

Tracking AI Advisory Drift Before It Reaches the Farms: Measuring Contextual Reliability in LLM-Generated Advice across the Agricultural Landscape

Jerry H. Martin, Tianyi Li

School of Applied and Creative Computing
Purdue University
West Lafayette, IN 47907, USA
{mart2202,li4251}@purdue.edu

Abstract

Agricultural advice intended to protect the land is sensitive to local context because a best management practice (BMP) designed for one location might be risky elsewhere. This risk is addressed in the U.S. by the Cooperative Extension experts stationed in different geographic regions who can benefit from the ability of general-purpose large language models (LLMs) to produce plausible answers quickly at a low marginal cost; however, these answers must remain locally reliable.

We evaluated local reliability by examining whether advisory behaviors in LLM responses repeatedly conformed to relevant BMPs. We analyzed this conformance in 288 LLM responses to eight publicly available Cooperative Extension inquiries. These inquiries were submitted by people in different locations who were concerned that local water could harm plants. Because the locations differed in plant types, water composition, soils, and how the inquiry was framed, responses from twelve LLMs at three low-temperature settings allowed us to examine the potential for unnecessary risk. We analyzed this risk using descriptive text metrics and an inductively developed advisory scorecard to identify BMP nonconformances, including unsupported claims, omissions, uncorrected naïve question framing, and contextually misframed solutions.

To characterize the risks in context-sensitive advice, we introduce *contextual advisory reliability* as a scorecard-bound diagnostic metric for assessing whether LLM-generated land-management advice consistently protects against risk. Risk is screened by evaluating repeated responses to calculate contextual advisory reliability. The ability to quantify reliability across different geographical regions supports responsible deployment in public advisory domains by helping institutions audit the suitability of AI-generated advice and communicate those results to institutional experts, land managers, and LLM designers.

Introduction

High-quality land management advice must account for site-specific differences in microclimate, soil chemistry, and resources, because protective advice in one field can be risky advice in a neighboring one (Kahan 2008). Matching protective advice to the field is especially important for households

with limited economic margins, which struggle to absorb or reverse the consequences of high-risk land management advice (World Food Programme and Food Security Information Network 2023). Even infrequent risky advice, when repeated across many decisions and many landscapes, can produce consequences that accumulate into lasting effects on crop production, water use, and regional prosperity. As human populations grow, land managers and supporting institutions will face increasing pressure to reduce errors in land management reasoning.

Traditionally, the challenge of minimizing local errors in land management reasoning in the U.S. has been handled by the Cooperative Extension system. This system is a network of geographically distributed institutions whose experts in soils, water, crops, infrastructure, and related disciplines advise local land managers in their own regions. This locally situated expertise provides a mechanism for validating, contextualizing, and filtering advice before it is put into practice. Land managers who put into practice plausible advice from a general-purpose LLM need an equivalent mechanism, or errors across repeated advice may compound into avoidable sociotechnical risk.

We evaluate repeated LLM responses to variations of one common land-management question: Will my water harm my plants? It is a useful test case because the answer depends on local soil and water chemistry, plant hardiness, established irrigation practices, and how the question is framed locally. Drawing questions from real submissions to Extension experts in different parts of the U.S. lets us ask whether LLM guidance is protective or risky.

Risky guidance can come from hallucination (Ji et al. 2023; Maynez et al. 2020), but the two are not the same. A trivial or even absurd hallucination may pose little risk to the land, while a hallucination-free recommendation may still be risky because it can omit a BMP, accept a false premise embedded in the question, or apply a valid solution to the wrong land-management context. These errors may stay invisible to advice seekers, who often ask questions precisely because they lack the expertise to catch errors in the answers. An Extension expert is unlikely to pose the same question to an LLM, but would likely recognize errors and BMP nonconformances that the advice seeker cannot. To quantify this expert-observable, context-sensitive risk, we introduce the concept of *contextual advisory reliability*.

