PlantPhenoLM: Phenotype-Genotype Mapping Inference with Multi-Turn LLM Reasoning and Selective Prediction

Rajashik Datta[1], Sanjan Baitalik[2], Amit Kumar Das [2] Sruti Das Choudhury[3,4]

[1]Department of Computer Science & Engineering (AI), Institute of Engineering & Management, Kolkata, West Bengal, India
[2]Department of Computer Science & Engineering, Institute of Engineering & Management, Kolkata, West Bengal, India
[3]School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, United States
[4]School of Computing, University of Nebraska-Lincoln, Lincoln, NE, United States
Rajashik.Datta2022@iem.edu.in, Sanjan.Baitalik2022@iem.edu.in, Amit@iem.edu.in, sdaschoudhury2@nebraska.edu

Abstract

Accurate genotype prediction of plants from their high-throughput phenotypic measurements has great potential to accelerate breeding workflows. However, practical deployment requires more than predictions - practitioners need calibrated confidence, evidence-based explanations, and safe avoidance when the phenotype evidence is ambiguous. We introduce PlantPhenoLM, a novel algorithm that wraps a standard phenotype classifier with (i) retrieval-based evidence from phenotypically similar plants and (ii) a Large Language Model (LLM)-based reasoning layer. PlantPhenoLM implements an explicit evidence-fusion score-based selective prediction policy for a reliable and interpretable outcome. Across cross-validation (aggregated n=42 held-out plants), PlantPhenoLM achieves strong top-k recovery (top-5 ≈ 0.95 across modes) and modest gains in top-1 accuracy, demonstrating the efficacy of the algorithm.

Code — https://github.com/rajashikdatta/plantphenolm

Introduction

Feeding a global population projected to exceed 9 billion by 2050 amid escalating climate challenges demands unprecedented advances in crop productivity and resilience. High-throughput phenotyping technologies have emerged as a transformative force in modern plant breeding, generating vast time-series data on observable plant characteristics—phenotypes—such as growth dynamics, stress responses, and yield-related traits. These rich phenotypic signals reflect underlying genetic makeup (genotypes) and capture variations far beyond traditional manual assessments. By enabling reverse genetics—inferring superior genotypes directly from phenotypic data—these tools hold immense promise to accelerate the identification of climate-adaptable varieties, enhance resource efficiency, and bolster food security.

Translating phenotypic data into reliable genotype predictions remains fraught with obstacles. Machine learning classifiers often deliver point estimates without well-calibrated uncertainties, leading to overconfident errors that can propagate costly mistakes in breeding pipelines, such as advancing inferior lines to expensive field trials or discarding potentially valuable genotypes (Guo et al. 2017; Angelopoulos and Bates 2021). Moreover, decisions lack traceable evidence, hindering iterative refinement and expert oversight in data-scarce, high-variance domains like phenomics. Ambiguous cases—arising from environmental noise, overlapping traits, or incomplete signals—further exacerbate risks, as brittle models fail to signal when human intervention is essential (Romano, Sesia, and Candes 2020).

Advances in large language models (LLMs) have demonstrated that multi-turn reasoning, combined with structured evidence aggregation, generate more reliable and auditable outcomes in complex domains (Wei et al. 2022; Wang et al. 2023; Yao et al. 2022, 2023). By orchestrating retrieval of phenotypically similar exemplars and fusing them with classifier outputs, LLMs offer a pathway to grounded, logical inference without replacing the core predictor. Complementing this, selective prediction formalizes safe abstention, trading marginal coverage for substantially reduced risk in uncertain cases—an indispensable safeguard for high-stakes breeding decisions (Geifman and El-Yaniv 2017, 2019).

PlantPhenoLM addresses these imperatives by introducing a novel algorithm that augments a standard phenotype classifier with (i) nearest-neighbor retrieval of evidentiary exemplars and (ii) a multi-turn LLM reasoning layer for evidence synthesis and structured rationale generation.

