separate reported background from visual observations and from uncertain
↪  interpretations. State uncertainty when identification or a relationship is unclear. Do not invent unseen features or
↪  treatments. A potential identification may be useful, but it must remain distinguishable from the observations
↪  supporting it.

Write the case description only.

Reported background
<BACKGROUND>
```

The user message then supplies each image as a visual input preceded by its literal label, Image I1, Image I2, and so forth. Every listed image is included.

E.2 Agricultural Multiple-Choice Answering

System message.

```
Answer the agricultural multiple-choice question using the supplied case evidence and your agricultural knowledge.
```

User message.

```
Reported background
<BACKGROUND>

Machine-generated case description
<SAVED DESCRIPTION>

Question
<QUESTION>

Options
A. <OPTION A>
B. <OPTION B>
C. <OPTION C>
D. <OPTION D>

Choose exactly one of the four options. Return only a JSON object with one field "answer" containing its letter A, B, C
↪  or D.
```

The description heading and its content are omitted in the background and image-only conditions. Image conditions append the original numbered visual inputs. The crossed comparison changes the saved

description producer without changing the instruction. The answer schema allows only the field answer with one of A, B, C or D.

E.3 Hierarchical-Table Extraction

System message.

```
Extract source-linked agricultural table evidence for later questions.
```

User message.

```
Convert this hierarchical agricultural statistics table into reusable qualified data. You have no downstream question.
↪  Preserve every body value and its full row and column header ancestry.

Coordinates are zero-based positions in texts. The first top_header_rows_num rows and first left_header_columns_num
↪  columns are headers. Use top_root for column ancestry and left_root for row ancestry. Some internal hierarchy nodes
↪  also identify a data row or column. At each body row/column index, use the deepest node at that index and retain its
↪  complete path from the first real node. A virtual root at (-1,-1) is omitted.

Return one JSON object with row_paths, column_paths and cells. Each path item has index and path. A path is an ordered
↪  list of [row_index,column_index,exact_header_text] triples, from ancestor to the indexed node. Every index and row
↪  field must use the original texts matrix coordinate, including the header offset. Do not restart the data region at
↪  zero. For row_paths, index equals the row_index in the final path triple. For column_paths, index equals the
↪  column_index in the final path triple. In cells, row is the same original row index as the matching row_paths item.

Include every body row in row_paths and cells and every body column in column_paths. Each cells item has row and values,
↪  with all body-column values in their original order. Copy strings exactly, including blanks, symbols, units and
↪  punctuation. Preserve zero values. Do not round, summarise or infer missing values. Return only the requested JSON.

Source table
<SERIALIZED SOURCE TABLE>
```

E.4 Agricultural Table Answering

System message.

```
Answer the agricultural statistics question using the supplied evidence.
```

User message.

```
Evidence
<SERIALIZED EVIDENCE>

Question
<QUESTION>

Answer the question from the supplied agricultural statistics evidence. The evidence may be a hierarchical table or an
↪  expansion into row paths, column paths and body values. Coordinates are zero-based original table positions. Every
↪  header path runs from ancestor to the indexed node. A null path or value is unavailable evidence. Do not invent an
↪  unavailable value.

Return only a JSON object with one field "answer", a nonempty list of strings. Put a numeric answer in the units
↪  requested by the question. Do not include explanations or units that the question does not ask to output. For
↪  multiple answers, use original table order unless the question specifies another order. If the supplied evidence does
↪  not determine an answer, use ["unknown"].
```

The extraction schema requires row paths, column paths and cells as described in the instruction. The answering schema requires a nonempty list of strings in the answer field. Greedy and reasoning conditions use identical prompt wording.

E.5 Cultivar Requests and Common Readers

The following messages are rendered from the prompt builders used for Tables A4 and A5. The question states the intended zone, endpoint, maturity ceiling, disease list and grade cutoff. In the question template

below, the grade phrase is “为中抗或更好” or “不为高感”, preceded by “均” for multiple diseases. The context contains one serialized object per anonymous profile.