We define *contextual advisory reliability* as expert-observable reliability within a specified advisory scope, assessed with an inductively derived scorecard that evaluates whether advice exhibits protective behaviors and avoids high-risk advisory errors according to relevant BMPs. Reliability therefore reflects the consistent absence of known, scorecard-defined risks rather than conformity to any particular response. Advice may vary within and across contexts without penalty simply for differing from a reference response, but it is penalized when it introduces known risks. We examine *contextual advisory reliability* through three research questions:

- RQ1: What recurring protective and risky advisory behaviors appear in LLM-generated responses to context-sensitive agricultural questions?
- RQ2: How prevalent are these advisory behaviors in context-sensitive LLM responses generated at low-temperature settings using models accessible to land managers?
- RQ3: What kinds of advisory failures are missed by descriptive text metrics or hallucination-centered evaluation?

We used eight questions submitted to U.S. Cooperative Extension experts between 2013 and 2024 to prompt 12 LLMs at three low-temperature settings, yielding 288 responses. We analyzed these responses using descriptive textual metrics (word count, Flesch-Kincaid reading level, vocabulary rarity, and lexical similarity between each inquiry and its response) as well as an advisory scorecard designed to encode the presence or absence of protective and risky advisory behaviors.

Advisory risk appears both within individual responses and across repeated responses. Within a response, it takes four forms: unsupported recommendations, omitted qualifications, misframing of the advisory question, and premature application of context-dependent guidance. Across repeated generations, advisory behavior was sometimes consistent and at other times unstable, resembling guess-like selection. The descriptive text metrics characterize surface form but do not capture scorecard-coded failures.

We make three contributions. First, we use a case study of 288 responses to eight real U.S. Cooperative Extension questions concerning irrigation water suspected of harming plants to provide empirical evidence of advisory failures in a problem highly sensitive to local context. Second, we show how to structure a scorecard around unsupported recommendations, omitted qualifications, misframing of the advisory question, and premature application of context-dependent guidance. Third, we introduce a diagnostic measure of contextual advisory reliability for evaluating the consistent absence of known, scorecard-defined risks within a specified advisory scope.

Background

Risks of Context-Insensitive LLM Advice for Land Management Land managers uncertain about an agricultural decision can get on-demand advice from an LLM (Tzachor

et al. 2023), potentially bypassing the local infrastructure built to put the right humans in the loop. Such use requires guardrails because LLMs tend to optimize linguistic plausibility over truth (Bender et al. 2021). Mismatches between plausibility and truth are often analyzed as hallucination or confabulation (Maynez et al. 2020; Ji et al. 2023), including false claims or true claims presented in a misleading form that distort advisory information (Aïmeur, Brassard, and Sallami 2023). Such distortions should be identified and filtered before land managers act on them and incur crop losses, soil degradation, or economic losses.

Human-in-the-Loop Infrastructure for Filtering Agromomic Risk Before LLMs Existed Long before LLMs, agricultural communities relied on agricultural societies to judge the local applicability of new farming practices through fairs and contests (Watson 1820). Prize-winning crops, livestock, and implements were placed on public display, making local results visible, comparable, and open to community judgment, and those judgments then circulated through informal social networks that filtered the local applicability of a practice before it spread.

As Extension pioneer Seaman Knapp observed, “the mere dissemination of printed information sometimes does not accomplish this result”: new agricultural knowledge must also be demonstrated under local field conditions before farmers can confidently apply it in decisions affecting their livelihoods (Martin 1921).

The importance of social trust in this process is evident in studies of hybrid seed-corn adoption (Ryan and Gross 1943) that Everett Rogers later drew on in developing the diffusion of innovations framework (Rogers 1962). Agricultural innovations, such as higher-yielding hybrid corn, spread not only through technical information but through local demonstration, trusted relationships, and observation of results. Designers of LLM-based agricultural systems should therefore preserve these trust-building functions rather than allow the distribution of information to bypass local evaluation.

Agromomic Risks When LLMs Can Bypass Humans in the Loop The risks of deploying AI in agriculture include misinformation, erosion of shared digital agromomic knowledge, and over-reliance on artificial intelligence (De Clercq et al. 2024). An advisory tool deployed in Utah illustrates the problem by confidently suggesting treatments for Peach Leaf Curl, a disease that rarely appears in the region’s arid climate (Hill and Narine 2026). An LLM-based system evaluated at Wisconsin-Madison generated soybean management plans that omitted key management steps (Mourtzinis et al. 2026), and a tool tested with irrigated-rice farmers in Nigeria gave inaccurate advice about planting dates, seed rates, and fertilizer management (Ibrahim, Senthilkumar, and Saito 2024). The concerns raised in these studies echo a long history predating AI, in which agricultural communities already relied on humans in the loop, serving in advisory roles to filter new knowledge through local observation, contextual adaptation, social trust, and institutional review before putting new recommendations into practice.