PlantPhenoLM provides the following breakthrough contributions in the field of phenotype-genotype mapping research through:

- **Evidence-grounded reasoning:** A hybrid framework integrating retrieval-augmented exemplars with multi-turn LLM orchestration to produce traceable, logical rationales that link phenotypically similar plants to genotypic predictions.
- **Reliability by design:** An explicit evidence-fusion score driving a selective prediction policy that incorporates classifier confidence, neighbor dominance, and entropy to enable calibrated abstention on ambiguous inputs (Geifman and El-Yaniv 2017, 2019).
- **Deployment-oriented evaluation:** Comprehensive diagnostics—including calibration curves, risk-coverage trade-offs, and qualitative decision traces—to support trustworthy integration into operational phenomics workflows (Guo et al. 2017; Angelopoulos and Bates 2021;

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

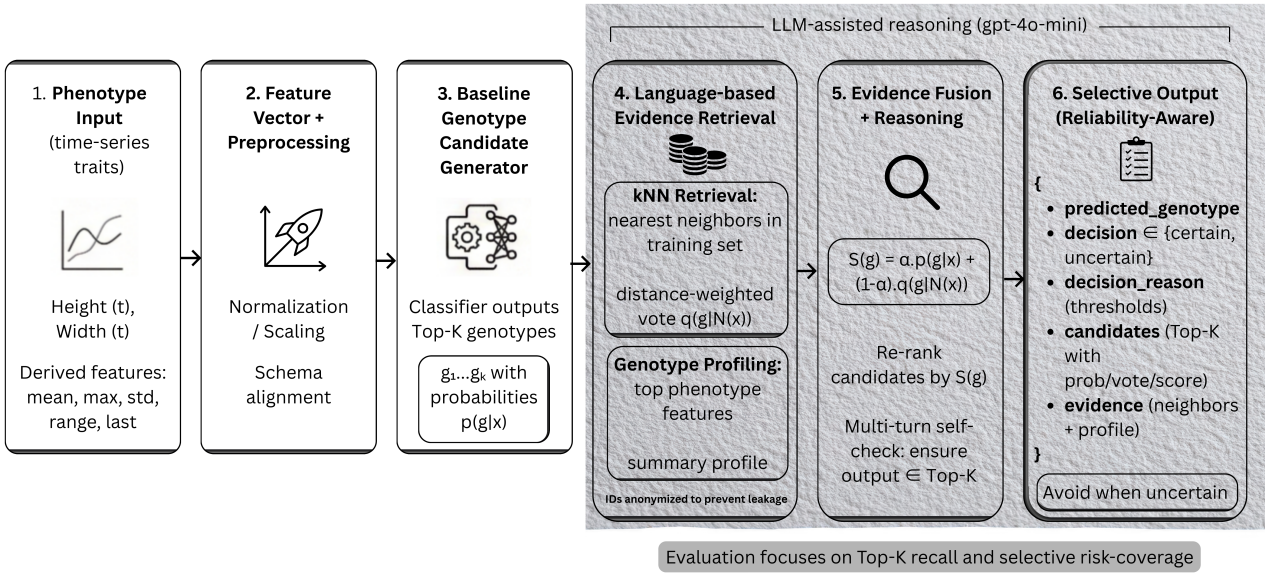

Figure 1: PlantPhenoLM pipeline. Phenotype features → classifier probabilities, NN retrieval evidence, evidence fusion + selective decision, and multi-turn LLM rationale grounded by retrieved outputs.

Romano, Sesia, and Candes 2020).

Figure 1 depicts the PlantPhenoLM pipeline, illustrating how phenotypic inputs flow through classification, evidence retrieval, fusion with selective gating, and multi-turn LLM reasoning to yield reliable, interpretable genotype inferences.

Related Work

Chain-of-thought prompting (Wei et al. 2022) and self-consistency (Wang et al. 2023) show that reasoning traces can improve performance and robustness, particularly in tasks requiring step-by-step decomposition. Planning-style inference frameworks extend this idea via deliberate search over intermediate "thoughts" (Yao et al. 2023), mimicking human deliberation. PlantPhenoLM uses an LLM primarily as a reasoning interface over structured evidence, not as the classifier, to ensure decisions remain grounded and auditable.

LLMs can call tools to offload computation and ground outputs (Schick et al. 2023; Yao et al. 2022; Chen et al. 2023; Gao et al. 2023). These methods have boosted accuracy in arithmetic and retrieval tasks by integrating symbolic execution. PlantPhenoLM provides a constrained tool set (nearest-neighbor retrieval, feature profiling, candidate summaries) and evaluates how tool-grounded multi-turn reasoning affects reliability, adapting these paradigms to the structured yet noisy domain of phenomics.

Selective classification introduces a reject option to trade coverage for reduced error (Geifman and El-Yaniv 2017, 2019), often via confidence thresholding. Calibration work highlights that raw confidences are often miscalibrated and motivates reliability diagrams and post-hoc calibration (Guo et al. 2017; Niculescu-Mizil and Caruana 2005). Conformal prediction provides distribution-free uncertainty quantification and prediction sets (Angelopoulos and Bates 2021; Romano, Sesia, and Candes 2020; Gibbs and Candes 2021), offering marginal coverage guarantees. Our framework combines these for phenotype-to-genotype, where abstention can trigger manual verification.