```
种植者的要求：在“<ZONE>”条件下种植，<ENDPOINT>不超过<DAYS>天，并且对<DISEASE LIST>的抗性<GRADE PHRASE>。
↪  抗性等级从好到差依次为：高抗、抗、中抗、中感、感、高感。某病害没有抗性记录时视为不满足。
问题：该品种是否满足这些要求？
```

Anonymous-profile reader. *System message.*

```
你是农作物品种推荐助手，依据给定证据判断品种是否满足种植者的要求。
```

User message.

```
品种：<CULTIVAR>（<CROP>，审定编号<REGISTRATION>）
以下保留该品种的全部生产条件记录，生产条件名称未记录。各编号仅区分记录，不能据编号判断生产条件。
同一编号内的生育期和抗性来自同一条生产条件记录。请求的生产条件属于这些记录之一。
抗性等级编码：0= 高抗，1= 抗，2= 中抗，3= 中感，4= 感，5= 高感。一个病害列出多个等级时，以最差等级为准。
<SERIALIZED PROFILE; REPEATED ONCE PER PROFILE>

<QUESTION>
请简要说明判断依据，最后输出 JSON：{"answer": " 是"} 或 {"answer": " 否"}。
```

Source reader. *System message.*

```
你是农作物品种推荐助手，依据给定证据判断品种是否满足种植者的要求。
```

User message.

```
品种：<CULTIVAR>（<CROP>，审定编号<REGISTRATION>）
审定公告中的品种特性原文：
<SOURCE PASSAGE>

<QUESTION>
请简要说明判断依据，最后输出 JSON：{"answer": " 是"} 或 {"answer": " 否"}。
```

Verbalized confidence. *User message.*

```
品种：<CULTIVAR>（<CROP>，审定编号<REGISTRATION>）
以下保留该品种的全部生产条件记录，生产条件名称未记录。各编号仅区分记录，不能据编号判断生产条件。
同一编号内的生育期和抗性来自同一条生产条件记录。请求的生产条件属于这些记录之一。
抗性等级编码：0= 高抗，1= 抗，2= 中抗，3= 中感，4= 感，5= 高感。一个病害列出多个等级时，以最差等级为准。
<SERIALIZED PROFILE; REPEATED ONCE PER PROFILE>

<QUESTION>

已有回答：<ACTUAL GROUPED ANSWER>
请估计上述回答正确的概率，范围为 0 至 1。只输出 JSON：{"confidence": 概率}。
```

Sufficient Context autorater. *User message.*

```
You are an expert LLM evaluator that excels at evaluating a QUESTION and REFERENCES.
Consider the following criteria:
Sufficient Context: 1 IF the CONTEXT is sufficient to infer the answer to the question and 0
IF the CONTEXT cannot be used to infer the answer to the question
Assume the queries have timestamp 2025-12-31.
First, output a list of step-by-step questions that would be used to arrive at a label for the
criteria. Make sure to include questions about assumptions implicit in the QUESTION.
Include questions about any mathematical calculations or arithmetic that would be required.
Next, answer each of the questions. Make sure to work step by step through any required
mathematical calculations or arithmetic. Finally, use these answers to evaluate the criteria.
Output the ### EXPLANATION (Text). Then, use the EXPLANATION to output the ###
EVALUATION (JSON)
EXAMPLE:
### QUESTION
In which year did the publisher of Roald Dahl’s Guide to Railway Safety cease to exist?
```

