RESEARCH ARTICLE

MaizeEviReason for Auditable Evidence Applicability and Conservative Release in Maize Disease Decision Support

Ning Zhang[1,2,3,4,5], **Huarui Wu**[1,2,3,4], **Huaji Zhu**[1,2,3,4] **and Yisheng Miao**[1,2,3,4,*]

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

Abstract

Agricultural evidence can be topically relevant yet conflict with the conditions of a query, or lack the information needed to support an action. We present MaizeEviReason, an auditable interface that keeps source conditions, distinguishes applicable, inapplicable and undetermined states, and separates applicability assessment from deterministic release checks. The core evaluation measures how much information is lost when conditions are projected onto fixed fields and what coverage is given up when supplied contracts are enforced. Matched learning experiments evaluate applicability prediction, and a public-pipeline diagnostic examines whether the interface is activated. On 49 constructed assignments from five documents, a six-field projection wrongly accepted 13 conflicting or unresolved assignments and reached an accuracy of 0.7347, against 0.8367 for direct interpretation of the text contract. The complete rule agrees perfectly with its reference, as expected, because those labels were derived from the same contracts. On 120 constructed release records, deterministic gating reduced observed symbolic violations from 7 to 0, while coverage fell from 55.83% to 50.00%. The prespecified source-family comparison of decision accuracy was not significant (Holm-adjusted $p = 0.75$), and the eight-record source-bound test covered only one family. On ten permanently held-out documents, the condition-differential evidence adapter (CDEA) obtained a macro-F1 of 0.4949 with the state loss and 0.2518 with its full objective. In the public evaluation the condition interface was never activated, so its effect under usable condition inputs remains untested. Taken together, these descriptive diagnostics locate information loss in the evaluated condition representation and mark the operating scope of applicability assessment and output checking; they are not evidence of general decision-support effectiveness.

Submitted: 10 June 2026
Revised: 30 October 2026
Accepted: 31 October 2026
Published: 08 November 2026

Vol. 2, **No.** 4, 2026.
10.62762/DIA.2026.954088

***Corresponding author:**
✉ Yisheng Miao
miaoys@nercita.org.cn

Citation
Zhang, N., Wu, H., Zhu, H., & Miao, Y. (2026). MaizeEviReason for Auditable Evidence Applicability and Conservative Release in Maize Disease Decision Support. *Digital Intelligence in Agriculture*, 2(4), 191–202.

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

Keywords: maize disease, evidence applicability, condition representation, conservative release, retrieval-augmented generation, source-separated evaluation.

1 Introduction

Digital tools for crop disease management have to connect observations of plant health with knowledge that is relevant to a particular production setting. Machine learning and deep learning support disease recognition, crop monitoring and prediction [1, 2], and research on agricultural decision support stresses that the information delivered to users must be relevant and usable [3]. In maize disease management, a retrieved recommendation may depend on the affected organ, growth stage, region or management objective. Checking these conditions is therefore a necessary step between finding relevant information and using it to support a decision.

Retrieval-augmented generation (RAG) grounds language-model responses in external knowledge [4]. Dense passage retrieval learns query-passage matching [5], while multilingual embeddings and graph-based memory widen the available retrieval mechanisms [6, 7]. Agricultural applications combine these capabilities with domain information. CornCare integrates visual recognition and knowledge graphs for maize disease diagnosis [8]; a condition-aware retrieval framework for tomato cultivation incorporates cultivation context [9]; Time-Aware Retrieval-Augmented Generation (TARAG) addresses temporal applicability in crop protection [10]; and the Structured Multimodal Agricultural Retrieval-Augmented Transformer (SMART) uses structured multimodal guidance for plant disease diagnosis [11]. These systems show that retrieved evidence needs to be linked to the agricultural setting in which it will be used.

General-purpose RAG research has also examined how reliably evidence is used. Self-RAG learns when to retrieve and how to critique its own output [12], and Corrective Retrieval Augmented Generation (CRAG) scores retrieved material to trigger corrective actions [13]. Gao et al. [14] separate answer correctness from citation quality in their evaluation framework, and word-level analysis of unsupported or contradictory claims in RAG responses shows that faithfulness failures arise even when relevant passages are retrieved [15]. Retrieval quality, citation support and contextual conditions thus act on different parts of the evidence-use process. The question left open for the present study is how to carry the necessary conditions of a source through the evidence interface and make their role in applicability and output decisions explicit.

The question matters most when source statements contain compound requirements or when the query gives incomplete context. A statement that requires either of two alternatives together with a third condition cannot be represented faithfully by listing the same conditions without their logical grouping. Likewise, a missing growth stage signals neither agreement nor contradiction. A condition-aware interface has to keep these distinctions and tie them to an explicit output policy, and its evaluation has to tell apart errors from incomplete condition recovery or representation and errors in prediction and release control.

We therefore introduce MaizeEviReason, an auditable framework for evidence applicability and conservative release in maize disease decision support. It stores source spans, necessary conditions, extraction status and decision reasons, and labels evidence as applicable, inapplicable or undetermined. Deterministic checks then decide whether an output may be released under the supplied conditions. Using constructed tests and source-bound records, we examine information loss under condition projection and extraction, the coverage cost of enforcing supplied contracts, and applicability prediction by a condition-differential evidence adapter (CDEA) under matched learning controls. A public-pipeline diagnostic checks whether the condition interface actually takes part in judgment. The evidence concerns condition compliance and interface behaviour; agronomic effectiveness would require independent validation.

