

ORIGINAL RESEARCH

Open access

Benchmarking Agricultural AI Without Field Experiments: A Conceptual Framework for Model Readiness, Data Quality, Deployment Risk, and Decision Validity

Olivia Bennett^{1*}, Harry Collins², Jack Foster¹, Ethan Wright³

Abstract

Agricultural artificial intelligence is commonly compared through benchmark datasets and technical metrics before any field experimentation occurs. Such comparisons can help identify promising models, reveal methodological weaknesses, and organize evidence for further testing, but they cannot establish field readiness, agronomic validity, operational safety, or deployment suitability. This article proposes a conceptual framework for benchmarking agricultural AI without field experiments. The framework separates five linked assessment domains: model readiness, technical validation, data quality, contextual fit, and deployment risk and decision validity. It addresses benchmark-task clarity, metric justification, dataset representativeness, provenance, label and ground-truth uncertainty, structured validation, domain shift, robustness, calibration, uncertainty, error patterns, bias, explainability, human oversight, and monitoring requirements. Non-empirical evaluation is treated as a structured screening process based on documentation, reported evidence, cross-context reasoning, and explicit claim boundaries. The framework produces a conditional benchmark-readiness statement and a field-validation gap register rather than a deployment decision. Its principal contribution is to distinguish benchmark performance from decision validity and deployment readiness while giving researchers, developers, evaluators, and agricultural stakeholders a transparent structure for comparing evidence before field testing. The framework is conceptual, does not replace field experiments, and should be updated as new data, validation evidence, user feedback, and operational observations become available.

Keywords Agricultural AI, Deployment risks, Benchmarking, Model readiness, Data quality, Decision validity

*Correspondence:

Olivia Bennett

olivia.bennett@nottingham.ac.uk

¹ Department of Agricultural AI Benchmarking and Readiness, Faculty of Agriculture, University of Nottingham, Nottingham, United Kingdom

² Department of Data Quality and Deployment Risk, Faculty of Agriculture, University of Southampton, Southampton, United Kingdom

³ Department of Model Validity and Decision Assessment, Faculty of Agriculture, Newcastle University, Newcastle, United Kingdom

Introduction

Agricultural AI now supports diverse activities across precision agriculture, smart farming, digital agriculture, crop monitoring, irrigation, soil assessment, livestock observation, and advisory systems. These applications depend on interconnected data flows, analytical pipelines, and organizational arrangements, while their machine-learning tasks differ substantially in inputs, targets, users, and consequences [1, 2]. Early comparison is therefore useful, but a benchmark result is meaningful only within the task, dataset, metric, and decision context that produced it.

Benchmark datasets and model scores offer a practical basis for comparing algorithms before field trials, yet they can encourage false equivalence between technical performance and agricultural usefulness. Deep-learning results are shaped by dataset design, sample composition, preprocessing, validation procedures, and reporting choices [3]. A model may perform well on a curated test set while remaining untested under different farms, seasons, sensors, crops, management practices, or operational constraints.

The challenge is also socio-technical. Digital agriculture affects users, advisory relationships, institutional responsibilities, and the distribution of benefits and burdens, none of which can be inferred from prediction metrics alone [4]. Field readiness therefore differs from model readiness, and deployment readiness differs from both. Technical validation may support a bounded claim about performance under specified conditions, but it does not establish farmer acceptance, agronomic correctness, safe operation, or responsible integration.

This article develops an original conceptual framework for benchmarking agricultural AI without field experiments. It organizes evidence around benchmark-task definition, model and data documentation, technical validation, representativeness, contextual fit, uncertainty, robustness, deployment risk, decision validity, human oversight, monitoring, and field-validation gaps. The framework is intended for early-stage comparison and risk screening. It does not certify systems, prescribe deployment, or replace local field validation.

Agricultural AI Benchmarking Challenges

Agricultural benchmarks often use datasets that are convenient, public, or easy to label rather than representative of intended use. In plant-disease recognition, effectiveness depends not only on sample volume but also on visual variety, symptom expression, background, capture conditions, and class composition [5, 6]. Consequently, benchmark performance may reflect dataset regularities rather than transferable biological or agronomic information. Task definitions can also be too narrow, converting complex decisions into simplified classification or prediction problems.

Dataset availability is uneven across precision-agriculture tasks, crops, regions, and acquisition settings [7]. This creates spatial, temporal, seasonal, regional, crop-context, and farm-context biases. A dataset may overrepresent particular production systems, growth stages, sensors, or environmental conditions while omitting small farms, uncommon cultivars, rare events, or low-connectivity settings. Large sample counts do not correct systematic absence, and a benchmark task can remain poorly matched to the target context.