```
### References
Roald Dahl's Guide to Railway Safety was published in 1991 by the British Railways Board.
The British Railways Board had asked Roald Dahl to write the text of the booklet, and
Quentin Blake to illustrate it, to help young people enjoy using the railways safely. The
British Railways Board (BRB) was a nationalised industry in the United Kingdom that
operated from 1963 to 2001. Until 1997 it was responsible for most railway services in Great
Britain, trading under the brand name British Railways and, from 1965, British Rail. It
did not operate railways in Northern Ireland, where railways were the responsibility of the
Government of Northern Ireland.
### EXPLANATION
The context mentions that Roald Dahl's Guide to Railway Safety was published by the
British Railways Board. It also states that the British Railways Board operated from 1963 to
2001, meaning the year it ceased to exist was 2001. Therefore, the context does provide a
precise answer to the question.
### JSON
{"Sufficient Context": 1}
Remember the instructions: You are an expert LLM evaluator that excels at evaluating a
QUESTION and REFERENCES. Consider the following criteria:
Sufficient Context: 1 IF the CONTEXT is sufficient to infer the answer to the question and 0
IF the CONTEXT cannot be used to infer the answer to the question
Assume the queries have timestamp 2025-12-31.
First, output a list of step-by-step questions that would be used to arrive at a label for the
criteria. Make sure to include questions about assumptions implicit in the QUESTION
Include questions about any mathematical calculations or arithmetic that would be required.
Next, answer each of the questions. Make sure to work step by step through any required
mathematical calculations or arithmetic. Finally, use these answers to evaluate the criteria.
Output the ### EXPLANATION (Text). Then, use the EXPLANATION to output the ###
EVALUATION (JSON)
### QUESTION
<QUESTION>
### REFERENCES
品种：<CULTIVAR>（<CROP>，审定编号<REGISTRATION>）
以下保留该品种的全部生产条件记录，生产条件名称未记录。各编号仅区分记录，不能据编号判断生产条件。
同一编号内的生育期和抗性来自同一条生产条件记录。请求的生产条件属于这些记录之一。
抗性等级编码：0= 高抗，1= 抗，2= 中抗，3= 中感，4= 感，5= 高感。一个病害列出多个等级时，以最差等级为准。
<SERIALIZED PROFILE; REPEATED ONCE PER PROFILE>
```

E.6 Verification and Source Retry

The verifier receives the same evidence and the actual candidate response. Its checks apply to both the initial profile answer and the recovered source answer.

```
Evaluate the candidate answer using the QUESTION and EVIDENCE. Check three criteria independently. RELEVANCE: the
↪  evidence concerns the requested subject and requirements. GROUNDING: the answer follows from the evidence without
↪  unsupported assumptions. ADEQUACY: the answer resolves the original question and all its constraints. A negative
↪  answer can be valid; use the missing-grade convention stated in the question. Explain briefly, then output a final
↪  JSON object with integer labels 0 or 1 for relevance, grounding, adequacy. If any check fails, also give a
↪  failure-aware rewritten question in rewrite: clarify the entity for relevance, request evidence-grounded reasoning
↪  for grounding, or restore missing constraints for adequacy. Preserve the original crop, production system, diseases,
↪  grade threshold and duration limit. Use an empty rewrite if all checks pass.
QUESTION:
<QUESTION>
EVIDENCE:
<EVIDENCE>
CANDIDATE ANSWER:
<CANDIDATE ANSWER>
Final JSON format: {"relevance":1,"grounding":1,"adequacy":1,"rewrite":""}
```

The retry uses the common reader system message. Its user message is the source context above, followed by the text below. If the verifier output cannot be parsed, the refinement is the literal sentence "Re-evaluate the original question using only the supplied evidence and resolve every constraint."