The study makes three contributions. First, it defines an evidence interface that keeps the represented condition structure and unresolved states open to inspection and separates applicability assessment from output checking. Second, it compares representations on identical condition assignments to diagnose information loss caused by projection, and complements this with an analysis of automatic condition recovery. Third, it evaluates controlled release by both symbolic violations and output coverage, with separate learning and activation diagnostics that clarify where the mechanism stops working. These contributions give a basis for judging when retrieved agricultural knowledge can be used under stated conditions and what information is still

Table 1. Core evaluation cohorts and the limits of their reference standards. Populations overlap or use different units and must not be summed as independent agricultural cases.

Track	Population / unit	Reference and limit
Predicate diagnostic	5 documents; 49 assignments	Supplied predicates; development diagnostic
Automatic extraction	32 cases; 128 variants; 160 unique jobs	20-case cohort: 80 state judgments
Symbolic release	120 synthetic + 8 source-bound records	Given contracts; sourced test: 1 family
Synthetic learning	100 / 20 / 30 base cases; 120 test variants	Constructed contracts; 5 test families
Literal-scope learning	12 / 4 / 4 base cases; 16 test variants	Literal scope; 4 test groups
Permanent holdout	10 documents; 40 scope records	Constructed literal states; 2 task-bridge components

needed when a decision has to be deferred.

2 Data and Dataset Construction

2.1 Condition references and source separation

Table 1 lists the evaluation populations. A contract is an explicit specification of the conditions required by a source statement, and its reference label records consistency with that specification rather than an independently established agronomic outcome. Constructed conditions, source-bound literal scope and official question answers are kept as separate reference types throughout. Following the transparency rationale of Datasheets for Datasets [16], we document cohort composition, reference provenance and intended use, so that it is clear what each evaluation can support.

The synthetic pilot contains 100 training, 20 development and 30 test base cases drawn from 12, 3 and 5 source families, respectively. Each base case has an applicable, a conflicting, a missing-value and a meaning-preserving variant, all kept in the same split. The 120 test records therefore contain 60 applicable, 30 inapplicable and 30 undetermined states. Repeated variants and seeds do not add independent sources.

The literal-scope pilot starts from 20 source statements split into 12 training, 4 development and 4 test base cases. Each has four variants that retain the task and material qualifiers of the source, so the 16 test records represent four test groups. A later fixed-budget diagnostic uses ten permanently held-out documents, each with four constructed scope states. These documents were never used for training, epoch selection or seed selection, but they form only two conservative task-bridge components and therefore give a bounded transfer diagnostic rather than broad validation on unseen families.

2.2 Representation, extraction and release cohorts

The complete-predicate diagnostic uses supplied logical contracts from five documents and 49 constructed assignments (20 satisfied, 14 conflicting and 15 unresolved). It compares the complete rule, a six-field projection and direct interpretation of the text contract on the same assignments. Because the reference labels of the rule are generated from the supplied contracts, perfect agreement is only an internal-consistency check.

The automatic-extraction diagnostic contains 32 base cases, 128 variants and 160 unique query/evidence extraction jobs. A 20-case literal-scope cohort within this diagnostic supplies 80 state judgments for checking whether the extracted conditions are sufficient; these records are not additional CDEA training examples. Fixed-candidate release uses the 120 synthetic records and a separate eight-record source-bound literal test from one source family. The two release tracks together comprise 128 records ($120+8$), shared across all mechanisms; the eight-record release test should not be confused with the 16-record learned pilot.

2.3 Supporting data and interface diagnostics

Supporting image and scenario data come from two maize disease dataset releases [17, 18]. Visual recognition, calibration, topical retrieval and the original 80-case maize scenario study serve only as upstream context and are not analysed further here. Their references are source-image labels, silver relevance judgments or constructed context, with no independent case-level reference for action support, so they define the scope of upstream controls and local adaptations rather than providing an end-to-end test of the present mechanism. The public-interface diagnostic uses the pinned release

