

A Comprehensive Review on Multi-Modal Large Language Models and Agentic Architectures for Precision Agricultural Decision Support

Tanmay Bhamre^{1*}, Devendra Yeole¹, Yash Barde¹, Gajanan Chaudhari¹, Amarsinh Varpe¹

¹Department of Computer Science and Engineering, Sandip University, Nashik, Maharashtra, India

ABSTRACT

This article reviews recent developments in multi-modal large language models and agentic architectures for agricultural decision support. Forty-five representative studies published between 2023 and 2026 were selected across perception, language, orchestration, soil intelligence and governance, organised under a four-layer taxonomy, and classified against nine functional capabilities using a stated, reproducible procedure. The comparative analysis exposes a pronounced imbalance. Crop and disease perception is covered by 36 per cent of the selected studies, while dose-bearing prescription appears in 7 per cent, profit-aware recommendation in 4 per cent, and persistent per-field memory in a single study. Advisory tools in this literature therefore remain unable to exchange state: image-based scanners are rarely conditioned on measured soil chemistry, and crop-recommendation models optimise agronomic suitability without reference to farm economics. On this basis the article synthesises a conceptual reference architecture from the reviewed literature. It is a literature-derived synthesis rather than an implemented system, and calibrated abstention, evaluation protocols for longitudinal memory, and vernacular voice coverage emerge as the most urgent gaps.

KEYWORDS - Multi-modal large language models; agentic architectures; precision agriculture; smart agriculture; foundation models; retrieval-augmented generation; crop disease diagnosis; explainable artificial intelligence.

1. INTRODUCTION

In South Asia, agriculture supports the livelihoods of most people, and small and marginal farms dominate the holdings on which a single season's output determines household income. Three decisions recur: what to plant, how much fertiliser to apply, and whether a discoloured leaf indicates disease or a nutrient deficiency. They are still made largely on existing practice and retailer advice, since formal extension is scarce and the soil report a farmer receives is rarely tied to a specific crop programme. For roughly a decade the computational answer was task-specific deep learning, a body of work now large enough to be studied bibliometrically in its own right [1]. Such models are efficient where labelled data are plentiful, but each answers only one question, needs a labelled corpus that is expensive to build in agriculture [2], and breaks when deployment conditions differ from training. Between 2023 and 2026 the substrate changed: large language, vision and multi-modal models brought few-shot generalisation, cross-modal reasoning and the decomposition of compound requests [3], and a recent systematic review confirmed both their promise and their limited penetration into agriculture [4]. Two developments matter most: retrieval-augmented generation, which grounds output in an external curated corpus rather than in parametric memory [5], and agentic orchestration, which composes heterogeneous pipelines behind one conversational interface, with multi-agent formulations now examined for agricultural workflows [6].

Whether that reaches a farmer is a separate question. Studies on adoption find that connectivity, cost and trust are the binding rural constraints rather than accuracy, and that digital tools are discarded when infrastructure cannot support them [7], [8]. Agricultural artificial intelligence also carries a distinctive ethical weight, since those affected have little room to absorb its mistakes [9]. Offline operation, transparency and accountability are therefore functional requirements.

It leaves a particular space that needs to be filled with new reviews. For agricultural use, a survey of large models report capabilities and compare the performance of large and traditional models [4]; for disease detection, a survey of the perception accuracy is reported [10], [11]. No one asks what happens between a farmer's observation and what action a farmer takes in the field, all take as an object of analysis the model of the farmer. None then reports about a quantity of recent systems don't report a treatment class, economic recommendations or anything that can be retained in a record rich of a field between interactions. This article is intended to help you set those figures and get a better understanding of what they mean.

The present review contributes four aspects to the body of the literature on surveys. It sets up a representational collection of 45 documents from 2023 to 2026 and specifies the selection and classifiability in detail. Bathed in a four-layer taxonomy, it poses questions about similarity of vision, language and agronomy on common grounds. It presents a capability-coverage analysis for 9 dimensions that gives the impression that the field is perception based and turns it into a fact. And it builds from the analysis a conceptual reference architecture with the accounting requirements needed by the governance community. No new experiments, new data or new empirical results are reported.

Section 2 introduces the methodology and the survey by taxonomic level, compares existing literature and identifies the gaps. Section 3 sets out the conceptual architecture and Section 4 the comparative analysis, while Sections 5 and 6 discuss the advantages and social welfare of the synthesis. Section 7 concludes and Section 8 sets out future directions.

2. LITERATURE SURVEY

2.1 Review methodology and classification procedure

The corpus was compiled based on the terms "large language model," "vision-language model," and "foundation model" with "agentic model" to form one set of texts; and crop disease, crop phenotyping, soil fertility, fertiliser and crop recommendation and extension on the other set of texts as search queries were used in Scopus, IEEE Xplore, ScienceDirect, Web of Science and arXiv. The window is open in January of 2023 as this is when the skills tested in this window did not become available in usable form. Records were sorted by title and abstract for agricultural relevance and a decision support component (only relevant to title and abstract) and then by full text for methodological content. Screening applied five criteria. Only records published between January 2023 and the first quarter of 2026 were admitted; venues were restricted to peer-reviewed journals, indexed conference proceedings and influential archival preprints; the domain had to be agricultural or directly transferable high-stakes decision support; the contribution had to be a method, system, benchmark, empirical study or methodologically detailed review rather than a position piece; and where a contribution appeared more than once, only the most complete version was retained.