Disease-recognition datasets further illustrate how class imbalance, annotation inconsistency, label uncertainty, and limited field representation can distort evaluation [8]. Ground truth may itself be uncertain when symptoms overlap, expert judgments differ, or measurements contain error. Aggregate metrics can conceal poor performance on rare but consequential classes. Benchmark designers must therefore justify metrics, report disaggregated errors, document annotation procedures, and distinguish reference uncertainty from model error.

Other threats include non-independent observations, leakage between training and testing, repeated tuning against public test sets, domain shift, and benchmark overfitting. Random splits may place images from the same field, season, plant, animal, or acquisition sequence in both development and evaluation sets. Even careful non-field tests cannot reveal all operational failures. The framework therefore requires explicit no-field-readiness and no-deployment-readiness boundaries whenever field evidence, workflow testing, monitoring, and stakeholder evaluation remain absent.

Model Readiness and Technical Validation

Model readiness begins with documentation. The intended task, prediction unit, time horizon, inputs, outputs, preprocessing, model version, training process, excluded uses, and known limitations must be stated before scores are compared. Internal validation should use protected data and prevent preprocessing or feature selection from accessing

evaluation information. Where data are spatially, temporally, or hierarchically structured, resampling must respect those dependencies because random cross-validation can produce misleading estimates [9].

External validation should use independently sourced data and explain how the external context differs from model-development conditions. Validation design should reflect the intended prediction target rather than methodological convenience [10]. For spatial applications, the evaluator must identify the quantity being estimated; a spatial cross-validation statistic is not automatically a valid estimate of map accuracy or transfer performance [11]. Metric selection must follow the evaluation question and the consequences of error.

Out-of-distribution testing, robustness checks, and generalization analysis extend validation beyond average performance. Evaluators should examine novel sensors, regions, seasons, crops, missing data, image quality, and plausible input disturbances. For broad spatial prediction, conventional validation may not reveal where predictor conditions fall outside training support [12]. Sensitivity analysis, and ablation analysis where appropriate, can clarify influential assumptions and model components, while structured error analysis can reveal systematic failure patterns hidden by aggregate metrics.

Uncertainty estimation and calibration should be assessed separately from discrimination or accuracy. Bias analysis should examine uneven coverage and errors across relevant agricultural contexts; fairness review is appropriate where benefits or burdens differ among stakeholders or farm types. Explainability should be evaluated only for a defined user and purpose, not treated as proof of correctness. Model readiness must therefore be reported as a bounded evidence profile, not as a guarantee of field validity or deployment suitability.

Table 1 summarises model readiness and technical validation criteria for benchmarking agricultural AI systems without field experiments.

Table 1. Model Readiness and Technical Validation Criteria for Benchmarking Agricultural AI Without Field Experiments: Model Documentation, Task Definition, Metric Justification, Internal Validation, External Validation, Out-of-Distribution Testing, Robustness, Generalization, Calibration, Uncertainty, Error Analysis, Bias, Fairness, Explainability, and Field-Validation Boundaries

Model readiness or validation component	Core readiness question	Required evidence or procedure	Technical validation function	Potential output	Failure or overclaiming risk	Additional assessment need	Frame ro
Model documentation	Is purpose and scope clear?	Versioned task, model, input, output, training, and limitation record	Enables reproducibility	Documentation profile	Hidden assumptions	Independent audit	Entry
Task definition	Does the task match the agricultural problem?	Target, unit, horizon, user, and exclusions	Establishes evaluation object	Task specification	Proxy-task mismatch	Decision-context review	Task
Metric justification	Do metrics reflect relevant errors?	Rationale and disaggregated results	Aligns metrics with questions	Metric adequacy statement	Aggregate scores conceal failure	Consequence review	Metric

Internal validation	Are unseen in-domain samples protected?	Isolated preprocessing, tuning, and testing	Detects overfitting	Internal evidence profile	Leakage	Reproducibility audit	Tech lay
Structured cross-validation	Are dependencies respected?	Grouped, spatial, temporal, or nested folds	Reduces optimistic estimates	Resampling profile	Non-IID leakage	External testing	Split-in che
External validation	Is evidence independently sourced?	Independent dataset and context comparison	Tests transfer	External-validity profile	External data remain too similar	Local validation	Transfe
Out-of-distribution testing	What happens beyond training support?	Held-out domains and novelty checks	Defines applicability limits	Out-of-distribution risk profile	Artificial shifts omit real change	Field monitoring	Bour che
Robustness and generalization	Is performance stable across plausible conditions?	Stress tests and subgroup or cross-domain results	Identifies fragility	Robustness profile	Narrow stress scenarios	Context-specific testing	Relia lay
Uncertainty and calibration	Are confidence estimates meaningful?	Uncertainty method and calibration evidence	Evaluates confidence quality	Uncertainty profile	False assurance	Target-context recalibration	Uncer lay
Sensitivity and ablation	Which assumptions or components matter?	Planned variations or component removal	Diagnoses dependencies	Contribution profile	Dataset-specific conclusions	Independent replication	Diagr lay
Error analysis	Which cases fail and why?	Case, class, subgroup, and residual review	Reveals systematic errors	Error taxonomy	Rare failures remain hidden	Agronomic review	Failure
Bias and fairness	Are errors or burdens uneven?	Context-relevant subgroup analysis	Screens disparities	Bias or fairness profile	Missing metadata	Stakeholder assessment	Respor lay
Explainability	Is explanation useful for a defined user?	Fidelity, stability, purpose, and limitation evidence	Supports oversight	Explainability profile	Plausibility mistaken for truth	User evaluation	Condi che