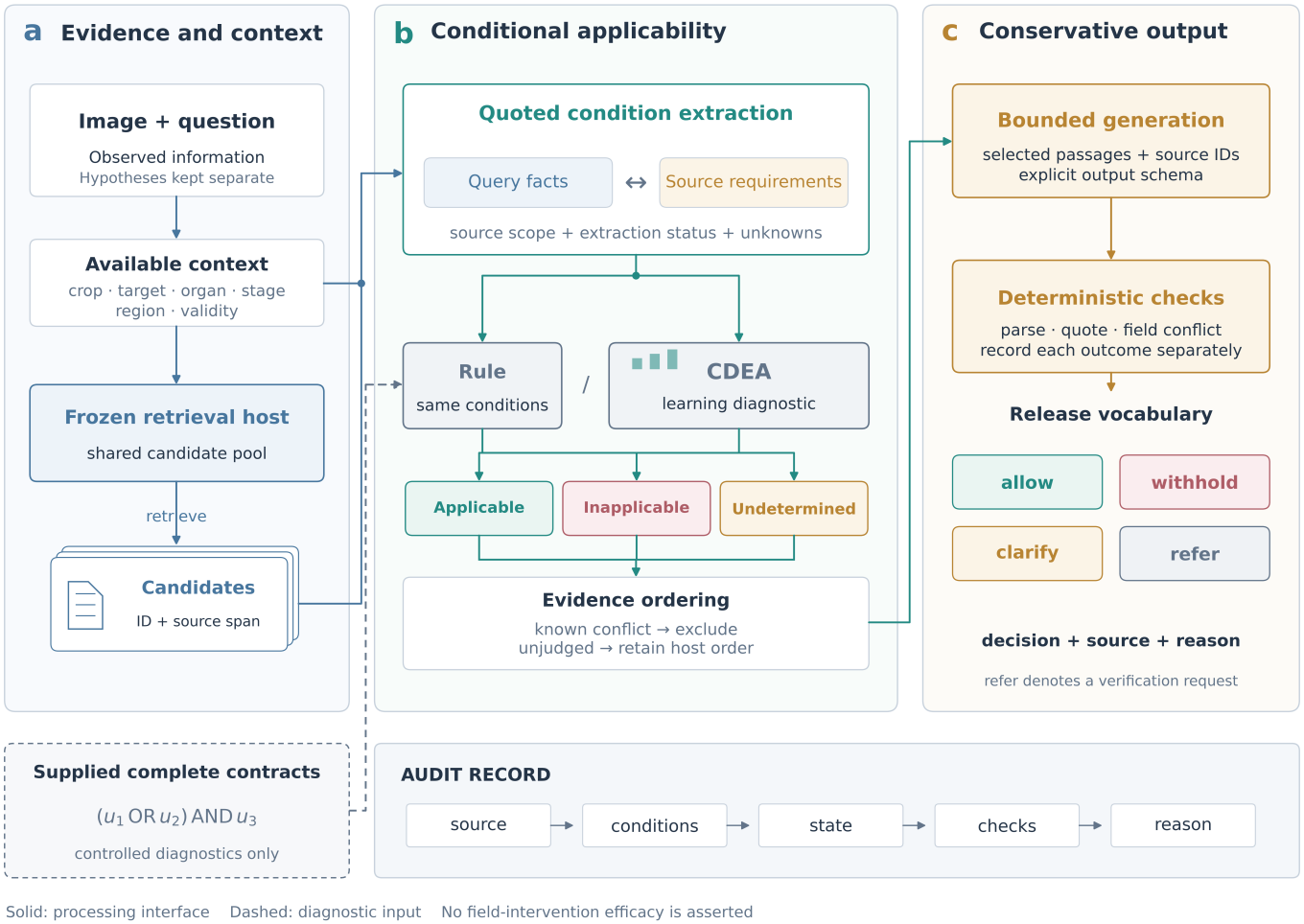


Figure 1. Evidence interface and its evaluation boundaries. Panels show (a) evidence and context, (b) conditional applicability and (c) conservative output. The core framework preserves source conditions, three-state judgments and independently recorded release checks. The grey condition-differential evidence adapter (CDEA) branch denotes a learned diagnostic; ID denotes identifier. The dashed input supplies complete contracts only for controlled tests and does not represent automatic recovery of compound predicates. Public reranking keeps unjudged candidates in host order, whereas the separate symbolic-release task can withhold unresolved actions. The interface is not an end-to-end validated agricultural intervention pipeline.

of the Agricultural Multimodal Understanding (AgMMU) benchmark [19].

3 Proposed Method

3.1 An auditable evidence interface

Figure 1 separates source acquisition, applicability assessment and release control. Each candidate keeps its document identifier, verbatim source span, necessary conditions, extraction status and resulting applicability state. The image, question and available context supply observations and provisional hypotheses; a predicted disease is kept apart from an observed fact, and the frozen retrieval host provides the shared candidate pool.

The quoted-field extractor admits only directly supported facts and necessary or prohibited conditions. It does not infer them from filenames, answer options or background knowledge, and ambiguous alternatives stay unknown. Empty requirements and extraction failure are recorded with different reasons, and neither implies applicability. Automatic extraction is treated as a possible point of failure, not as an oracle.

3.2 Preserving logical scope and unresolved states

A conditional contract specifies what must hold for a source statement to apply. The original span and its logical scope are kept separate from any reduced, field-based representation. For example, $(u_1 \vee u_2) \wedge u_3$ requires either of two alternatives together with a third condition, and a list of the same atoms without their grouping loses this requirement. The example illustrates representation semantics and

is not an additional agricultural test result. The complete-predicate diagnostic supplies the logical structure explicitly and compares it with a six-field projection on identical assignments; we do not claim a general automatic predicate parser.

We use three applicability labels. *Applicable* means that the available context satisfies the recovered or supplied necessary conditions under their stated logic. *Inapplicable* means that known context violates those conditions. *Undetermined* means that a condition contract is available but missing or ambiguous context prevents its truth value from being resolved. *Unjudged* is a processing status, not a fourth class: no applicability label is issued because no evaluable condition representation was obtained, for example after an empty or invalid extraction. For compound contracts, labels follow the full logical expression rather than isolated mismatches between atoms.

The six-field same-input rule first checks for known contradictions in the necessary fields, which yield inapplicable. Without a contradiction, missing necessary information yields undetermined; otherwise every necessary field must match for the record to be applicable. Complete predicates are evaluated under their supplied logic. The complete-tree rule, the projected-field rule and the learned field-based adapter thus receive explicitly distinguished representations, and perfect agreement of the tree rule with a reference derived from the same contract measures internal consistency only.

Fact Extraction and VERification (FEVER) distinguishes supported, refuted and insufficiently evidenced claims [20]. Our states concern instead whether the necessary conditions of a source are met; they say nothing about whether the source statement itself is true. The distinction between missing knowledge and intrinsic variability drawn in uncertainty analysis [21] is also relevant here: the undetermined state marks missing information needed to evaluate the contract, not a calibrated probability of error.