Each retained source was classified against the nine capabilities reported in Section 4.2. The rule was deliberately conservative: a capability was credited only where specified in the system description or evaluated in the results, never where it appears solely in motivation, related work or future work. A study reporting a fertiliser class without a quantity was not credited with prescription; one caching context within a session was not credited with persistent memory; one noting multilingual potential without an evaluated speech path was not credited with vernacular interface. Categories are not mutually exclusive, so a source addressing several was credited for each and the counts do not sum to forty-five. Classification was performed once per source against the full text; the boundary cases are treated as a threat to validity in Section 4.4.

2.2 Task-specific deep models: the pre-large-model baseline

The preceding paradigm produced highly optimised single-task detectors: citrus fruit detection using global context fusion [12] and insect detection for field deployment [13] pursue accuracy at low inference cost, while hybrid convolution-transformer designs extend the line to hyperspectral classification from unmanned aerial platforms [14], showing that transformer components were absorbed into the specialist tradition rather than displacing it. The same argument underpins agricultural robotics [15], [16], and related sensing work maps the diagnostic envelope [10], [11].

Compared rather than listed, these divide by what is optimised: [12] and [13] optimise inference cost at fixed accuracy, [10], [14] and [11] information content at fixed acquisition effort, and [15] and [16] closed-loop latency. None optimises the interpretive step that follows. Two observations carry forward: specialists remain appropriate where labelled data are plentiful and latency critical, and sensing modality sets the diagnostic ceiling, since a photograph supports lesion morphology but does not measure soil chemistry.

2.3 Large vision models and foundation perception

Foundation models altered the labelling economics of agricultural perception. Promptable segmentation showed that one pre-trained model can segment unseen objects without task-specific training [17]; later work adapted vision foundation models to plant phenotyping [18]; and a spectral foundation model reported strong reconstruction on agricultural material [19], which matters because spectral data carry nutrient and stress signal RGB imagery does not.

Set against Section 2.2, these trade in the opposite direction, buying label efficiency and generalisation at the price of a raised computational floor. The trade is unresolved in the corpus; what is visible instead is a division of labour, with large models used offline for annotation and distillation while compact models serve the real-time path. Notably [19] is the only selected source bringing foundation-scale modelling to a modality carrying direct nutrient signal, leaving the soil side conspicuously under-served.

2.4 Language models for agricultural text and decision tasks

Language-model work began with text and moved toward decision support: cross-linguistic classification of agricultural text using a general conversational model achieved competitive performance without task-specific training [20]; unsupervised named-entity recognition addressed extraction of structure from unstructured advisory text [21]; and language models have contributed to engineering itself, including agricultural robotic design [22].

These differ in where the model sits. In [20] it replaces a trained classifier, in [21] it supports an extraction pipeline, in [22] it advises a human designer, so delegated authority rises across them with no corresponding increase in verification. Domain specialisation is the recurring finding, corroborated externally in medicine, where a domain-adapted model outperformed substantially larger general models [23]. The limitation is simple: a language model reasoning over text cannot see a leaf.

2.5 Multi-modal models for crop diagnosis

Multi-modal models close that gap and form the densest region of the corpus. A survey frames the architecture in which a language model coordinates encoders for other modalities [24], and visual instruction tuning established how such models acquire instruction-following behaviour over images [25]. In agriculture, an image-text-label formulation for cucumber disease recognition showed that combining visual with textual and categorical supervision materially improves recognition under small-sample conditions [26] — the standard condition in agricultural datasets.

The agricultural instantiations are best read against one another. The wheat work [27] prioritises robustness to uncontrolled acquisition and the tomato work [28] segmentation precision under a fused language-vision objective, while the potato work [29] carries the pipeline past identification into preventive advice, making it one of the few reviewed systems whose deliverable is a recommendation rather than a label. AgroGPT [30] instead expert-tunes a compact model, addressing the deployment cost [17] and [19] incur. Evaluation infrastructure has begun to follow [31], and exploratory work extends multi-modal reasoning into breeding [32].

What none of these share is conditioning. The visual model almost always interprets an image in isolation, without measured properties of the soil the plant grows in; only [29] moves toward the prescriptive step, and without a soil profile. This matters, because nutrient-deficiency symptoms and early fungal infections are visually adjacent, yet the correct response to one is a soil amendment and to the other a fungicide. The subsection therefore shows strong image understanding, improving field realism, weak prescriptive depth and no persistence.

2.6 Soil, nutrient and recommendation intelligence

Soil work is methodologically distinct from the vision work. Explainable machine learning has been applied to predicting soil nitrogen, phosphorus and potassium, with attribution exposing which inputs drive each prediction [33]; internet-of-things instrumentation has been combined with learned models for real-time crop prediction from soil and weather data [34]; generative augmentation with interpretable modelling has served yield prediction under scarce data [35]; and time-series foundation models have been evaluated for soil-moisture forecasting [36]. Economic modelling appears rarely, forecasting of collective agricultural economic development [37] being among the few contributions treating farm economics as a target.