Claim boundaries	What remains unproven?	Gap register and explicit limitations	Prevents readiness inflation	Conditional readiness statement	Benchmark score becomes approval	Field validation	Fir bour
------------------	------------------------	---------------------------------------	------------------------------	---------------------------------	----------------------------------	------------------	-------------

Data Quality and Contextual Fit

Data quality assessment begins with documentation and provenance. Evaluators should record why data were collected, by whom, with which instruments or procedures, under what permissions, and through which transformations. Agricultural AI data management depends on integration, metadata, standards, validation, usability, and traceability [13]. Completeness must be assessed against the defined task because missing variables, periods, locations, or metadata can change both model performance and interpretation.

Representativeness requires more than overall dataset size. The benchmark should describe spatial, temporal, seasonal, regional, crop, cultivar, production-system, and farm-context coverage. Labels and annotations require protocols, quality controls, and, where relevant, agreement or adjudication records. Ground-truth uncertainty should be stated when reference measurements or expert diagnoses are imperfect. Class imbalance must be linked to per-class performance and error consequences rather than hidden within a single aggregate result.

Contextual fit asks whether the development evidence resembles the intended conditions of use. Mobile crop-disease classification under uncontrolled capture demonstrates why illumination, device, background, and symptom presentation matter when moving beyond curated images [14]. Domain shift should be examined explicitly when sensors, image domains, regions, or management conditions change, as transfer can degrade agricultural vision performance [15]. Non-IID structures must also be documented so that farms, fields, seasons, plants, animals, or image sequences are not split inappropriately.

Controlled-condition success does not independently establish field performance, particularly when development and field data differ materially [16]. The proposed framework therefore produces a contextual-fit profile and mismatch register rather than a guarantee. Data governance should also document access, stewardship, permitted use, retention, accountability, and change control. These checks can identify evidence gaps and support field-validation planning, but they cannot confirm local agronomic validity, farmer usability, operational reliability, or deployment readiness without direct evaluation.

Deployment Risk and Decision Validity

Agricultural AI deployment risk cannot be inferred from prediction error alone. System outputs may interact with ecological processes, production decisions, labor arrangements, data infrastructures, and institutional responsibilities, creating downstream effects not represented by benchmark metrics. Responsible evaluation must therefore examine systemic risks, ethical safeguards, accountability, and multiple dimensions of potential impact rather than reducing readiness to a single performance score [17-19]. A non-empirical benchmark can identify plausible hazards and unresolved responsibilities, but it cannot determine their frequency or severity under actual operating conditions.

Decision validity concerns whether a model output is relevant, timely, interpretable, and sufficiently supported for a particular agricultural decision. These conditions vary across farm systems, technological pathways, production structures, and stakeholder groups [20]. The proposed framework therefore requires decision-context mapping: who receives the output, what decision it informs, which alternatives are available, what evidence is missing, and what consequences follow from false, delayed, or misunderstood outputs. Predictive accuracy remains distinct from agronomic validity because a statistically accurate target may still be an incomplete or inappropriate basis for action.

Human mediation is especially important when AI outputs enter farmer-facing or adviser-facing decision-support systems. Agricultural advisers may combine digital outputs with local knowledge, observations, production objectives, and professional judgment, while farmers determine whether information is relevant to their operational circumstances [21]. Benchmarking should consequently document interpretation responsibilities, escalation routes, override authority, uncertainty communication, and conditions requiring independent verification. Human oversight must be substantive rather than nominal: the designated reviewer needs adequate information, competence, time, and authority to question or reject an output.

Deployment-risk screening should also define operational and safety boundaries, provisional risk classifications, monitoring needs, and updating requirements. Potential model drift and data drift should be anticipated when climates, seasons, sensors, crop varieties, management practices, or user populations change. Feedback mechanisms require controlled provenance and revalidation so that corrections do not introduce new biases. These criteria support early risk identification only; they do not guarantee decision validity, operational safety, or deployment readiness.