Applicability is also distinct from the decision to release an action. In public reranking, explicit failures are excluded and unjudged candidates keep their original relative order, so retaining a passage does not assert that it supports an action. In the separate fixed-candidate release task, only a reference-applicable symbolic item is eligible for release. The two policies serve different tasks and their outcomes are never pooled.

3.3 Deterministic checks and symbolic release

The interface records parsing status, evidence identifiers, literal quotation checks and explicit field conflicts separately. The procedural output vocabulary is *allow*, *withhold*, *clarify* and *refer*, where *refer* denotes a need for external verification rather than an observed expert intervention. A literal quotation check verifies copied text but not whether the text entails a proposed agricultural action.

The fixed-candidate study compares soft prompting, pre-generation gating, gating with explicit deferral, and deterministic checking of the soft output. Each record supplies one symbolic action and a complete reference contract. Post-checking reuses the soft response, which isolates the effect of the check from that of new generation. The outcomes are release coverage, violation rate over all records and completion of reference-applicable records, and they describe the supplied symbolic task only. Selective prediction, exemplified by SelectiveNet [22], likewise evaluates accepted predictions together with coverage. Our release study relies on explicit contracts rather than a learned reject option, but the connection motivates reporting withheld output alongside observed violations.

3.4 CDEA as a matched learning diagnostic

CDEA is evaluated as a learning diagnostic. It combines fixed query and evidence embeddings with explicit condition comparisons (Figure 2). Let q and e denote the 1,024-dimensional query and evidence embeddings produced by BGE-M3 (BAAI General Embedding, M3) [6]. Their concatenation with the elementwise product forms the semantic vector a. The 42-dimensional condition vector c covers crop, target, organ, stage, region and validity: for each field it contains four comparison indicators (match, conflict, unknown and irrelevant) and three indicators that record whether the field is necessary and whether the query and evidence values are present. GELU denotes the Gaussian error linear unit; W and b denote trainable weight matrices and bias vectors.

$$a = [q;\, e;\, q \odot e], \quad a \in \mathbb{R}^{3072}, \quad c \in \mathbb{R}^{42} \tag{1}$$

$$z = \mathrm{GELU}(W_p a + b_p), \quad g = \mathrm{sigmoid}(W_g c + b_g) \tag{2}$$

$$h = z + W_{\uparrow}\big[g \odot \mathrm{GELU}(W_{\downarrow} z + b_{\downarrow})\big] + b_{\uparrow} \tag{3}$$

$$p = \mathrm{softmax}\big(W_h[h;\, c] + b_h\big), \quad p = [p_A, p_I, p_U] \tag{4}$$

Here $z, h \in \mathbb{R}^{64}$, $g \in \mathbb{R}^{16}$, and p_A, p_I and p_U are the predicted probabilities of the applicable, inapplicable