Compared with Sections 2.3 and 2.5, this literature is more disciplined about interpretability, since [33] and [35] report attribution as a first-class output where almost no reviewed vision study does, and about temporality, since [34] and [36] treat soil state as evolving. Against that, it is almost disconnected from the imagery literature: very few contributions fuse a soil profile with a crop image, and crop-suitability systems ignore input cost, market price and the previous season's crop. A recommendation can therefore be agronomically defensible and economically uninformative at once.

2.7 Agentic orchestration, retrieval grounding and tool use

The orchestration literature supplies the mechanism by which the preceding fragments might be joined. A framework in which a language model plans a task, selects specialised models and composes their outputs established the coordinating role such a model can occupy [38], while retrieval-augmented generation ties the resulting narration to a verifiable corpus [5]. Agricultural instantiations have begun to appear: a multi-modal agent for plant disease diagnosis [39], and a review of multi-agent formulations decomposing long agricultural workflows [6].

The general and agricultural cases differ instructively. The general framework [38] assumes a rich registry of interchangeable models, the agricultural agent [39] a narrow inventory, and the multi-agent review [6] largely describes prospective rather than evaluated configurations. Across all three, persistence of state between sessions is seldom examined, and where a quantity is computed the arithmetic is typically performed by the language model rather than an auditable procedure. Where a mis-stated fertiliser quantity carries direct financial and environmental cost that is hard to defend, and Section 3 treats it as a synthesis constraint.

2.8 Vernacular and voice-first interaction

Interface work is thinly represented but consequential. Large-scale weakly supervised training produced robust multilingual speech recognition generalising across languages and acoustic conditions without dataset-specific fine-tuning [40], making a voice-first advisory interface practical where it previously was not, while the adoption literature identifies barriers of usability and trust as much as infrastructure [7], [8]. Yet only two selected sources treat vernacular or voice interaction as a functional requirement of the diagnostic path rather than an accessibility addendum — a striking imbalance, given that the enabling capability was available throughout the review window.

2.9 Responsible deployment, ethics and governance

Governance is comparatively well developed in the corpus, which is itself a finding. Work on responsible plant-science artificial intelligence sets out the obligations arising when models influence agronomic practice [41]; an examination of responsibility allocation addresses accountability when an automated recommendation causes harm [42]; scoping of large language models in education documents over-trust, opacity and unequal access as failure modes transferring directly to extension [43]; and the case for regulatory oversight in healthcare supplies a risk-tiering template for agricultural applications of differing consequence [44]. A distinctively agricultural concern is the under-representation of plant knowledge in language models [45].

These intervene at different levels: research conduct [41], liability [42], user harm [43] and statutory oversight [44]. Together they converge on four requirements — transparency of evidence, calibrated uncertainty, farmer ownership of field data, and equitable access — none satisfied by any single technical contribution reviewed in Sections 2.2 to 2.7. The survey is organised throughout under a four-layer taxonomy of perception, reasoning, action and governance.

2.10 Comparison and analysis of existing studies

Read across the reviewed systems, two patterns recur. Capability accumulates as model classes advance but verification does not: convolutional detectors are efficient and strong on local lesion morphology yet single-task and label-hungry; vision transformers and hybrids add global context and label efficiency at a high computational floor; language models add few-shot and cross-lingual competence but no visual grounding; multi-modal models add joint image-text reasoning yet are rarely conditioned on soil and shallow in prescription; and agentic systems plan and compose specialised components while persisting little state and leaving arithmetic unaudited. At each step generality and autonomy grow, while the reported evidence for calibration, grounding and safety fails to keep pace. The second pattern is that no reviewed system carries a persistent agronomic memory, only the forecasting work [36] carrying state of any kind, so every other reviewed system re-derives its understanding of a field at each interaction.

A third pattern concerns explainability: present in the soil literature [33], [35], largely absent from multi-modal diagnosis, and demanded throughout the governance literature [41], [42], [44]. That mismatch — the strand with the most decision-critical outputs being the least interpretable — is the sharpest tension the corpus contains, and no individual study acknowledges it.

2.11 Research gaps and challenges

Consolidating the observations recorded at the end of each subsection, the reviewed literature leaves a coherent set of problems unresolved. Deficiency-disease disambiguation is the outstanding technical difficulty and, on this evidence, is not solvable from imagery alone. Evaluation is the outstanding methodological difficulty, since neither longitudinal memory nor prescription safety has an accepted protocol, so neither can be compared across systems. The gaps may be stated as follows.

- No persistent per-field memory: one source retains state across interactions, and that state is temporal rather than agronomic.
- Weak soil-image fusion: diagnosis is almost never conditioned on measured soil chemistry, leaving adjacent conditions unresolved.
- Limited economic reasoning: recommendations optimise agronomic suitability, seldom cost, price or expected margin.
- Weak multilingual and voice support: robust speech recognition exists but is rarely integrated into the diagnostic path.
- Poor explainability where it matters most: attribution is standard in soil work, largely absent from multi-modal diagnosis.
- No safety-aware recommendation: no reviewed source reports a dosage-safety envelope or an out-of-envelope rate.
- Absent calibration and abstention: accuracy is reported widely, calibration and refusal almost never.
- Unoperationalised data governance: field data are agreed sensitive, but ownership, inspection and erasure are seldom specified.