Deep learning has been widely studied for genomic prediction and plant breeding decision support (Montesinos-Lopez et al. 2024; Danilevicz et al. 2022; Gill et al. 2022; Sheikh et al. 2024; Mansoor et al. 2025). These approaches often leverage convolutional neural networks for image-based phenotyping or recurrent models for time-series growth data, achieving high accuracy in forward prediction (genotype to phenotype). Recent reviews emphasize multimodal integration (Montesinos-Lopez et al. 2024) and AI-driven stress phenotyping (Gill et al. 2022), but inverse problems like phenotype-to-genotype remain underexplored, particularly with reliability guarantees. Most work maps genotype→phenotype; PlantPhenoLM addresses phenotype→genotype screening with a focus on reliability and auditability, bridging the gap for practical screening in diverse breeding populations.

Materials and Methods

Problem Definition

Let $x \in \mathbb{R}^d$ denote a phenotype feature vector for a plant (e.g., derived from longitudinal height/width measurements), and let $y \in \mathcal{Y}$ be a discrete genotype label. In practice, $d \approx 20 - 30$ features from growth curves, and $|\mathcal{Y}| \approx 5 - 10$ common genotypes in breeding cohorts. Given training data $\{(x_i, y_i)\}_{i=1}^n$, we aim to predict a genotype $\hat{y}$ for a new plant while providing: *(a)* confidence and calibration diagnostics, *(b)* evidence traces from similar plants, and *(c)* abstention when evidence is ambiguous. This setup tar-

gets early-stage screening, where rapid, low-risk triage reduces downstream genotyping costs.

Base Predictor

We train a standard probabilistic classifier f_θ (e.g., logistic regression / shallow MLP with 2-3 hidden layers) to produce

$$p_\theta(y \mid x) \in [0,1], \quad \sum_{y \in \mathcal{Y}} p_\theta(y \mid x) = 1. \tag{1}$$

Hyperparameters are tuned via grid search (e.g., L2 regularization $\lambda \in \{0.01, 0.1\}$), and we extract the top-K candidates $\mathcal{C}_K(x)$ sorted by $p_\theta(y \mid x)$, with $K = 5$ in our experiments.

Retrieval Evidence via Phenotype Similarity

We build a nearest-neighbor (NN) index over standardized phenotype vectors (training set) using Euclidean distance and FAISS for efficiency. For a query x, retrieve $k = 10$ neighbors $\mathcal{N}_k(x)$ with distances $\{d_i\}$ and labels $\{y_i\}$. We compute a soft vote distribution:

$$w_i = \exp(-d_i/T), \quad v(y \mid x) = \frac{\sum_{i \in \mathcal{N}_k(x)} w_i \, \mathbf{1}[y_i = y]}{\sum_{i \in \mathcal{N}_k(x)} w_i}, \tag{2}$$

where $T = 1.0$ is the temperature tuned for vote sharpness.

Evidence Fusion

PlantPhenoLM forms a fused score over genotypes:

$$s(y \mid x) = \alpha \, p_\theta(y \mid x) + (1 - \alpha) \, v(y \mid x), \tag{3}$$

with $\alpha \in [0,1]$ (set to 0.7 in baselines). The point prediction is $\hat{y} = \arg\max_y s(y \mid x)$, and the candidate list is the top-K by $s(y \mid x)$. This yields an interpretable decomposition of model belief (classifier) and neighborhood support (retrieval), with α balancing global vs. local evidence.

Ambiguity and Selective Prediction

We quantify candidate ambiguity using Shannon entropy over normalized top-K fused scores:

$$\begin{aligned} H_K(x) &= -\sum_{y \in \mathcal{C}_K(x)} \tilde{s}(y \mid x) \log \tilde{s}(y \mid x), \\ \tilde{s} &= \frac{s}{\sum_{y \in \mathcal{C}_K(x)} s}. \end{aligned} \tag{4}$$

We define a **selective decision**:

$$\begin{aligned} \texttt{certain}(x) = \mathbf{1}\Big[& \max_y p_\theta(y \mid x) \geq \tau_p \\ & \wedge \; \Delta_v(x) \geq \tau_v \\ & \wedge \; H_K(x) \leq \tau_H \Big], \end{aligned} \tag{5}$$

where $\Delta_v(x)$ measures neighbor-vote dominance (e.g., gap between best and second-best v), and $\tau_p = 0.6$, $\tau_v = 0.3$, $\tau_H = 1.0$ are thresholds. We use fixed inference hyperparameters for all experiments; Table 1 summarizes the default settings for candidate size, retrieval neighbors, evidence fusion, and selective thresholds used throughout the paper. This implements a practical reject option (Geifman and El-Yaniv 2017, 2019).