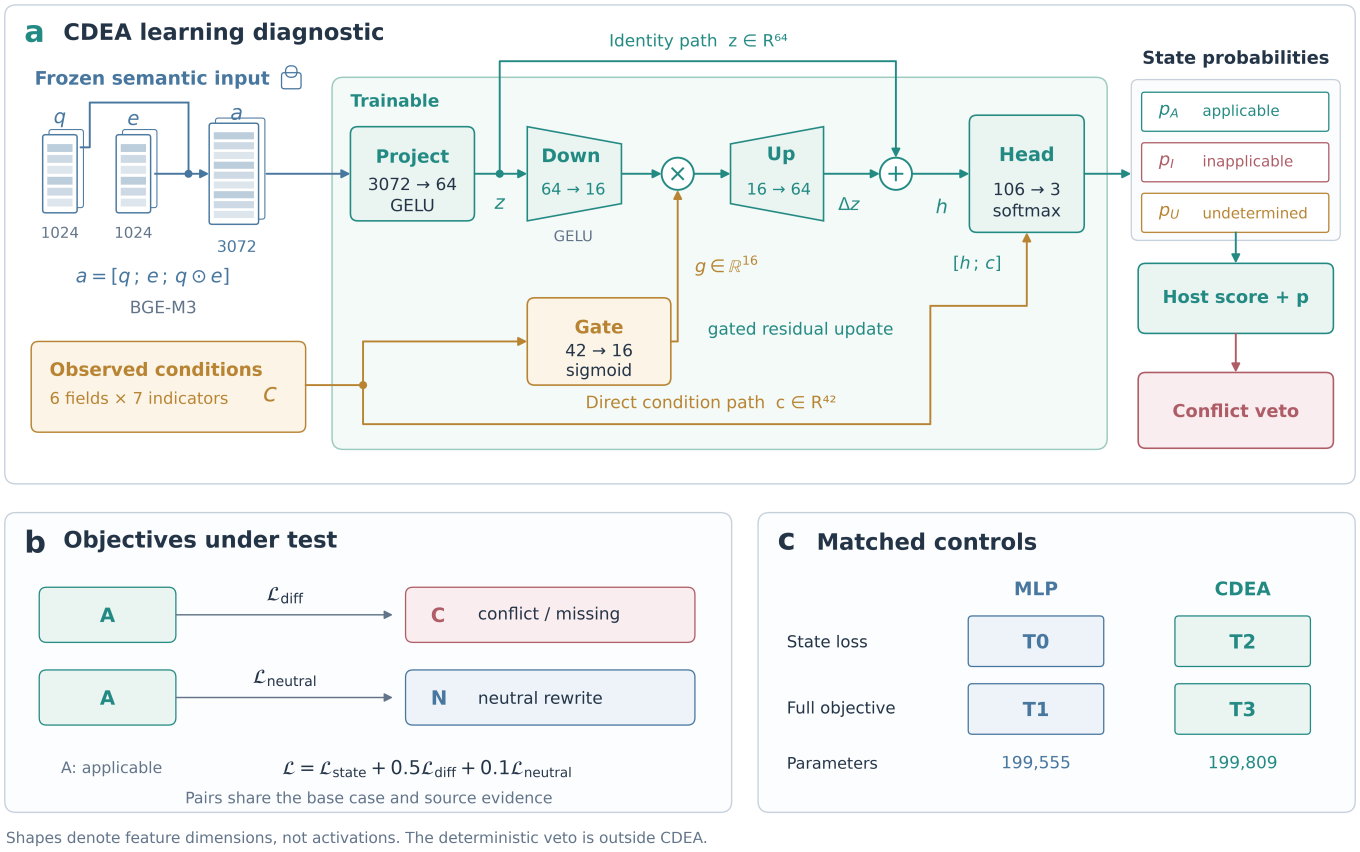


Figure 2. CDEA as a matched learning diagnostic. Panels show (a) CDEA computation, (b) paired objectives and (c) matched architecture and loss controls. CDEA denotes condition-differential evidence adapter, MLP multilayer perceptron and GELU Gaussian error linear unit. A denotes an applicable case, C a conflict or missing-condition variant and N a neutral rewrite. The identity path and the direct 42-dimensional condition path are retained. T0 is the MLP with state loss, T1 the MLP with the full objective, T2 CDEA with state loss and T3 CDEA with the full objective; the deterministic veto lies outside the learned model. The full objective is a tested hypothesis, not a configuration shown to be superior.

and undetermined states. The semantic projection has width 64 and the gated residual bottleneck has width 16. The final classifier takes the 64-dimensional adapted vector together with the direct 42-dimensional condition input. All CDEA linear layers are trainable and BGE-M3 is frozen, giving 199,809 trainable parameters. The comparator concatenates a and c, passes them through one 64-unit hidden layer with GELU activation and predicts the same three states with 199,555 parameters, about 0.13% fewer than CDEA. A deterministic rule receives the same exposed condition information.

The four controls T0-T3 in Figure 2 cross two architectures, CDEA and a capacity-matched multilayer perceptron (MLP), with two objectives, state loss and the full objective. The paired terms contrast changed conditions with meaning-preserving rewrites. All four arms are reported; T2 was not chosen after inspection to replace the originally frozen public T3 checkpoint. Learned scores cannot override explicit conflicts, and the controlled release comparison isolates deterministic enforcement rather than CDEA.

4 Experimental Protocol

4.1 Controlled release and original endpoints

Coverage is the number of released records divided by the total number of input records. The violation rate is the number of records released despite an inapplicable or undetermined reference label, divided by the total number of input records. Completion is the number of correctly released reference-applicable records divided by the total number of reference-applicable records.

The four mechanisms share the 128 input records. Generating decisions for the three generative mechanisms requires 384 requests (3×128), whereas deterministic post-checking reuses the soft outputs. The ledger keeps 383 received responses and counts one interrupted, unverifiable request as a failure. The

archived protocol names decision accuracy as the primary inferential endpoint; this report concentrates on descriptive coverage and violation. A closed-corpus comparison with CRAG [13] is not reported, because its single-candidate setting does not match the present task.

4.2 Matched learning and independent units

The original T0-T3 pilots use seeds 17, 29, 43, 71 and 101, AdamW with a learning rate of 0.001 and a weight decay of 0.0001, a batch size of 64 and at most 50 epochs. Early stopping uses a patience of five epochs and selects the earliest epoch that attains the maximum development macro-F1. A later diagnostic with a fixed budget of 50 epochs holds samples, architecture, losses, optimizer, seeds and batch order constant to examine the original stopping behaviour. It was designed after the pilots and is not a preregistered confirmatory experiment.

Applicability is scored by accuracy, macro-F1 over the three fixed states, condition-change response and neutral-change rate. A paired response counts as correct only if both ends receive the correct state, so a constant predictor can be stable under neutral changes while being insensitive to condition changes. Seeds quantify training variability. Statistical independence is assessed at the source or base-case level; variants, predictions and seeds are not counted as new sources. The invariance and directional-expectation tests of CheckList [23] are a methodological precedent for separating responses to meaningful changes from stability under neutral ones, and the risk that contaminated evaluation units inflate apparent performance motivates respecting source-level independence when defining evaluation populations [24].

The synthetic pilot averages repeated seeds within cases and uses its five test families for paired bootstrap and exact sign-flip comparisons. Its three contrasts compare T3 with T1, T2 and the rule, with Holm correction [25]. The four literal test groups and the two task-bridge components of the permanent holdout limit population inference.

5 Results

5.1 Condition completeness limits applicability assessment

On the 49 supplied-contract assignments, the six-field projection reaches an accuracy of 0.7347 and a macro-F1 of 0.7259 (Table 2). It wrongly accepts six conflicting and seven unresolved assignments. Direct interpretation of the text contract reaches an accuracy of 0.8367 and a macro-F1 of 0.8176; it wrongly accepts seven unresolved assignments, and its one remaining error labels an unresolved assignment as conflicting. The complete rule agrees perfectly with the reference generated from its own predicates. These differences are descriptive comparisons within a constructed diagnostic: they show information loss in the chosen projection but do not validate the agronomic correctness of the predicates.