```
Original question:
<QUESTION>
```

```
Verification feedback and refinement:
<VERIFIER REWRITE>
The original requirements are binding. Explain briefly and finish with JSON {"answer":" 是"} or {"answer":" 否"}.
```

MBA-RAG trains a routing encoder and uses the same readers after selection, without introducing an additional generation prompt.

F Data, Code and Experimental Settings

F.1 Material Available for Verification

The reproducibility materials available on request include the parsed approval profiles and source spans, enumerated requests and documentary answers, cross-validation fold assignments, saved model responses and token counts, policy predictions, and analysis code for Tables A3–A6. They also include the derived claims used for cross-zone prediction and the script for the numerical management example in Section 7.4. Official documents are identified by announcement number and version in the bibliography, model checkpoints are specified below, and public image, case, table and field datasets are cited in Appendices B and C under their original licenses.

F.2 Documentary Workload and Bounds

Table A3 uses all 157,522 enumerated candidate–request combinations in the six specified type zones. The maturity ceilings are the observed duration breakpoints, and the two disease cutoffs accept codes 0–2 and 0–4. Source parsing, condition evaluation and the four association ablations are deterministic.

Tables A4 and A5 use a fixed stratified sample of 1,000 requests from 229 registrations, comprising 400 coarse-bound disagreements, 300 other multi-profile requests and 300 single-profile requests. Each request carries its stratum expansion weight, and each reported metric divides its weighted sum by the total weight of the enumerated workload. Five cross-validation folds are assigned by a SHA-256 hash of the normalized cultivar name with trailing ASCII suffixes removed so that each cultivar family remains in a single fold. For test fold f, fold $(f+1) \bmod 5$ is reserved for calibration and the remaining three folds are used for fitting.

F.3 Readers and Learned Routing

The common reader is `Qwen/Qwen3.8-27B` (checkpoint revision `1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0`) run in bfloat16 precision with no adapter [72], temperature 0, generation seed 20260927 and native thinking disabled. Output limits are 2,048 tokens for the reader, 4,096 tokens for the Sufficient Context autorater and 256 tokens for verbalized confidence. Each invoked call contributes its recorded input and output tokens, excluding shared profile construction, encoder inference and router training.

The Sufficient Context adaptation uses the one-shot autorater prompt from commit `801b465eb3be5bf5216f be3301e61daee85ff99c` over the request and anonymous profile values. The two-signal variant fits a logistic regression combiner ($C = 1$, L-BFGS, at most 1,000 iterations, seed 20260927, with fit weights normalized to mean one) to the autorater decision and the reader's reported confidence, routing invalid signals to the source. Calibration selects thresholds for access targets of 0, 1.2, 5.2, 10, 30, 50 and 100% after accounting for mandatory accesses while keeping tied-score groups together.

The MBA-RAG adaptation uses the selected-action squared-error loss from commit `1779c029dfd8a0056f e15cc4184e13964ec8312e` with `google-bert/bert-base-chinese` (revision `8f23c25b06e129b6c986331a13d8d025 a92cf0ea`). The router receives the cultivar identity and request (256 tokens) and selects between the anonymous-profile reader and the source reader. Each fit runs for one epoch with batch size 1, gradient accumulation over 16 examples, AdamW (learning rate 10^{-5}, weight decay 0.01, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$) and a linear schedule without warm-up, using seed $20260927 + f$ on the three training folds of test fold f. At training step s, exploration chooses a random action with probability $\min\{1, 1/\log(s + 10^{-5})\}$ and otherwise selects the higher-scoring action. The final partial accumulation also produces an optimizer update. The target reward is 1 for a correct profile answer, $1 - \lambda$ for a correct source answer and 0 for an incorrect answer under workload weights normalized within the fit folds, evaluated at action penalties $\lambda \in \{0.01, 0.1, 0.3\}$.

The A2RAG-style controller checks relevance, grounding and answer adequacy with the common model at temperature 0, seed 20260927 and a 2,048-token output limit. When verification triggers source access, the controller reads the registration's approval passage and allows at most two source-answer attempts (an initial source answer and one refined-question retry), returning an abstention if verification remains unresolved.