3. PROPOSED SYSTEM: A CONCEPTUAL REFERENCE ARCHITECTURE SYNTHESIZED FROM THE REVIEWED LITERATURE

The framework below is a synthesis of the reviewed literature, not an implemented or evaluated system. Each element corresponds to a requirement the selected studies establish as necessary or leave unaddressed, and the discussion describes what such a framework would need to integrate rather than what any artefact does. The workflow in Section 3.4 is likewise a reading of existing studies, not a trace of an executed process.

3.1 Design principles derived from the survey

Five principles follow from Section 2. Fusion before assertion: since deficiency and disease are visually adjacent, the reviewed evidence indicates that visual diagnosis should be conditioned on measured soil chemistry rather than asserted from the image alone [26], [27]. Grounding before generation: agronomic claims are better retrieved from a curated, versioned corpus and cited than produced from parametric memory [5], [45]. Rule-based derivation where consequence is quantitative: dosage is more defensibly obtained from an auditable procedure clamped to published safe limits [41], [42]. Persistence as structure: the field, not the interaction, is the natural unit of state. Graceful degradation: the diagnostic path should survive without connectivity, a binding constraint in the adoption literature [7], [8] and an attainable one in the efficient-model literature [30].

3.2 The synthesized framework

The conceptual reference architecture synthesised from the reviewed literature is best described as a set of dimensions rather than as software components. The interaction dimension covers image capture, document upload, vernacular voice and a longitudinal view of a field. The perception dimension concerns normalising heterogeneous inputs into a common structured observation, drawing on compact crop models [12], [13], soil texture inference, report parsing and speech transcription [40]. The orchestration dimension corresponds to the plan-perceive-retrieve-act-reflect pattern of [6] and [38], within which routing and abstention would be decided.

The agronomic reasoning dimension is, on this reading, the element that should not be learned. Matching a crop's requirement envelope against a fused soil profile, indexing nutrients against published critical limits, and deriving a corrective quantity from the measured shortfall and the carrier's nutrient content are deterministic operations, and the governance literature gives good reason to keep them so [41], [42]. The memory and retrieval dimension unifies two things the corpus treats separately: a longitudinal record of a field, and retrieval over a curated advisory corpus [5]. The tool and action dimension implies a small, schema-validated inventory, since [38] and [39] report better behaviour under constrained tool sets. The model and infrastructure dimension stays abstract, because the evidence does not settle whether hosted or on-device inference is preferable.

3.3 Accountability dimensions reported in the literature

Running across the framework are the accountability requirements of Section 2.9. A calibrated confidence would accompany each diagnosis, and below a threshold the reviewed evidence favours abstention and referral to a human expert over selection of the likelier candidate. Each agronomic claim would display its retrieved supporting passage, each quantity would stay traceable to the rule and safety limit producing it, and each tool invocation would be recorded. Stored field data would be inspectable, correctable and erasable by the farmer generating it, the practical form of the ownership question raised in [42]. Each corresponds to a failure mode documented in the corpus.

3.4 Illustrative workflow derived from existing studies

An illustrative case exercises fusion and persistence together. It is constructed from capabilities reported separately across the literature rather than describing an executed run. A farmer photographs a maize leaf showing pale interveinal striping and asks, in a regional language, what is wrong and what to apply. A vision component of the kind reported in [27] and [30] returns crop, growth stage and two candidate conditions — a zinc deficiency and an early foliar blight — with a confidence gap too narrow to resolve from the image, the failure mode of Section 2.5. A stored soil profile of the kind produced by [33], recording available zinc below its critical limit alongside a two-season cereal sequence, would resolve the ambiguity by conditioning; no reviewed system makes that retrieval. A rule-based step would then derive and clamp the corrective quantity, with the supporting advisory passage [5] retrieved alongside and the response conveyed by speech [40] as well as on screen.

The case is chosen because none of the reviewed systems could produce this answer as reported: without the stored soil profile the ambiguity does not resolve, and without rotation history the proactive advisory does not arise. Each gap identified in Section 2.11 maps onto an element of the synthesised framework: unconditioned diagnosis onto deficiency inference conditioned by a retrieved soil profile, generated claims onto retrieval over a versioned advisory corpus, model-produced quantities onto rule-based dosage with a published safety clamp, statelessness onto a longitudinal field record with trend rules, and confident output under ambiguity onto calibrated confidence with abstention and referral.

4. RESULTS AND DISCUSSION: COMPARATIVE ANALYSIS OF THE REVIEWED LITERATURE

The findings below are the outcome of the comparative analysis of the forty-five reviewed sources. They characterise the selected literature, not the performance of any implemented system; no experiments, simulations or measurements were conducted for this article.

4.1 Composition of the reviewed corpus

The corpus comprises thirty-two journal articles, nine conference papers and four archival preprints, or 71, 20 and 9 per cent. Its temporal distribution is front-loaded: twenty-one sources from 2023, thirteen from 2024 and eleven from 2025 onward. The apparent decline reflects indexing lag and selective inclusion rather than contraction of the field, and the qualitative trend runs the other way, with agentic and benchmark contributions clustering in 2025. A structural shift is visible: 2023 is dominated by single-modality models and first applications of conversational models to agricultural text, 2024 by field-realistic multi-modal diagnosis, and 2025 by agents, benchmarks and efficiency-oriented adaptation.