Table 2. Complete-predicate diagnostic on the same 49 constructed assignments from five documents. Perfect complete-rule agreement follows from the reference construction. False acceptance means that a conflicting or unresolved assignment is classified as satisfied.

Representation	Acc.	Macro-F1	False acceptance
Six-field projection	0.7347	0.7259	13 (6 conflict + 7 unresolved)
Direct text interpretation	0.8367	0.8176	7 unresolved
Complete supplied rule	1.0000	1.0000	0 by construction

Automatic extraction adds a further bottleneck. In the 20-case literal cohort, the query fields match their contract references and the extraction outputs are structurally valid, yet the recovered sets of necessary evidence conditions do not match their references, so all 80 records receive the unjudged processing status. The extractor accepts only explicitly quoted necessary or prohibited conditions; descriptive mentions and ambiguous alternatives are not promoted to requirements. A structurally valid output can therefore lack the condition information needed for judgment. The result points to insufficient condition recovery in the evaluated interface, not to an absence of applicable conditions in the source statements.

5.2 Deterministic checks and the coverage cost

On the 120 constructed records, soft prompting releases 67 records, seven of which are violations. Pre-gating and post-checking each release 60 records with no observed violation. Coverage thus falls from 55.83% to 50.00%, a decrease of 5.83 percentage points, while completion of reference-applicable records stays at 100%. Gating with explicit deferral releases 59 records and reaches a completion of 98.33%, because the retained failure is not counted as successful assistance (Table 3 and Figure 3).

Table 3. Fixed-candidate release results. Released and violation entries are counts over all input records; coverage is the released count divided by the total number of input records, and completion is the correctly released applicable count divided by the number of reference-applicable records. Synthetic records contain 60 applicable items and source-bound records contain four. Post-checking reuses the soft outputs; the interrupted request remains a failure.

Track / mechanism	Released	Violations	Coverage (%)	Completion (%)
Synthetic / Soft prompt	67/120	7/120	55.83	100.00
Synthetic / Pre-gate	60/120	0/120	50.00	100.00
Synthetic / Gate + defer	59/120	0/120	49.17	98.33
Synthetic / Post-check	60/120	0/120	50.00	100.00
Source-bound / Soft prompt	6/8	2/8	75.00	100.00
Source-bound / Pre-gate	4/8	0/8	50.00	100.00
Source-bound / Gate + defer	4/8	0/8	50.00	100.00
Source-bound / Post-check	4/8	0/8	50.00	100.00

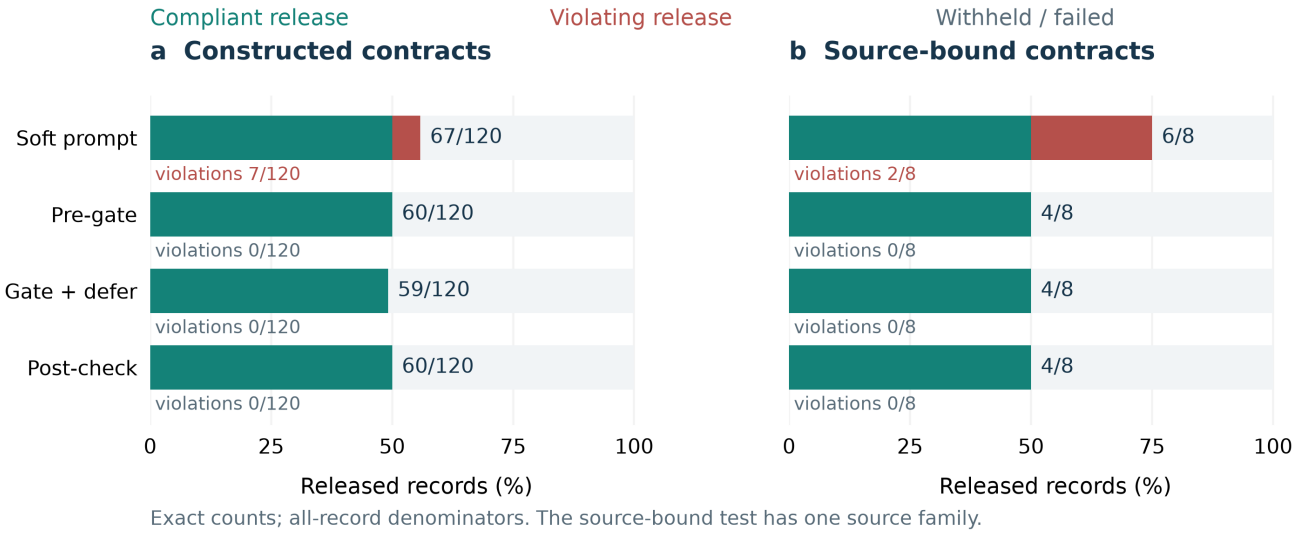


Figure 3. Observed coverage-violation trade-off under supplied contracts. Panels show (a) 120 constructed records and (b) eight source-bound records from one source family. Bars distinguish compliant and violating releases; pale grey segments denote withheld or failed output. Labels report exact counts and all-record denominators. No population confidence interval or agronomic safety interpretation is implied.

On the eight source-bound records, soft prompting releases six, including two violations, and each deterministic mechanism releases four with no observed violation. Coming from a single source family, this result cannot be used to estimate a population violation rate. The reductions are attributable to enforcement of the supplied contracts, not to CDEA.

For the original decision-accuracy endpoint, the source-family comparison of post-checking against soft prompting gives a Holm-adjusted p-value of 0.75. The comparison therefore shows no significant gain in decision accuracy, and non-significance should not be read as equivalence. The zero observed violation count is a separate, descriptive finding.