Reliability Through Repetition An advice-generating system earns trust when avoidance of known advisory risks is demonstrated through repetition. In experimentation, repetition tests whether a claim contains error, while repeating the same message tests whether the communication channel introduces error. A message-based advisory system therefore depends on both reliable claims and a reliable communication channel.

Experimental repetition has been formalized since Fisher’s work at Rothamsted, where variation within agricultural land obscured the causal interpretation of isolated results (Fisher 1925, 1935). Subsequent work identified Type I errors (detecting an absent effect), Type II errors (missing a present effect) (Neyman and Pearson 1928, 1933), Type III errors (“giving the right answer to the wrong problem”) (Kimball 1957), and Type IV errors (“the incorrect interpretation of a correctly rejected hypothesis”) (Marascuilo and Levin 1970). These errors show how repeated results can appear sound yet remain operationally misleading.

In LLM-generated agricultural advice, experimental errors can manifest as advisory response errors. An LLM may assert a claim unsupported by local evidence (Type I), omit guidance that a locally informed advisor would provide (Type II), answer a question a locally informed advisor could identify as the wrong question (Type III), or apply otherwise valid knowledge before the relevant local conditions have been established (Type IV).

Type II errors are especially difficult to detect because of the file drawer problem: null findings are less likely to be written up or submitted for publication. Models trained on the published record may therefore have limited access to the evidence needed to avoid repeating Type II errors (Moniz, Druckman, and Freese 2025; Lee 2026).

Extension experts have traditionally delivered guidance through conventional channels, such as printed bulletins, phone lines, and static web pages, where the message is expected to remain fixed in transmission. LLM-based systems differ because repeating fixed prompts can produce different yet fluent responses, some of which may introduce errors. In conventional communication channels, the potential for message error can be characterized using entropy (Shannon 1948). However, introducing an LLM turns a conventional channel into a generative channel in which variation does not necessarily introduce error, so a different diagnostic tool is needed to identify errors that introduce advisory risk.

Methods

This study uses a repeated, cross-model advisory evaluation design to examine scorecard-defined risk in LLM responses to context-sensitive agricultural questions.

Question Selection and Pre-processing

The Cooperative Extension platform “Ask Extension”¹ hosts publicly searchable questions submitted by users from across the U.S. and answered by agricultural experts. We searched the platform for questions about the same underlying advisory problem, “*Will my water harm my plants?*”

¹<https://ask.extension.org/>

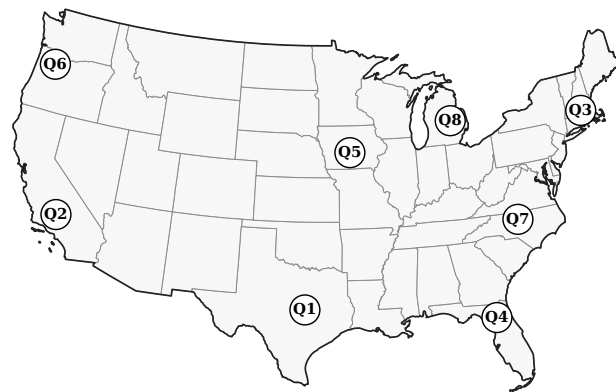


Figure 1: Geographic origins of questions with labels corresponding to Table 1.

Using one advisory problem allows responses to be evaluated across similar but non-identical situations. In each situation, the asker’s location helps determine the relevant context, alongside soil, water, plant, and infrastructure conditions. For each context-sensitive question, we obtained multiple LLM responses to evaluate scorecard-defined advisory risk.