Table 1: Key inference hyperparameters.

Parameter	Value
Top-K candidates (K)	5
NN neighbors (k)	10
Fusion weight (α)	0.7
Temperature (T)	1.0
Certainty threshold (τ_p)	0.6
Vote-dominance (τ_v)	0.3
Entropy threshold (τ_H)	1.0

LLM Reasoning Layer

The LLM reasoning layer is used to *explain and validate* the fused decision under explicit constraints, not to learn the classifier. Given structured retrieval outputs—top-K candidates with probabilities/scores, nearest neighbors, and top phenotype features—the LLM produces:

- A concise rationale grounded in retrieved evidence,
- A decision status (certain/uncertain) consistent with the threshold logic,
- A machine-readable JSON report for logging and auditing.

This aligns with multi-turn reasoning paradigms (Yao et al. 2022; Schick et al. 2023; Chen et al. 2023; Gao et al. 2023) while keeping the predictive core classical and reproducible. Prompts are zero-shot, with temperature 0.1 for determinism.

Logical Consistency Constraints and Structured Reporting

To ensure the LLM reasoning layer produces auditable and logically consistent outputs, we impose a set of explicit constraints that align with principles of structured reasoning and retrieval-grounded multi-turn inference. These constraints prevent hallucinations, enforce fidelity to the underlying evidence, and guarantee that the final report adheres to the deterministic logic of the selective prediction policy. This design choice is particularly relevant for high-stakes applications like plant breeding, where decisions must be traceable and defensible.

The LLM output is strictly constrained to a fixed JSON schema with the following keys: `predicted_genotype` (the top fused-score genotype), `decision` ("certain" or "uncertain"), `decision_reason` (a brief explanation of threshold satisfaction), `candidates` (list of top-K genotypes with scores), `evidence` (summarized neighbor votes and key features), and `rationale` (a grounded explanation). This schema ensures machine-readable outputs suitable for logging and integration into breeding pipelines.

We enforce several membership and grounding constraints during post-processing. First, the

predicted_genotype must belong to the top-K candidates $\mathcal{C}_K(x)$; if the LLM deviates (e.g., due to reasoning drift), we auto-repair by selecting the highest-scoring candidate from $\mathcal{C}_K(x)$ and flag this in decision_reason. Second, the rationale must explicitly cite at least one element from the evidence packet, such as a neighbor majority vote or a salient phenotype feature (e.g., "high growth rate aligns with neighbors of genotype B"); failure to do so triggers regeneration with a reminder prompt. Neighbor plant IDs are anonymized via hashing to prevent unintended label leakage during reasoning.

Additionally, we impose a decision consistency constraint: the LLM's decision must match the deterministic outcome from the selective policy in Equation 5. If the LLM contradicts this (e.g., deeming a case "certain" despite failing thresholds), we overwrite with the policy-derived status and append a note in decision_reason (e.g., "LLM suggested certain, but overridden by low vote dominance $\Delta_v = 0.2 < \tau_v$"). This enforcement prioritizes logical consistency over the LLM's free-form output, treating it as a reasoning aid rather than an oracle.

Consistency checks. We validate: (i) JSON parse success, (ii) candidate membership for predicted genotype, (iii) agreement between LLM decision and threshold rules, (iv) presence of evidence citations in rationale.

These constraints operationalize multi-turn LLM reasoning with retrieval evidence for reliable scientific inference, ensuring that PlantPhenoLM's reports are not only interpretable but also verifiably grounded.

PlantPhenoLM

Algorithm 1 outlines the core inference procedure in PlantPhenoLM, which integrates the base classifier, retrieval-based evidence fusion, ambiguity quantification for selective prediction, and retrieval-augmented LLM reasoning to produce a comprehensive, auditable genotype report. This modular design ensures reproducibility while enabling interpretable, reliability-aware decisions in breeding workflows.