Non-Empirical Evaluation Criteria

Non-empirical evaluation begins with inspectable documentation. Benchmark-task descriptions should identify the target, unit of prediction, intended user, time horizon, exclusions, and relationship to the agricultural use case. Dataset documentation should explain motivation, composition, collection, preprocessing, intended uses, and limitations so that assumptions and potential inappropriate uses can be audited [22]. Model documentation should similarly record versions, dependencies, training procedures, evaluation conditions, uncertainty methods, and declared boundaries.

Evaluation integrity requires explicit examination of leakage, contamination, and dependence structures. Information may cross training and evaluation boundaries through preprocessing, repeated samples, grouped observations, feature construction, or repeated use of a public test set. Such leakage can inflate reported performance and weaken reproducibility [23]. A non-empirical assessment should therefore inspect split design, grouping variables, preprocessing isolation, tuning history, and test-set governance before interpreting any benchmark comparison.

Robustness, uncertainty, and external-validity evidence should be evaluated separately from average predictive performance. Uncertainty quantification can distinguish sources of limited confidence and support cautious interpretation, but uncertainty estimates may themselves be unreliable under distribution shift [24]. Reviewers should examine calibration, sensitivity, plausible perturbations, subgroup errors, out-of-distribution behavior, and the extent to which validation conditions resemble intended contexts. Missing evidence should be recorded rather than converted into a favourable readiness assumption.

Bias, fairness, and explainability reviews should be applied where relevant to the use case and available evidence. Explainability assessment should identify the intended user, explanatory purpose, method, fidelity, stability, and limitations because a plausible explanation does not necessarily describe the model faithfully or establish correctness [25]. Decision-consequence mapping, human-oversight planning, deployment-risk screening, and monitoring planning complete the non-empirical review. The resulting evidence profile identifies priorities for field validation; it does not replace field experimentation or authorize deployment.

Table 2 maps data-quality, contextual-fit, deployment-risk, decision-validity, and non-empirical evaluation criteria for agricultural AI benchmarking.

Table 2. Data Quality, Contextual Fit, Deployment Risk, Decision Validity, and Non-Empirical Evaluation Criteria for Agricultural AI Benchmarking: Representativeness, Provenance, Label Quality, Domain Shift, Contextual Fit, Risk Classification, Decision Consequence, Human Oversight, Monitoring, and Field-Experiment Boundaries

Evaluation criterion or risk component	Core screening question	Required evidence or documentation	Evaluation or risk-screening function	Potential output	Residual or failure risk	Final validation oversight
Data provenance and documentation	Are origins, transformations, permissions, and intended uses traceable?	Versioned dataset record and lineage	Supports reproducibility and auditability	Provenance profile	Undocumented upstream errors	Data-rev
Representativeness	Are intended crops, regions, seasons, farms, and conditions covered?	Source–target coverage comparison	Identifies sampling gaps	Representativeness matrix	Unknown target variation	Local v
Label and ground-truth quality	Are labels and reference measurements sufficiently reliable?	Protocols, agreement, adjudication, and measurement uncertainty	Separates model and reference error	Label-quality profile	Biological ambiguity remains	Agron
Class imbalance and subgroup bias	Are rare or underrepresented cases adequately evaluated?	Class counts and disaggregated errors	Detects concealed failure	Bias-risk profile	Missing metadata limits review	Target colle
Domain shift and non-IID structure	Do target distributions or dependencies differ from development data?	Domain comparisons and grouping information	Screens transfer and leakage risk	Shift-risk register	Unanticipated shifts remain	Exter field
Contextual fit	Do crop, sensor, farm, workflow, and decision contexts align?	Context profile and mismatch register	Integrates applicability evidence	Conditional-fit statement	Similarity may not yield transfer	Fa
Decision validity and agronomic validity	Is the output relevant and sufficient for the intended decision?	Decision pathway, rationale, alternatives, and missing evidence	Separates prediction from action	Decision-evidence gap register	Accurate outputs may remain unusable	Agron
Decision consequence and risk classification	What follows from error, delay, or misuse?	Severity, exposure, reversibility, and detectability scenarios	Prioritises further assessment	Provisional risk profile	Rare systemic consequences may be omitted	Multidis

Uncertainty and robustness evidence	Are confidence and stability evaluated under relevant conditions?	Calibration, stress tests, sensitivity, and limitations	Screens unreliable confidence	Reliability profile	Tested conditions remain incomplete	Context validity
Explainability and fairness	Are explanations and distributional effects appropriate to the use case?	User, purpose, fidelity, subgroup, and burden analysis	Supports responsible oversight	Conditional review profile	Persuasive explanations may be incorrect	Stakeholder evaluation
Human oversight	Who interprets, verifies, overrides, and escalates outputs?	Roles, competence, authority, and fallback procedures	Establishes accountable control	Oversight plan	Nominal oversight may fail	Workload tests
Monitoring and updating	How will drift, incidents, overrides, and feedback be handled?	Indicators, triggers, owners, and change-control process	Defines lifecycle evidence needs	Monitoring specification	Degradation may remain undetected	Long-term evaluation
Field-experiment boundary	Which important claims remain untested?	Field-validation gap register	Prevents non-field substitution	Explicit limitation statement	Evidence may still be overinterpreted	Field experiment results