4.2 Capability coverage within the reviewed corpus

Applying the procedure of Section 2.1, each source was assessed against nine capabilities an end-to-end advisory system would need. Visual crop and disease perception is addressed by sixteen sources, or 36 per cent of the corpus; responsible artificial intelligence and governance by eight, or 18 per cent; soil and nutrient assessment by seven, or 16 per cent; agentic orchestration and tool use by six, or 13 per cent; knowledge grounding and citation by five, or 11 per cent; prescription with dose computation by three, or 7 per cent; vernacular or voice interface and profit-aware recommendation by two each, or 4 per cent; and persistent per-field memory by a single source, or 2 per cent. The capabilities are not mutually exclusive, so the counts do not sum to forty-five.

This is a quantitative fact that this article shows is extremely one-sided in the passage analyzed. Perception adds more than a third of the selected studies and, for an applied topic, governance is particularly rich: three sources state a dose-bearing prescription, two an economical recommendation, and one a state about a field. The action side has only a few sources: three compute a dose-bearing prescription rather than name a class of treatment, two make a state about the economy, and one, the state about a field. It

does not seem to be an issue of maturity, because the capabilities are not sequential prerequisites in any respect; it is more likely an issue of incentives—and as perception admits public benchmarks, it admits nothing like the prescription safety that would ensure the longitudinal memory that serves the driver. The result, as the literature reviewed and quoted demonstrates, is that, until now, scant literature has made the journey from observation to bounded, justified action. A farmer who would adopt the most highly reviewed system would know what the leaf contained; but at best he has no information about class of intervention; about quantity to be applied; cost that would incur on application; likelihood of return that would accrue through application; and whether a deficiency had occurred before.

4.3 Convergence and disagreement in the reviewed literature

There are three that are agreed upon. Results of multi-modal work independently presented for scale vs domain specialization for agriculture tasks [26, 30] and confirmed for neighbouring regulated domains [23] proved that domain specialization outflanks scale in agricultural tasks. Laboratory evaluation is exaggerated the field performance, which requires a move towards uncontrolled acquisition in recent diagnosis work [27] [28]. Remedial suggestions for factual drift given in literature on grounding [5] and used in agricultural reviews [4, 6] are the retrieval grounding. There are two issues that are still litigated. The first is where computation should take place: the efficiency tradition favors compact models at the edge [12, 13, 15], while the foundation-model tradition favors performing the computation on the floor since efficiency goes for generalisation, and labelled efficiency goes for models. This is not decided by the evidence, and it is a division of labour (pragmatic synthesis) [4, 30]. The second one is the locus of accountability, assigned differently by developers, deploying institutions and regulators without convergence [41] (convergence is a policy question with a technical consequence), not having a system that provides evidence for its outputs being able to hold any accountable.

4.4 Threats to validity

These findings are limited by the following: The most important is selection bias, as the corpus was not gathered by a rigorous protocol but merely a keyword search and judgemental selection, only the most recent contributions were included, and a single reviewer made the assessment of retention. The composition reported above therefore reflects those retention and selection choices as well as the field itself. The corpus deliberately has a limited scope: a broader corpus would increase the absolute numbers, but the coverage distribution would be unlikely to invert, as the perception emphasis keeps returning over and over again. Classifications will be based on evaluations and the conservative interpretations of Section 2.1 will have under credited studies that implement a capability without describing it. Last but not least, English-language study indexing is not as extensive as it could be around geographic research, at which vernacular interface work tends to be found, so this figure is a lower bound.

5. ADVANTAGES OF THE SYNTHESIZED REFERENCE ARCHITECTURE

In comparison to the system in corpus, the synthesised one would have following advantages, which can be traced to a survey finding: These are only analytical expectations and not measurable outcomes.

Additionally, diagnostic disambiguation should be improved, since the deficiency would be determined by measured soil chemistry in the past, and not based on inferred pixel data, as noted in Section 2.11. The presentation is transformed to a structural form of accountability, in that every claim is attached to a retrieved source, and the science-related quantity to some rule and safety limit, so that such an accountability scheme may constitute part of the governance literature's call for a structure of accountability. Worst case behavior is limited to restricting the language model to being arithmetic only: if the generative component fails, it will be an ill-formed description about what went wrong, not a dangerous number.

Advice would compound: When the field is the “state” instead of the session, the length of the record could make things get better, and conditions like rotation and deficiency-trend rules can be expressed that none of the stateless systems in the corpus can. An economic relevance is restored from candidate crops which are ranked according to their expected margins under exposed assumptions and is a direct result of the gap in Section 2.6, while deployability follows a direct route from a model-agnostic interface and an on-device path, as restated in [7] and [8]. This makes the division of perception, reasoning, memory and action inspectable and testable independently.

6. SOCIAL WELFARE OF THE SYNTHESIZED REFERENCE ARCHITECTURE

Welfare case involves those that have to pay the price for the mistakes this synthesis attempts to correct. Farmers pay for fertiliser for various reasons, some of which include the damage caused to soil as a result of fertiliser misapplication; and soil deficiencies of phosphorus, potassium and micronutrients, including zinc, sulphur and boron, which is the trend in many areas. A plan using the results of a soil test and feeding relationships that is both well-balanced and based on nutrient and tonnage retentions would not only use less input, but would reduce nutrient runoff as well, and have more benefits than the person that follows the advice would receive.