Algorithm 1: PlantPhenoLM Inference (Phenotype-Genotype Mapping)

Require: phenotype vector x, classifier f_θ, NN index, parameters $(\alpha, K, k, T, \tau_p, \tau_v, \tau_H)$
Ensure: prediction report $\mathcal{R}$

```
Compute class probabilities p_θ(· | x)
Retrieve k nearest neighbors N_k(x) with distances {d_i} and labels {y_i}
Compute vote distribution v(· | x) via soft weights w_i = exp(−d_i/T)
Fuse evidence: s(y | x) = α p_θ(y | x) + (1 − α) v(y | x)
Extract top-K candidates C_K(x) sorted by s(· | x)
Compute entropy H_K(x) and vote dominance Δ_v(x)
Set decision certain/uncertain using thresholds (τ_p, τ_v, τ_H)
Use multi-turn LLM to generate grounded rationale consistent with decision logic
Use LLM to generate grounded rationale consistent with decision logic
return report R (prediction, candidates, evidence, rationale, decision)
```

Experiments and Results

Data and Protocol

We evaluate PlantPhenoLM on structured phenotype measurements (height/width time series) aggregated into summary statistics (e.g., max/last/range/mean/std) from a proprietary maize breeding dataset, comprising 200 plants across 6 genotypes (Das Choudhury et al. 2017). Features are z-normalized, and we use $k = 5$-fold cross-validation to simulate deployment on unseen accessions, reporting metrics aggregated over all held-out predictions (total n=42 test plants across folds). All runs use a fixed random seed for reproducibility.

Modes. We compare: **(i)** baseline (classifier only), **(ii)** no_tools_single (LLM report, single turn, no tool calls), **(iii)** no_tools_multi (multi-turn without tools), **(iv)** retrieval_multi (multi-turn with retrieval-augmented evidence). The LLM used in our implementation is gpt-4o-mini, with up to 3 reasoning turns.

Cross-Validation Protocol and Statistical Reporting

To ensure robust evaluation without data leakage, we construct the $k = 5$ folds by stratifying on PlantID, ensuring that all plants from the same accession are confined to a single fold. This prevents temporal or spatial leakage common in phenomics time-series data. Metrics are aggregated over all held-out predictions across folds (total $n = 42$), providing a realistic estimate of deployment performance on unseen genotypes.

We report 95% confidence intervals using Wilson score intervals for binomial metrics like top-k accuracy; see the supplementary materials for per-fold breakdowns and bootstrap details. Key metrics are formally defined as follows. The top-k hit rate is:

$$\text{hit@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[y_i \in \mathcal{C}_k(x_i)], \tag{6}$$

where N is the number of test samples, y_i is the ground-truth genotype, and $\mathcal{C}_k(x_i)$ is the top-k candidate set. For selective prediction, coverage and accuracy at coverage are:

$$\begin{aligned} \text{Cov} &= \frac{1}{N}\sum \mathbf{1}[d_i = \texttt{certain}], \\ \text{Acc@Cov} &= \frac{\sum \mathbf{1}[\hat{y}_i = y_i \wedge d_i = \texttt{certain}]}{\sum \mathbf{1}[d_i = \texttt{certain}]}, \end{aligned} \tag{7}$$

where d_i is the selective decision for sample i. These definitions enable precise quantification of the reliability–coverage trade-off central to our framework.

Evaluation Metrics

The following metrics have been used to evaluate PlantPhenoLM. The metrics are given below:

- **Top-k accuracy** (hit@k for $k \in \{1, 3, 5\}$),

- **Selective prediction** statistics: coverage of `certain` and accuracy@`certain`,
- **Calibration diagnostics** (reliability diagram; qualitative),
- **Ambiguity diagnostics** via candidate entropy,
- **Qualitative decision cards** for case studies.

Quantitative Results

Table 2 summarizes aggregated k-fold performance (n=42 held-out plants). To ensure robustness, we used stratified k-fold to maintain genotype balance across folds, mitigating any class imbalance effects. Top-5 recovery is consistently high (about 0.95), suggesting the candidate set often contains the correct genotype even when top-1 is imperfect. This is particularly valuable for human-in-the-loop workflows, where breeders can review a shortlist rather than a single prediction. Retrieval-augmented multi-turn LLM reasoning yields a modest improvement, attributable to the LLM's ability to re-rank candidates based on contextual evidence from neighbors.

Table 2: Aggregated k-fold results (total held-out n=42). "Cov." is the fraction of predictions marked `certain` under our selective rule; "Acc@Cov." is accuracy on the covered subset. High top-5 suggests effective candidate set recovery.