Proposed Benchmarking Framework

The proposed framework begins with agricultural AI use-case definition, benchmark-task definition, and decision-context mapping. These stages specify the agricultural problem, target, prediction unit, time horizon, intended user, operational exclusions, and potential consequences. Responsibility must be assigned explicitly because trust and accountability depend on identifiable actors who remain answerable for interpreting, communicating, and acting on AI outputs [26]. The framework does not treat the model as an autonomous decision-maker.

The second stage reviews model and data documentation. Model review addresses architecture description, versioning, training procedures, preprocessing, intended use, excluded use, and limitations. Data review addresses provenance, composition, collection conditions, completeness, labels, reference uncertainty, and governance. Human-centered requirements are considered alongside technical evidence because usable oversight depends on interaction design, explanation, user control, and operational context [27]. Documentation gaps become explicit readiness limitations rather than assumed strengths.

The third stage assesses data quality, representativeness, technical validation, and external validity. Data-centric assessment is necessary because curation, labeling, cleaning, augmentation, metadata, and lifecycle management can influence reliability independently of model architecture [28]. Technical review examines split integrity, metric justification, structured cross-validation, external data, out-of-distribution tests, robustness, generalization, uncertainty, calibration, sensitivity, ablation where supported, and error patterns. Outputs remain multidimensional evidence profiles rather than a universal readiness score.

The fourth stage integrates bias, fairness, explainability, decision validity, and deployment risk. Benchmark readiness is communicated as a bounded technical judgment rather than evidence of wider agricultural benefit, inclusion, sustainability, or successful transformation [29]. The evaluator maps error consequences, agronomic evidence gaps, contextual mismatches, operational boundaries, safety boundaries, oversight responsibilities, and monitoring requirements. Critical unresolved risks cannot be averaged away by stronger performance in another domain.

The final stage produces a benchmark-readiness statement, field-validation gap register, human-oversight and monitoring plan, and framework-update schedule. Agriculture-specific explainability evidence may support review and communication, but it cannot independently prove model correctness, decision validity, or safe use [30]. Readiness statements must therefore identify supported contexts, unsupported contexts, uncertainty, unresolved risks, and evidence still required from field experiments, user evaluation, workflow testing, and longitudinal monitoring.

Table 3 presents the proposed benchmarking framework for agricultural AI model readiness, data quality, deployment risk, and decision validity without field experiments.

Table 3. Proposed Benchmarking Framework for Agricultural AI Without Field Experiments: Use-Case Definition, Benchmark Task Definition, Decision-Context Mapping, Model Documentation, Data Documentation, Data Quality, Representativeness, Technical Validation, External Validity, Uncertainty, Robustness, Bias, Explainability, Decision Validity, Deployment Risk, Human Oversight, Field-Validation Gaps, and Deployment Boundaries

Framework stage	Core purpose	Required input	Benchmarking function	Potential output	Validation or assessment need	Failure risk
Use-case and task definition	Establish agricultural purpose and evaluation object	User, target, horizon, context, and exclusions	Aligns benchmark with intended problem	Use-case and task specification	Agronomic and stakeholder review	Proxy-task mismatch
Decision-context mapping	Separate prediction from decision support	Decision owner, alternatives, timing, and consequences	Identifies evidence needed for action	Decision pathway	User and agronomic evaluation	Accuracy becomes advice
Model documentation review	Establish reproducibility and scope	Model version, training, inputs, outputs, and limitations	Exposes assumptions and dependencies	Documentation profile	Independent technical audit	Hidden development choices
Data documentation and quality	Establish traceability and fitness for task	Provenance, composition, labels, missingness, and governance	Identifies data defects	Data-readiness profile	Data-steward and domain review	Poor data are treated as adequate