The second dimension is accessibility. A voice-first, vernacular interface removes literacy and language as preconditions for agronomic guidance, which matters because the population least served by formal extension is also least served by text-first, English-first tools; keeping the diagnostic path usable without a network connection extends the same reasoning to infrastructure. The third is institutional leverage: extension officers are vastly outnumbered, and a system that triages routine cases and leaves only ambiguous ones to the officer extends their reach rather than replacing them, which is the safest posture for a system whose mistakes cost money.

These directions relate to several United Nations Sustainable Development Goals. Improved diagnostic accuracy and yield stability contribute to SDG 2 on zero hunger; quantity-specific fertiliser advice bounded by published safety limits to SDG 12 on responsible consumption and production; reduced nutrient runoff and adaptation to variable seasons to SDG 13 on climate action; and the accessibility argument above to SDG 10 on reduced inequalities. Two welfare risks deserve statement. An advisory framework that recommends specific inputs can become a hidden sales channel, which argues for generic carriers and brand-neutral alternatives. And aggregated field data could be used against the interests of those who generated them, which argues for a farmer-owned, inspectable and erasable record. Both responses are partial: they are governance commitments, not technical guarantees.

7. CONCLUSION

This article examined forty-five sources published between 2023 and 2026 on multi-modal large language models and agentic architectures for agricultural decision support, categorised them under four layers — perception, reasoning, action and governance — and compared them across nine capabilities using the classification procedure stated above. Within the corpus the analysis found a strong emphasis on perception, a governance literature unusually well developed for an applied domain, and a sparse action side: dose-bearing prescription in three sources, economically framed recommendation in two, and persistent per-field memory in one.

The key component of this review is in terms of the process it is examining. Previous reviews have focused on models and their capabilities when used for the tasks they are designed to solve [4] not the distance between what a model can help a farmer do from observing to taking action, which this article measures and reports. The literature, however, is aligned in suggesting objectives that any future work should be striving for: domain specialisation is better for agricultural work than scale is; where task is evaluated in the labs, this leads to an over prediction of its performance in the field; while it is true that grounds the retrieval and, therefore, any other grounded generation problem, it remains a partial fix since all work is not grounded.

The article then combined a conceptual reference architecture that incorporates perception yet linked to soil chemistry from a time series of measurements in the field, built agronomic claims with references to the field measurement retrieved from memory and set dosage rules within a published safety envelope; included requirements for accountability across all dimensions. That architecture is a synthesis from literature in no way implemented and evaluated here, a scope of a survey which has to be kept, and to which it is the main constraint. The evidence, however, does clearly assert that an emphasis should be placed somewhere. Most of the elements of an integrated agricultural assistant can be found in the reviewed corpus. What is missing is not a better classifier but the connective structure — fusion, persistence, bounded action and traceable accountability — together with the evaluation protocols for memory, calibration and prescription safety, without which that structure cannot be compared and is therefore unlikely to be built.

8. FUTURE ENHANCEMENT

The gaps consolidated in Section 2.11 point in several directions, grouped here by near, medium and long horizons. The most pressing near-term need concerns reporting: the corpus reports accuracy widely and calibration almost never, yet where a confidently wrong answer costs a season, a well-calibrated system that abstains appropriately may be worth more than a more accurate one that

never does. Equally near-term are a direct comparison of soil-conditioned against image-only diagnosis, which would test the central claim of Section 2.5, and the integration of already-available multilingual speech into the diagnostic path rather than around it. In the medium term, an evaluation protocol for longitudinal memory is the single change most likely to move the field: a community-accepted metric — for instance the drop in recommendation quality when a field's history is withheld, adjudicated by agronomists — would make the capability comparable and therefore publishable, and its absence is the clearest reason persistent memory appears in one source out of forty-five. A prescription-safety benchmark is the analogue on the action side, since multi-modal benchmarking has begun for understanding tasks [31] and the corresponding dosage instrument would report out-of-envelope rates against published safe limits. Broadening sensing from the leaf to the plot [11], [19] belongs to the same horizon.

Longer-term aims are less definite. Currently, on-device learning relies on raw images being transferred to a cloud server to improve the model, but with federated on-device learning, the models will be able to improve without the raw imagery ever leaving a farmer's device. For the last several years the site has been instrumented with continuous soil telemetry, which could also be used for point-in-time soil laboratory cards, continuing the trend of [36]. The autonomous farmer advisory loop may proactively move toward seasonally scheduled advising autonomously based on principles indicated in the multi-agent literature [6]. Two additional extensions are often mentioned, but seldom realized: the vernacular breadth should include dialect and code-switching; there should also be linkage to market, credit, insurance and scheme interfaces, which would link diagnosis to the measures that will determine if a recommendation can be realized.

ACKNOWLEDGEMENT

The authors thank the Department of Computer Science and Engineering, Sandip University, Nashik, for the academic support and infrastructure provided during the preparation of this review. The authors are also grateful to the faculty reviewers whose comments on an earlier draft improved the structure of the survey and the comparative analysis presented here.