Table A5 combines the saved reader outputs with multi-profile or profile-agreement routing and with model reading or direct condition execution. Registration-level differences are averaged with request weights within each registration before computing the mean and sample standard deviation across the 229 registrations, whereas the overall token difference weights the full request workload.

F.4 Cross-Zone Prediction

Table A6 uses five cultivar-grouped fold assignments generated with seeds 0, 1, 2, 3 and 4. For each seed, a NumPy random generator permutes the integers from 0 to $N_{\text{cult}} - 1$ and assigns them modulo 5 in order to the sorted cultivar names, where N_{cult} is the number of distinct cultivars. Numeric features and one-hot crop, disease and zone codes follow Appendix A.4, with numeric scaling for logistic regression fitted on training observations only and repeated point grades in a cultivar–disease–zone cell summarized by their median rounded to the nearest integer (ties to even).

Logistic regression uses L2 regularization ($C = 0.3$), L-BFGS, tolerance 10^{-4} and at most 5,000 iterations. The random forest uses 500 trees, Gini splits, bootstrap sampling, square-root feature subsampling, unrestricted depth, a minimum leaf size of 2 and a minimum split size of 2. Histogram gradient boosting uses log loss, learning rate 0.05, 200 boosting iterations, at most 31 leaves per tree, at least 20 observations per leaf, 255 bins and zero L2 regularization. Automatic early stopping is inactive because the training sets have fewer than 10,000 rows. Both tree ensembles use the fold-assignment seed. Means and sample standard deviations are computed across the five seeds, and the exact two-sided sign tests are evaluated on the per-cultivar accuracies.

F.5 Software and Hardware

For Tables A3–A6, data processing and model training used Python 3.12.3, NumPy 2.4.3, pandas 3.0.3, scikit-learn 1.9.0, PyTorch 2.12.0 and Transformers 5.10.2. Common-reader inference used vLLM 0.27.1 with PyTorch 2.13.0+cu132 and Transformers 5.15.1 (8,192-token context and batched-token limits, up to 32 concurrent sequences, and 0.72 GPU-memory fraction; MBA-RAG router training used one GPU and four CPU threads). All experiments ran on a single server equipped with two Intel Xeon Platinum 8260 processors at 2.40 GHz (48 physical cores, 96 logical CPUs), 503 GiB RAM and four NVIDIA A100-SXM4 GPUs (80 GiB each).



Yusen Yang received the M.Sc. degree in computer engineering from the College of Design and Engineering, National University of Singapore (NUS), Singapore. He is an Engineer with the National Engineering Research Center for Information Technology in Agriculture (NERCITA) and the Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences (BAAFS). His research interests include smart agriculture, multi-agent reinforcement learning (MARL), agricultural foundation models, and personalized water and fertilizer management decision support systems. (Email: workpsn202610@126.com)



Youlin Cui received his education in agricultural engineering and agronomy. He is a Senior Agronomist and researcher with the Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences (BAAFS) and the National Engineering Research Center for Information Technology in Agriculture (NERCITA). His research focuses on whole-process mechanization and intelligent equipment for open-field vegetable production, unmanned farm operation systems, intelligent harvesting robotics, and digital village technology applications. (Email: cuiyl007@nercita.org.cn)



Peng Hao received his education in computer science and agricultural information engineering. He is a Senior Engineer with the Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences (BAAFS) and the National Engineering Research Center for Information Technology in Agriculture (NERCITA). His research interests center on agricultural Internet of Things (IoT) networking, multi-view projection clustering and reconstruction of agricultural IoT data, opportunistic transmission coordination, and intelligent agricultural knowledge services. (Email: haop007@nercita.org.cn)



Huaji Zhu received the Ph.D. degree in cartography and geographic information systems from the Institute of Remote Sensing Applications, Chinese Academy of Sciences (CAS), Beijing, China, in 2006. He is a Research Professor and Master's Supervisor with the National Engineering Research Center for Information Technology in Agriculture (NERCITA) and the Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences (BAAFS), where he serves as Deputy Director of the Big Data Intelligence Department. His research focuses on agricultural big data governance, agricultural foundation models, intelligent knowledge cloud platforms, and precision management of crop growth environments. He has served as principal investigator for multiple national research projects. Corresponding author. (Email: zhuhuaji007@nercita.org.cn)