Representativeness and contextual fit	Compare development and target conditions	Crop, region, season, farm, sensor, and workflow profiles	Identifies coverage and mismatch	Applicability matrix	Local contextual review	Unsupported transfer
Technical validation	Examine methodological soundness	Splits, metrics, tuning history, external data, and errors	Screens leakage, overfitting, and transfer	Validation profile	Reproduction and independent testing	Inflated benchmark performance
Uncertainty, calibration, and robustness	Assess confidence and stability	Calibration, stress, sensitivity, and domain evidence	Identifies unreliable confidence	Reliability profile	Target-context evaluation	False assurance
Bias, fairness, and explainability	Review responsible-performance conditions	Subgroup, user, explanation, and burden evidence	Supports oversight and disparity screening	Conditional responsibility profile	Missing groups or misleading explanations	Stakeholder evaluation
Decision-validity screening	Examine relevance and agronomic sufficiency	Decision map, uncertainty, rationale, and consequence analysis	Identifies missing decision evidence	Validity-gap register	Prediction is mistaken for valid action	Agronomic and user testing
Deployment-risk classification	Organise technical and socio-technical risks	Consolidated risk register	Prioritises field and workflow assessment	Provisional risk profile	Simplified classification hides hazards	Multidisciplinary review
Oversight and monitoring planning	Define accountable control and lifecycle observation	Roles, overrides, indicators, triggers, and incident procedures	Specifies safeguards and evidence needs	Oversight and monitoring plan	Workflow and longitudinal testing	Nominal controls fail
Field-validation gap identification	State what remains unknown	All evidence profiles and unresolved questions	Converts gaps into research needs	Field-validation agenda	Field-study design review	Gaps are understated
Benchmark-readiness statement	Communicate bounded pre-field evidence	Domain profiles, limits, and critical gaps	Summarises readiness without aggregation	Conditional narrative statement	Independent review	Statement becomes approval

Framework updating	Maintain validity over model and context changes	New data, versions, drift, incidents, and user feedback	Supports controlled revision	Updated framework record	Change-impact assessment	Prior findings are reused silently
--------------------	--	---	------------------------------	--------------------------	--------------------------	------------------------------------

Figure 1 presents the proposed benchmarking framework for agricultural AI without field experiments, linking use-case definition, benchmark task design, model readiness, data quality, contextual fit, deployment risk, decision validity, field-validation gaps, human oversight, and framework updating.

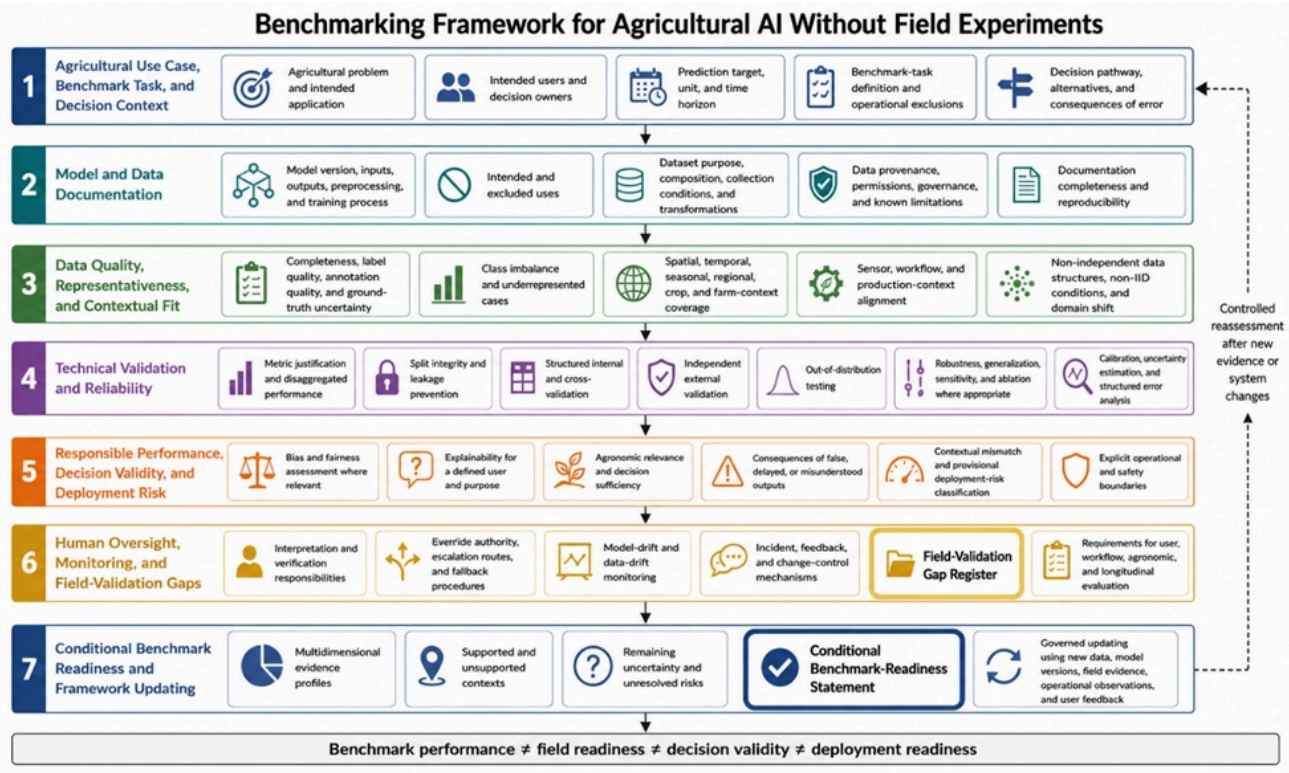


Figure 1. Benchmarking Framework for Agricultural AI Without Field Experiments

Alt-text description:

A conceptual benchmarking framework showing agricultural AI use-case definition, benchmark task design, model documentation, data quality assessment, technical validation, contextual fit, deployment risk, decision validity, human oversight, field-validation gaps, and benchmark readiness boundaries.

Practical Implications and Future Research

For researchers, developers, dataset curators, and benchmark designers, the framework supports clearer task definitions, documented data provenance, representativeness analysis, protected evaluation splits, disaggregated error reporting, and explicit uncertainty statements. Dataset documentation should be updated as collection, labeling, or permitted uses change [22]. Benchmark publications should identify unsupported contexts and explain what additional external, agronomic, and field evidence would be needed before stronger claims could be considered.

For platform providers, advisers, farmers, operators, technology evaluators, and journal reviewers, the framework provides questions for interpreting evidence rather than selecting products or prescribing actions. Particular attention

should be given to contextual mismatch, uncertainty communication, human authority, fallback procedures, and benchmark-to-field validity gaps. Future research should examine how uncertainty and calibration behave across crops, seasons, sensors, and user settings, while avoiding the assumption that technically quantified uncertainty automatically improves decisions [24].

Longitudinal research is needed on model drift, data drift, feedback updating, oversight effectiveness, and changes in performance after operational conditions evolve. Comparative field studies should investigate how benchmark evidence relates to agronomic outcomes, workflow integration, user interpretation, and unintended consequences. Governance research should examine responsibility, traceability, and accountability across developers, providers, advisers, and users [17]. These are conceptual implications and proposed research directions, not agronomic, legal, policy, procurement, certification, model-selection, platform-selection, or deployment guidance.

Conclusion

This article proposes a conceptual framework for benchmarking agricultural AI without field experiments by linking use-case definition, benchmark-task design, model readiness, technical validation, data quality, contextual fit, deployment risk, decision validity, non-empirical evaluation, human oversight, monitoring, field-validation gaps, and framework updating. Its central contribution is a disciplined separation between benchmark performance and field readiness. Agricultural AI evidence should be interpreted through task relevance, documentation, representativeness, uncertainty, generalization limits, decision context, risk, and unresolved validation needs rather than through technical scores alone. The framework can support early comparison and research planning, but it cannot establish agronomic correctness, operational safety, farmer acceptance, or deployment readiness. Such claims require contextual evaluation, field experimentation, human-centered assessment, and longitudinal monitoring.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 07 Sep 2025

Revised: 28 Oct 2025

Accepted: 14 Dec 2025

Published online: 25 May 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Wolfert S, Ge L, Verdouw C, Bogaardt MJ. Big data in smart farming—A review. *Agric Syst*. 2017;153:69-80.
<https://doi.org/10.1016/j.agry.2017.01.023>.
- Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: A review. *Sensors (Basel)*. 2018;18(8):2674.
<https://doi.org/10.3390/s18082674>.
- Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: A survey. *Comput Electron Agric*. 2018;147:70-90.
<https://doi.org/10.1016/j.compag.2018.02.016>.
- Klerkx L, Jakku E, Labarthe P. A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda. *NJAS Wageningen J Life Sci*. 2019;90-91:100315.
<https://doi.org/10.1016/j.njas.2019.100315>.
- Barbedo JGA. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Comput Electron Agric*. 2018;153:46-53.
<https://doi.org/10.1016/j.compag.2018.08.013>.
- Barbedo JGA. Factors influencing the use of deep learning for plant disease recognition. *Biosyst Eng*. 2018;172:84-91.
<https://doi.org/10.1016/j.biosystemseng.2018.05.013>.
- Lu Y, Young S. A survey of public datasets for computer vision tasks in precision agriculture. *Comput Electron Agric*. 2020;178:105760.
<https://doi.org/10.1016/j.compag.2020.105760>.
- Xu M, Park JE, Lee J, Yang J, Yoon S. Plant disease recognition datasets in the age of deep learning: Challenges and opportunities. *Front Plant Sci*. 2024;15:1452551.
<https://doi.org/10.3389/fpls.2024.1452551>.
- Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Aroita G, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*. 2017;40(8):913-29.
<https://doi.org/10.1111/ecog.02881>.
- Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environ Model Softw*. 2018;101:1-9.
<https://doi.org/10.1016/j.envsoft.2017.12.001>.
- Wadoux AMJC, Heuvelink GBM, de Bruin S, Brus DJ. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecol Modell*. 2021;457:109692.
<https://doi.org/10.1016/j.ecolmodel.2021.109692>.
- Meyer H, Pebesma E. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nat Commun*. 2022;13(1):2208.
<https://doi.org/10.1038/s41467-022-29838-9>.
- Williamson HF, Brettschneider J, Caccamo M, Davey RP, Goble C, Kersey PJ, et al. Data management challenges for artificial intelligence in plant and agricultural research [version 2; peer review: 2 approved]. *F1000Res*. 2023;10:324.