REFERENCES

- [1] C. M. Badgujar, A. Poulouse, and H. Gan, "Agricultural object detection with You Only Look Once (YOLO) algorithm: A bibliometric and systematic literature review," *Computers and Electronics in Agriculture*, vol. 223, art. 109090, 2024, doi: 10.1016/j.compag.2024.109090.
- [2] J. Li, D. Chen, X. Qi, Z. Li, Y. Huang, D. Morris, et al., "Label-efficient learning in agriculture: A comprehensive review," *Computers and Electronics in Agriculture*, vol. 215, art. 108412, 2023, doi: 10.1016/j.compag.2023.108412.
- [3] J. Li, M. Xu, L. Xiang, D. Chen, W. Zhuang, X. Yin, et al., "Large language models and foundation models in smart agriculture: Basics, opportunities, and challenges," *arXiv:2308.06668*, 2023.
- [4] H. Zhu, S. Qin, M. Su, C. Lin, A. Li, and J. Gao, "Harnessing large vision and language models in agriculture: a review," *Frontiers in Plant Science*, vol. 16, art. 1579355, 2025, doi: 10.3389/fpls.2025.1579355.
- [5] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, et al., "Retrieval-augmented generation for large language models: A survey," *arXiv:2312.10997*, 2023.
- [6] Y. Zhao, J. Liang, B. Chen, et al., "Applications research progress and prospects of multi-agent large language models in agriculture," *Smart Agriculture*, vol. 7, no. 5, pp. 37-51, 2025, doi: 10.12133/j.smartag.SA202503026.
- [7] F. Da Silva, S. L. C. Da Silva, F. M. Machado, J. G. A. Barbedo, and F. G. Amaral, "Farmers' perception of the barriers that hinder the implementation of agriculture 4.0," *Agricultural Systems*, vol. 208, art. 103656, 2023, doi: 10.1016/j.agry.2023.103656.
- [8] T. Dibbern, L. A. S. Romani, and S. M. F. S. Massruha, "Main drivers and barriers to the adoption of digital agriculture technologies," *Smart Agricultural Technology*, vol. 8, art. 100459, 2024, doi: 10.1016/j.atech.2024.100459.
- [9] M. Ryan, "The social and ethical impacts of artificial intelligence in agriculture: mapping the agricultural AI literature," *AI & Society*, vol. 38, pp. 2473-2485, 2023, doi: 10.1007/s00146-021-01377-9.
- [10] R. Chin, C. Catal, and A. Kassahun, "Plant disease detection using drones in precision agriculture," *Precision Agriculture*, vol. 24, pp. 1663-1682, 2023, doi: 10.1007/s11119-023-10014-y.
- [11] E. Omia, H. Bae, E. Park, M. S. Kim, I. Baek, I. Kabenge, et al., "Remote sensing in field crop monitoring: A comprehensive review of sensor systems, data analyses and recent advances," *Remote Sensing*, vol. 15, no. 2, art. 354, 2023, doi: 10.3390/rs15020354.
- [12] Y. Lin, Z. Huang, Y. Liang, Y. Liu, and W. Jiang, "AG-YOLO: A rapid citrus fruit detection algorithm with global context fusion," *Agriculture*, vol. 14, no. 1, art. 114, 2024, doi: 10.3390/agriculture14010114.
- [13] N. Wang, S. Fu, Q. Rao, G. Zhang, and M. Ding, "Insect-YOLO: A new method of crop insect detection," *Computers and Electronics in Agriculture*, vol. 232, art. 110085, 2025.
- [14] X. Guo, Q. Feng, and F. Guo, "CMTNet: a hybrid CNN-transformer network for UAV-based hyperspectral crop classification in precision agriculture," *Scientific Reports*, vol. 15, art. 12383, 2025, doi: 10.1038/s41598-025-97052-w.
- [15] Q. Yang, X. Du, Z. Wang, Z. Meng, Z. Ma, and Q. Zhang, "A review of core agricultural robot technologies for crop productions," *Computers and Electronics in Agriculture*, vol. 206, art. 107701, 2023, doi: 10.1016/j.compag.2023.107701.
- [16] F. Visentin, S. Cremasco, M. Sozzi, L. Signorini, M. Signorini, F. Marinello, et al., "A mixed-autonomous robotic platform for intra-row and inter-row weed removal for precision agriculture," *Computers and Electronics in Agriculture*, vol. 214, art. 108270, 2023, doi: 10.1016/j.compag.2023.108270.
- [17] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023, pp. 4015-4026.
- [18] F. Chen, M. V. Giuffrida, and S. A. Tsafaris, "Adapting vision foundation models for plant phenotyping," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV) Workshops*, 2023, pp. 604-613.