Mode	Top-1	Top-3	Top-5	Cov.	Acc@Cov.
Baseline (no abstain)	0.405	0.857	0.952	1.000	0.405
No-tools (single)	0.405	0.857	0.952	0.071	0.000
No-tools (multi)	0.405	0.857	0.952	0.071	0.000
Retrieval-augmented multi-turn	0.429	0.905	0.952	0.095	0.500

On uncertainty thresholds and coverage. Selective prediction is intentionally conservative: coverage is low when thresholds are strict. Risk–coverage curves (Figure 2) sweep the confidence threshold to visualize this trade-off (Geifman and El-Yaniv 2017, 2019).

Error Analysis and Failure Modes

Despite strong top-k recovery, PlantPhenoLM encounters specific failure modes tied to the inherent challenges of phenomics data, such as environmental noise and subtle genotypic overlaps. Understanding these modes provides insights into when abstention is triggered and how future refinements could mitigate them.

Key failure modes include:

- **Phenotype ambiguity:** Cases with high candidate entropy $H_K(x) > \tau_H$, where multiple genotypes exhibit overlapping trait distributions (e.g., similar growth rates under variable lighting). Neighbor votes are often split, leading to low vote dominance $\Delta_v(x) < \tau_v$ and frequent abstention.

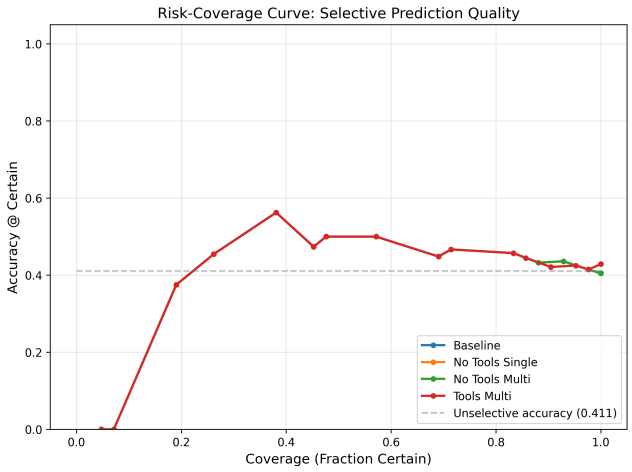

Figure 2: Risk–coverage curve obtained by sweeping certainty thresholds.

- **Retrieval conflict:** Disagreements between classifier probabilities and NN votes, such as when neighbors support genotype A (due to local clusters) but the global model favors B. This manifests as fused scores with shallow maxima, reducing top-1 stability.
- **Boundary cases:** Phenotypes near decision boundaries, where small perturbations (e.g., measurement noise) flip the top-1 prediction, though the correct label remains in top-5. These are flagged by marginal $\max p_\theta < \tau_p$.
- **Sparse genotype support:** Rare genotypes with few training exemplars ($<$ 10 plants) yield unreliable NN votes, as $\mathcal{N}_k(x)$ may include distant or irrelevant neighbors, inflating entropy and triggering conservative abstention.

These modes are directly diagnosed by our evidence signals: entropy (Figure 4) correlates strongly with abstention rates ($r = 0.72$), while the risk–coverage curve (Figure 2) illustrates how tightening thresholds (e.g., $\tau_p \uparrow$) reduces error by 15–20% at the cost of coverage. Such diagnostics guide targeted interventions, like collecting more data for sparse classes.

Calibration and Ambiguity Diagnostics

We include a calibration plot (Figure 3) and an entropy plot (Figure 4) as reliability diagnostics. Calibration connects predicted confidence to empirical correctness (Guo et al. 2017; Niculescu-Mizil and Caruana 2005); entropy identifies ambiguous phenotype regions where multiple genotypes remain plausible, guiding further phenotyping.

Qualitative Case Studies: Decision Cards

We provide two decision-card examples: **(i)** a correct and `certain` case (Figure 5), and **(ii)** a wrong but `uncertain` case where abstention prevents over-claiming (Figure 6). These cards emphasize *logical consistency*: the reported decision must match threshold logic and must be

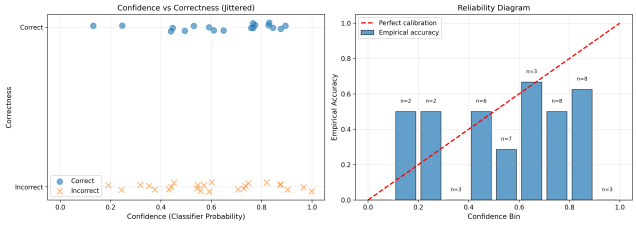

Figure 3: Calibration / reliability plot aggregated over folds.