<https://doi.org/10.12688/f1000research.52204.2>.

Picon A, Alvarez-Gila A, Seitz M, Ortiz-Barredo A, Echazarra J, Johannes A. Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. *Comput Electron Agric.* 2019;161:280-90.

<https://doi.org/10.1016/j.compag.2018.04.002>.

Rodriguez-Vazquez J, Fernandez-Cortizas M, Perez-Saura D, Molina M, Campoy P. Overcoming domain shift in neural networks for accurate plant counting in aerial images. *Remote Sens (Basel).* 2023;15(6):1700.

<https://doi.org/10.3390/rs15061700>.

Ahmad A, El Gamal A, Saraswat D. Toward generalization of deep learning-based plant disease identification under controlled and field conditions. *IEEE Access.* 2023;11:9042-57.

<https://doi.org/10.1109/ACCESS.2023.3240100>.

Tzachor A, Devare M, King B, Avin S, Ó hÉigeartaigh S. Responsible artificial intelligence in agriculture requires systemic understanding of risks and externalities. *Nat Mach Intell.* 2022;4(2):104-9.

<https://doi.org/10.1038/s42256-022-00440-4>.

Dara R, Hazrati Fard SM, Kaur J. Recommendations for ethical and responsible use of artificial intelligence in digital agriculture. *Front Artif Intell.* 2022;5:884192.

<https://doi.org/10.3389/frai.2022.884192>.

Mallinger K, Baeza-Yates R. Responsible AI in farming: A multi-criteria framework for sustainable technology design. *Appl Sci (Basel).* 2024;14(1):437.

<https://doi.org/10.3390/app14010437>.

Klerkx L, Rose D. Dealing with the game-changing technologies of agriculture 4.0: How do we manage diversity and responsibility in food system transition pathways?. *Glob Food Sec.* 2020;24:100347.

<https://doi.org/10.1016/j.gfs.2019.100347>.

Eastwood C, Ayre M, Nettle R, Dela Rue B. Making sense in the cloud: Farm advisory services in a smart farming future. *NJAS Wageningen J Life Sci.* 2019;90-91:100298.

<https://doi.org/10.1016/j.njas.2019.04.004>.

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H 3rd, et al. Datasheets for datasets. *Commun ACM.* 2021;64(12):86-92.

<https://doi.org/10.1145/3458723>.

Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N Y).* 2023;4(9):100804.

<https://doi.org/10.1016/j.patter.2023.100804>.

Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf Fusion.* 2021;76:243-97.

<https://doi.org/10.1016/j.inffus.2021.05.008>.

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion.* 2020;58:82-115.

<https://doi.org/10.1016/j.inffus.2019.12.012>.

Alexander CS, Yarborough M, Smith A. Who is responsible for 'responsible AI'? Navigating challenges to build trust in AI agriculture and food system technology. *Precis Agric.* 2024;25(1):146-85.

<https://doi.org/10.1007/s11119-023-10063-3>.

Holzinger A, Saranti A, Angerschmid A, Retzlaff CO, Gronauer A, Pejakovic V, et al. Digital transformation in smart farm and forest operations needs human-centered AI: Challenges and future directions. *Sensors (Basel)*. 2022;22(8):3043.
<https://doi.org/10.3390/s22083043>.

Benfenati D, Amalfitano D, Russo C, Tommasino C, Rinaldi AM. Data centric artificial intelligence for agrifood domain: A systematic mapping study. *Comput Electron Agric*. 2025;239:110847.
<https://doi.org/10.1016/j.compag.2025.110847>.

Rose DC, Wheeler R, Winter M, Lobley M, Chivers CA. Agriculture 4.0: Making it work for people, production, and the planet. *Land Use Policy*. 2021;100:104933.
<https://doi.org/10.1016/j.landusepol.2020.104933>.

Rajbongshi A, Johora FT, Hossain A, Sarker MS, Rahman MH, Rahman MW, et al. Leveraging explainable AI for sustainable agriculture: A comprehensive review of recent advances. *Artif Intell Rev*. 2026;59:105.
<https://doi.org/10.1007/s10462-025-11459-5>.