5.3 Learning diagnostics and limits of the full objective

All learned arms predict applicable for every record in the original synthetic test, which gives an accuracy of 0.5000, a macro-F1 of 0.2222 and a condition-change response rate of zero. In the literal pilot, T3 obtains a macro-F1 of 0.2869 ± 0.1093 and a condition-change response rate of 0.0500 ± 0.1118 (mean $\pm$ standard deviation over five training seeds); with four test groups, no stable advantage can be shown. The same-information rule is perfect against the supplied contract labels.

All 20 original synthetic training runs select epoch 1 and stop at epoch 6. The fixed-budget replay reproduces the first six epochs and then reaches a high

Table 4. CDEA and capacity-matched controls on 40 constructed scope records from ten permanently held-out documents. Values are five-seed means. CE denotes cross-entropy state loss; full denotes state loss plus the paired differential and neutral terms; MLP denotes multilayer perceptron; CDEA denotes condition-differential evidence adapter. The documents form two conservative task-bridge components; the rule is evaluated against supplied contract truth.

Arm	Epoch 1 accuracy	Epoch 1 macro-F1	Epoch 50 accuracy	Epoch 50 macro-F1
T0 MLP / CE	0.500	0.2222	0.250	0.1333
T1 MLP / full	0.500	0.2222	0.250	0.1333
T2 CDEA / CE	0.505	0.2351	0.545	0.4949
T3 CDEA / full	0.505	0.2351	0.320	0.2518
Contract rule	1.000	1.0000	1.000	1.0000

development macro-F1, which identifies the stopping rule as one optimization bottleneck; these development results are not independent test results. On the ten permanently held-out documents, T2 reaches a macro-F1 of 0.4949 at epoch 50, compared with 0.2518 for T3 and 0.1333 for both MLP controls (Table 4). The higher T2 score justifies further work within this small diagnostic, but with only two task-bridge components it supports neither a general advantage, nor the full objective, nor a causal account of why that objective failed. No arm is promoted to a new confirmatory result.

6 Discussion

6.1 Conditional information is the substantive object of evaluation

The representation and release experiments address two connected but distinct questions. The projection diagnostic asks whether an interface keeps the distinctions present in a supplied contract; the release study asks what enforcing that contract changes in a fixed output task. Keeping a crop or stage label is not enough if alternatives, conjunctions or unknown necessary values lose their meaning. This finding concerns the six-field projection that was evaluated; richer structured representations may avoid the loss.

The contribution is therefore more specific than applying a rule to its own reference. The study shows where the chosen representation accepts conflicting or unresolved assignments, keeps the source-level scope needed to inspect those errors, and reports both the benefit and the cost of checking generated releases. Perfect complete-rule agreement remains a consistency result, and zero observed violations remain a property of the evaluated records; neither validates the truth, completeness or agronomic consequences of the supplied conditions.

6.2 Prediction failure and interface inactivity require different explanations

CDEA has two separate limitations in the present evidence. Under supplied inputs, the original pilots collapse and the full objective lacks stable support on the later holdout. T2 scores higher than the matched MLP controls in that small diagnostic, which is a reason for further investigation rather than evidence of general superiority, and it says nothing in favour of the full objective. The post hoc ablations of the differential and neutral losses are limited in scope and cannot isolate a general cause of failure.

The quoted extractor does not recover sufficient source requirements in the literal cohort, and the public diagnostic likewise leaves all 15,420 candidate condition states unjudged. On 771 AgMMU questions, hybrid retrieval with and without the frozen CDEA answers 678 questions correctly, with identical top-20 candidate order and identical predicted answers. The condition interface therefore did not take part in this pipeline, and the comparison tests neither CDEA under usable condition inputs nor a gain on public questions. The unjudged cases reflect limited condition recovery and judgment coverage, not an observed applicability class. Keeping unjudged candidates in host order avoids unsupported applicability claims, but it also leaves the benefit of the module under usable conditions unestimated. Condition recovery, prediction given conditions and checked output therefore need separate evaluation on new sources after any change to losses, prompts or stopping rules.

6.3 Agricultural scope and evidence limits

For agricultural knowledge systems, the interface makes a missing growth stage, a conflicting source condition or an unresolved requirement visible and records why a decision was deferred. This is an audit function. Disease recognition, topical retrieval, applicability and action support remain separate claims; any improvement in recognition

from the supporting visual adaptation, which is not reported here, would leave its effect on downstream applicability and action support untested.

The contract studies rely on few independent sources, constructed states and a reference tied to the supplied conditions, and the source-bound release test covers only one family. The supporting relevance judgments are incomplete silver labels, the public retrieval corpus is restricted, and the source-image labels were not independently re-established. The later diagnostics were designed after earlier results were known, and the name of a model endpoint alone cannot confirm the provenance of its public weights. Any local re-implementation of CornCare, tomato cultivation retrieval or graph-based methods, and any generation-coverage result, is subject to implementation and scoring limitations.

No experiment measures treatment efficacy, yield, environmental effect or prospective benefit to farmers. The next priority is independently sourced conditions with an explicit reference scope, followed by evaluation of condition recovery and of the activated interface on held-out sources.