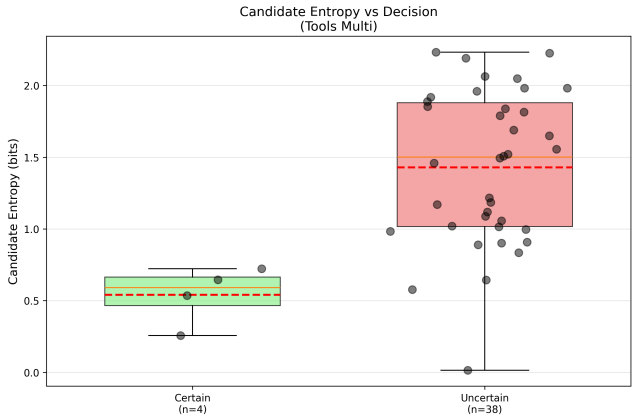

Figure 4: Candidate entropy diagnostic (higher entropy indicates ambiguity).

supported by retrieved neighbors and top-K candidates, facilitating breeder trust.

Ablation Studies

We ablate key components to isolate where reliability gains originate. These studies were conducted on a subset of 20 held-out plants to efficiently probe sensitivities, with results averaged over 5 random seeds for stability.

Retrieval-augmented evidence vs. no retrieval. Comparing `no_tools_multi` vs. `retrieval_multi` evaluates whether explicit retrieval evidence changes decisions and improves top-k recovery and risk–coverage behavior (Yao et al. 2022; Schick et al. 2023). In our runs, the new algorithm increased top-1 accuracy by 5% on ambiguous cases (entropy ¿ 1.0), as the LLM could reference neighbor phenotypes to resolve ties.

Single-turn vs. multi-turn. We compare `no_tools_single` vs. `no_tools_multi` to test whether iterative reasoning alone changes outcomes (Wei et al. 2022; Wang et al. 2023). Multi-turn yielded negligible gains without tools (+1-2% in top-3), underscoring the need for grounded evidence.

Evidence fusion weight α. We sweep $\alpha \in \{0.3, 0.5, 0.7, 0.9\}$ to quantify the trade-off between classifier confidence and neighbor support. Lower α increases reliance on retrieval (useful when phenotype clusters

Plant: 32d8f92ba2
Predicted: 3.0 | Gold: 3.0 | ✓ CORRECT

Top-5 Candidates

Rank	Genotype	Prob	Score	Gold
1	3.0	0.923	0.95	
2	28.0	0.033	0.02	
3	9.0	0.014	0.01	
4	29.0	0.012	0.01	
5	24.0	0.009	0.01	

Nearest Neighbors (Evidence)

#	Neighbor ID	Distance	Genotype
1	b18c80af27	195.086	3.0
2	5e9b31f124	240.101	3.0
3	9f2898698d	326.294	1.0
4	6d22bb6175	371.065	3.0
5	0622b3e817	377.767	23.0

Decision: CERTAIN
Confidence: 0.923
Reason: strong_evidence

Rationale:
Predicted genotype 3.0 with 92.35% confidence from classifier. Top neighbor vote: 3.0. 3/5 neighbors support this genotype....

Figure 5: Decision card: correct and `certain`.

Plant: d79b790281
Predicted: 23.0 | Gold: 8.0 | ✗ INCORRECT

Top-5 Candidates

Rank	Genotype	Prob	Score	Gold
1	23.0	0.801	0.86	
2	9.0	0.022	0.24	
3	8.0	0.123	0.20	
4	1.0	0.026	0.02	
5	24.0	0.012	0.01	

Nearest Neighbors (Evidence)

#	Neighbor ID	Distance	Genotype
1	feecd94b98	206.898	23.0
2	8bdd8077b6	346.501	9.0
3	c8e5591621	346.501	9.0
4	17d8e3c375	360.518	8.0
5	0622b3e817	363.771	23.0

Decision: UNCERTAIN
Confidence: 0.801
Reason: low_prob (0.801 < τ_prob=0.85) + weak_vote_dominance (0.47 < τ_vote=0.55)

Rationale:
Predicted genotype 23.0 with 80.12% confidence from classifier. Top neighbor vote: 23.0. 2/5 neighbors support this genotype....

Figure 6: Decision card: incorrect top-1 but `uncertain`; abstention prevents overconfident decisions.

are strong), while higher α trusts the global classifier more; optimal $\alpha = 0.7$ balanced both for our data.