- [19] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5227-5244, 2024, doi: 10.1109/TPAMI.2024.3362475.
- [20] B. Zhao, W. Jin, J. Del Ser, and G. Yang, "ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification," *Neurocomputing*, vol. 557, art. 126708, 2023, doi: 10.1016/j.neucom.2023.126708.
- [21] G. Veena, V. Kanjirang, and D. Gupta, "AGRONER: An unsupervised agriculture named entity recognition using weighted distributional semantic model," *Expert Systems with Applications*, vol. 229, art. 120440, 2023, doi: 10.1016/j.eswa.2023.120440.
- [22] F. Stella, C. Della Santina, and J. Hughes, "How can LLMs transform the robotic design process?," *Nature Machine Intelligence*, vol. 5, pp. 561-564, 2023, doi: 10.1038/s42256-023-00669-7.
- [23] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, "PMC-LLaMA: toward building open-source language models for medicine," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1833-1843, 2024, doi: 10.1093/jamia/ocae045.
- [24] J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu, "Multimodal large language models: A survey," in *Proc. IEEE Int. Conf. Big Data (BigData)*, 2023, doi: 10.1109/BigData59044.2023.10386743.
- [25] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 34892-34916.
- [26] Y. Cao, L. Chen, Y. Yuan, and G. Sun, "Cucumber disease recognition with small samples using image-text-label-based multi-modal language model," *Computers and Electronics in Agriculture*, vol. 211, art. 107993, 2023, doi: 10.1016/j.compag.2023.107993.
- [27] K. Zhang, L. Ma, B. Cui, X. Li, B. Zhang, and N. Xie, "Visual large language model for wheat disease diagnosis in the wild," *Computers and Electronics in Agriculture*, vol. 227, art. 109587, 2024.
- [28] Y. Hu, J. Zhu, G. Zhou, M. He, M. Lv, J. Wang, et al., "LVF: A language and vision fusion framework for tomato diseases segmentation," *Computers and Electronics in Agriculture*, vol. 227, art. 109484, 2024.
- [29] H. Zhu, W. Shi, X. Guo, S. Lyu, R. Yang, and Z. Han, "Potato disease detection and prevention using multimodal AI and large language model," *Computers and Electronics in Agriculture*, vol. 229, art. 109824, 2025.
- [30] M. Awais, A. H. S. A. Alharthi, A. Kumar, H. Cholakkal, and R. M. Anwer, "AgroGPT: Efficient agricultural vision-language model with expert tuning," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)*, 2025, pp. 5687-5696.
- [31] A. Gauba, I. Pi, Y. Man, Z. Pang, V. S. Adve, and Y.-X. Wang, "AgMMU: A comprehensive agricultural multimodal understanding benchmark," in *Advances in Neural Information Processing Systems (NeurIPS)*, Datasets and Benchmarks Track, 2025.
- [32] G. Yang, Y. Li, Y. He, Z. Zhou, L. Ye, H. Fang, et al., "Multimodal large language model for wheat breeding: a new exploration of smart breeding," *arXiv:2411.15203*, 2024.
- [33] T. Abekoon, H. Sajindra, N. Rathnayake, I. U. Ekanayake, A. Jayakody, and U. Rathnayake, "A novel application with explainable machine learning (SHAP and LIME) to predict soil N, P, and K nutrient content in cabbage cultivation," *Smart Agricultural Technology*, vol. 11, art. 100879, 2025, doi: 10.1016/j.atech.2025.100879.
- [34] M. S. Sharafat, N. D. Kabya, R. I. Emu, M. U. Ahmed, J. C. Onik, M. A. Islam, and R. Khan, "An IoT-enabled AI system for real-time crop prediction using soil and weather data in precision agriculture," *Smart Agricultural Technology*, vol. 12, art. 101263, 2025, doi: 10.1016/j.atech.2025.101263.
- [35] I. S. Pothapragada and R. Sujatha, "GANs for data augmentation with stacked CNN models and XAI for interpretable maize yield prediction," *Smart Agricultural Technology*, vol. 11, art. 100992, 2025, doi: 10.1016/j.atech.2025.100992.
- [36] B. Deforce, B. Baesens, and E. S. Asensio, "Time-series foundation models for forecasting soil moisture levels in smart agriculture," *arXiv:2405.18913*, 2024.
- [37] C. Zheng and H. Li, "The prediction of collective economic development based on the PSO-LSTM model in smart agriculture," *PeerJ Computer Science*, vol. 9, art. e1304, 2023, doi: 10.7717/peerj-cs.1304.
- [38] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2024.
- [39] L. Qin, X. Wu, X. Dong, H. Wang, T. Yang, and Q. Wang, "PDD-Agent: Multimodal large language model-driven AI agent for enhanced plant disease diagnosis," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2025, pp. 1271-1276, doi: 10.1109/ICIP55913.2025.11084359.
- [40] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, 2023, pp. 28492-28518.
- [41] A. L. Harfouche, V. Petousi, and W. Jung, "AI ethics on the road to responsible AI plant science and societal welfare," *Trends in Plant Science*, vol. 29, no. 2, pp. 104-107, 2024, doi: 10.1016/j.tplants.2023.12.016.
- [42] C. S. Alexander, M. Yarborough, and A. Smith, "Who is responsible for 'responsible AI': Navigating challenges to build trust in AI agriculture and food system technology," *Precision Agriculture*, vol. 25, pp. 146-185, 2024, doi: 10.1007/s11119-023-10063-3.
- [43] L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, et al., "Practical and ethical challenges of large language models in education: A systematic scoping review," *British Journal of Educational Technology*, vol. 55, no. 1, pp. 90-112, 2024, doi: 10.1111/bjet.13370.
- [44] B. Mesko and E. J. Topol, "The imperative for regulatory oversight of large language models (or generative AI) in healthcare," *npj Digital Medicine*, vol. 6, art. 120, 2023, doi: 10.1038/s41746-023-00873-0.
- [45] A. Geitmann and A. J. Bidhendi, "Plant blindness and diversity in AI language models," *Trends in Plant Science*, vol. 28, no. 10, pp. 1095-1097, 2023, doi: 10.1016/j.tplants.2023.06.016.