7 Conclusions

MaizeEviReason provides an auditable framework for assessing whether retrieved agricultural evidence applies to a stated context and for checking outputs against supplied conditions. The experiments yield two descriptive findings. On 49 constructed assignments from five documents, the six-field projection falsely accepts 13 conflicting or unresolved assignments, which shows that keeping field values alone can lose distinctions in the supplied logical scope. On 120 constructed release records, deterministic enforcement reduces observed symbolic violations from 7 to 0, while coverage decreases from 55.83% to 50.00%. These changes concern compliance with the supplied contracts and come at a measurable cost in output availability; the original source-family comparison shows no significant gain in decision accuracy (Holm-adjusted $p = 0.75$).

The contribution is a condition-preserving evidence interface together with a diagnostic evaluation of representation loss and controlled release. Source spans, necessary conditions, unresolved states and decision reasons stay available for inspection. Because condition recovery, applicability prediction and output checking are separated, their failures can be told apart, and effects of deterministic enforcement are not credited to the learned adapter. For agricultural knowledge systems, this gives a way to inspect the conditions attached to retrieved advice and to identify the information still needed before an action can be supported. The demonstrated scope is auditability within the evaluated tasks and explicit handling of incomplete evidence.

The findings are bounded by constructed states and few independent sources. CDEA shows no stable superiority with its full objective, and the public pipeline does not activate the condition interface. Agronomic safety, treatment efficacy and farmer benefit are not established. Further validation requires independently sourced conditions, reliable recovery of their scope and evaluation of an activated interface on held-out sources.

Data Availability Statement

Data will be made available on request.

Funding

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

Conflicts of Interest

Ning Zhang is affiliated with the Beijing Paide Weiye Technology Development Co., Ltd., Beijing, China. The authors declare that this affiliation had no influence on the study design, data collection, analysis, interpretation, or the decision to publish, and that no other competing interests exist.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors, 18*(8), 2674. [CrossRef]

[2] Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and electronics in agriculture, 147*, 70-90. [CrossRef]

[3] Rose, D. C., Sutherland, W. J., Parker, C., Lobley, M., Winter, M., Morris, C., ... & Dicks, L. V. (2016). Decision support tools for agriculture: Towards effective design and delivery. *Agricultural systems, 149*, 165-174. [CrossRef]

[4] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems, 33*, 9459-9474.

[5] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 6769-6781). [CrossRef]

[6] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 2318-2335). [CrossRef]

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

[8] Liu, Y., Xie, Y., Wu, X., & Wang, Q. (2026). CornCare: A knowledge-graph-enhanced multimodal diagnostic reporting system for corn diseases. *Agriculture, 16*(10), 1109. [CrossRef]

[9] Wang, Y., Zhao, K., Li, H., Wang, R., & Chen, W. (2026). A condition-aware retrieval-augmented decision support framework for tomato cultivation management. *Frontiers in Plant Science, 17*, 1802748. [CrossRef]

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

[11] Chan, Y. C., Fan, Y. C., Li, H. C., & Yang, M. D. (2026). SMART: Structured multimodal agricultural retrieval-augmented transformer for plant disease diagnosis. *Computers and Electronics in Agriculture, 250*, 111882. [CrossRef]

[12] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024, May). Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations* (Vol. 2024, pp. 9112-9141).

[13] Yan, S. Q., Gu, J. C., Zhu, Y., & Ling, Z. H. (2024). Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*. [CrossRef]

[14] Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6465-6488). [CrossRef]

[15] Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., ... & Zhang, T. (2024, August). Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 10862-10878). [CrossRef]

[16] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86-92. [CrossRef]

[17] Wang, Y., Xian, G., & Zhao, R. (2025). Image-Text Multi-Modal Dataset of Corn Leaf Diseases based on Manual Annotation and Contrast Generation Model. *Journal of Agricultural Big Data, 7*(3), 371. [CrossRef]

[18] CHEN, L., SHI, B., & YUAN, Y. IDADP-Corn Disease Image Description Dataset. *China Scientific Data, 10*(4), 1-10. [CrossRef]

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

[20] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018, June). FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1* (Long Papers) (pp. 809-819). [CrossRef]

[21] Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning, 110*(3), 457-506. [CrossRef]

[22] Geifman, Y., & El-Yaniv, R. (2019). SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning* (PMLR 97, pp. 2151-2159). https://proceedings.mlr.press/v97/geifman19a.html

[23] Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4902-4912). [CrossRef]

[24] Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., de Lacalle, O. L., & Agirre, E. (2023, December). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 10776-10787). [CrossRef]

[25] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70. https://www.jstor.org/stable/4615733



Ning Zhang received her master's degree from Beijing University of Agriculture in 2021. She is currently engaged in research on large models and autonomous systems for agriculture at Beijing Paide Weiye Technology Development Co., Ltd. (Email: zhangning@nercita.org.cn)



Huarui Wu is a research professor and received his Ph.D. degree in computer application technology from Beijing University of Technology. He is currently Deputy Director of the Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences. His research interests include performance optimization and control theory for the agricultural Internet of Things (IoT), intelligent services for agricultural big data, unmanned vegetable farms and digital village technologies. (Email: wuhr@nercita.org.cn)



Huaji Zhu received his Ph.D. degree from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, in 2006 and subsequently joined the Beijing Agricultural Information Technology Research Center. He is currently a research professor at the Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences. His research interests include temporal geographic information systems (GIS), agricultural information technology and rural informatization applications. (Email: zhuhj@nercita.org.cn)



Yisheng Miao received his Ph.D. degree in computer science and technology from Beijing University of Technology in 2022. He joined the Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, in 2010. His research focuses on agricultural big data and artificial intelligence technologies and their applications. (Email: miaoys@nercita.org.cn)