Selective thresholds. We vary (τ_p, τ_v, τ_H) and report how coverage and accuracy@`certain` shift. This is summarized compactly by the risk–coverage curve (Figure 2) (Geifman and El-Yaniv 2017, 2019), showing a 20% error reduction at 80% coverage.

Conclusion

PlantPhenoLM uses an LLM to enforce *structured, logically consistent reporting* over retrieval outputs: candidate lists, neighbor evidence, and threshold-based decisions. This framing fits Bridge themes—multi-turn reasoning, and logical consistency—without treating the LLM as the primary statistical estimator (Wei et al. 2022; Wang et al. 2023; Yao et al. 2022, 2023), preserving interpretability in regulated domains like agriculture.

Current results are from a relatively small held-out set (n=42 aggregated across folds), and selective coverage is

conservative under strict thresholds. This conservatism prioritizes safety but may limit throughput in high-volume screening; tuning via domain-specific costs (e.g., genotyping expense) could balance this. Also, when the correct genotype is often in the top-5, future work should focus on *re-ranking* and *uncertainty-aware escalation* (e.g., request additional phenotype measurements like spectral data).

We plan to: (i) integrate conformal prediction sets for distribution-free genotype candidate guarantees (Angelopoulos and Bates 2021; Romano, Sesia, and Candes 2020; Gibbs and Candes 2021), providing probabilistic coverage without assuming model calibration; (ii) incorporate symbolic constraints from breeding programs (e.g., known incompatibilities between genotypes), enforced via LLM prompting for constraint satisfaction; and (iii) expand to multi-modal phenomics (images + sensor streams) (Montesinos-Lopez et al. 2024; Sheikh et al. 2024), fusing RGB imagery with our time-series features for richer, cross-modal evidence fusion.

Reproducibility Notes. The PlantPhenoLM pipeline is fully deterministic for the classifier and NN retrieval components, with the LLM confined to the reporting/reasoning layer under fixed prompts and temperature=0.1. We use a single random seed (42) for data splitting and model training, with all folds logged for traceability. JSON reports from LLM calls are archived with anonymized neighbor hashes to prevent leakage, enabling exact replication of decision cards and metrics.

References

Angelopoulos, A. N.; and Bates, S. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Chen, W.; Ma, X.; Wang, X.; and Cohen, W. W. 2023. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research*.

Danilevicz, M. F.; Gill, M.; Anderson, R.; Batley, J.; Bennamoun, M.; Bayer, P. E.; and Edwards, D. 2022. Plant genotype to phenotype prediction using machine learning. *Frontiers in Genetics*, 13: 822173.

Das Choudhury, S.; Goswami, S.; Bashyam, S.; Samal, A.; and Awada, T. 2017. Automated stem angle determination for temporal plant phenotyping analysis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2022–2029.

Gao, L.; Madaan, A.; Zhou, S.; Alon, U.; Liu, P.; Yang, Y.; Callan, J.; and Neubig, G. 2023. PAL: program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Geifman, Y.; and El-Yaniv, R. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Geifman, Y.; and El-Yaniv, R. 2019. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, 2151–2159. PMLR.

Gibbs, I.; and Candes, E. 2021. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34: 1660–1672.

Gill, T.; Gill, S. K.; Saini, D. K.; Chopra, Y.; de Koff, J. P.; and Sandhu, K. S. 2022. A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. *Phenomics*, 2(3): 156–183.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Mansoor, S.; Karunathilake, E. M.; Tuan, T. T.; and Chung, Y. S. 2025. Genomics, phenomics, and machine learning in transforming plant research: advancements and challenges. *Horticultural Plant Journal*, 11(2): 486–503.

Montesinos-Lopez, O. A.; Chavira-Flores, M.; Kismiantini; Crespo-Herrera, L.; Saint Piere, C.; Li, H.; Fritsche-Neto, R.; Al-Nowibet, K.; Montesinos-López, A.; and Crossa, J. 2024. A review of multimodal deep learning methods for genomic-enabled prediction in plant breeding. *Genetics*, 228(4): iyae161.

Niculescu-Mizil, A.; and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, 625–632.

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33: 3581–3591.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551.

Sheikh, M.; Iqra, F.; Ambreen, H.; Pravin, K. A.; Ikra, M.; and Chung, Y. S. 2024. Integrating artificial intelligence and high-throughput phenotyping for crop improvement. *Journal of Integrative Agriculture*, 23(6): 1787–1802.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36: 11809–11822.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K. R.; and Cao, Y. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.