When selecting questions, we included those in which the asker was assessing whether water would harm their plants. We excluded questions that did not include both a water and a plant concern or for which a similar question from the same region had already been selected. Figure 1 shows where the questions came from, and Table 1 provides their contextual details. We removed any personal names before submitting questions to the LLMs. For multi-turn exchanges, we retained only the initial inquiry because subsequent user replies could not be replicated for LLM responses.

The original Extension answers are treated as institutionally situated reference responses rather than infallible ground truth. Extension advisors often apply best management practices codified in NRCS Conservation Practice Standards, such as the standard for Irrigation Water Management (Natural Resources Conservation Service 2020), and the Irrigation Guide within the National Engineering Handbook (Natural Resources Conservation Service 1997). These standards provide the basis for identifying where advisory practices may introduce or mitigate risk. We therefore use relevant BMPs as the baseline for identifying risk and the Extension answers as context-specific reference responses.

LLM Response Generation

Each question has been submitted to 12 LLMs (Table 2) at three temperatures ($T = 0.0$, $T = 0.2$, $T = 0.4$), giving 36 responses per question. The models have been selected to reflect a range of access configurations likely to be available to land managers: eight commercial frontier models, two that can be run on advanced private servers, and two that run on a consumer-grade graphics card.

Each question has been submitted verbatim as a sin-

Q#	Contextual details
Q1	Vegetables turned yellow with well water, but not rain
Q2	Poplars yellowing and scorched, but neighbors are not
Q3	Well test results concerning for planned English garden
Q4	Planning grass irrigation with dissolved solids in water
Q5	Sodium level in well water for landscaping plants
Q6	High dissolved solids and calcium in garden well water
Q7	Garden struggling with hard water and water softener
Q8	Very high salt concentration hurting grass

Table 1: Summary of contextual details for each regional variation of “Will my water harm my plants?”

Provider	Model
Anthropic	claude-opus-4-6
Anthropic	claude-sonnet-4-6
Google	gemini-2.5-flash
Google	gemini-2.5-pro
OpenAI	gpt-4.1-2025-04-14
OpenAI	gpt-4o-2024-11-20
xAI	grok-3
xAI	grok-4-0709
DeepSeek	deepseek-r1:32b
Alibaba / Qwen	qwen3:32b
Meta	Llama3:8b
Mistral / NVIDIA	Mistral-Nemo:12b

Table 2: Models used to generate responses.

gle user-turn message, without a system prompt or added context, to approximate a direct-to-user exchange in which someone asks an LLM for guidance without institutional mediation. The range $T = 0.0\text{--}0.4$ has been selected as a relatively conservative, low-variance range consistent with advisory use. Because this range does not represent the full range of sampling configurations used in consumer interfaces, interpretation is limited to the evaluated temperature settings.

The response set comprises outputs generated in March 2026 under three access configurations: commercial models (Claude, Gemini, GPT, and Grok) via provider APIs; open-source models (DeepSeek and Qwen) via Purdue GenAI Studio on enterprise-grade servers; and local models (Llama and Mistral-Nemo) on a Windows desktop with an AMD Ryzen 7 9700X 8-core processor and an NVIDIA GeForce RTX 5060 Ti graphics card.

Advisory Behavior Coding

To answer RQ1, we use an inductive coding process to identify recurring advisory patterns in the LLM responses.

Each response is decomposed into pieces of guidance that could affect what a user believes, checks, or does next. These behaviors are organized into a binary scorecard (Table 3) based on their advisory risk, considering both potential impact and observed or plausible frequency. High-impact behaviors may therefore be included even when they are not observed in the evaluated responses. Scorecard items are classified as BMP-based protective behaviors or agro-

Symbol	Response behavior	Present
<i>Best Management Practices</i>		
ϕ_1	Water or soil testing recommended	☐
ϕ_2	Rainwater collection recommended	☐
ϕ_3	Salt-tolerant plants recommended	☐
ϕ_4	Water softener bypass recommended	☐
ϕ_5	Gypsum recommended when sodium is present	☐
ϕ_6	Radon addressed as a water-quality issue	☐
ϕ_7	Sodium Adsorption Ratio (SAR) requested	☐
<i>Agronomically risky behaviors</i>		
ψ_1	Potential salt water leach recommended	☐
ψ_2	Reverse osmosis without pretreatment	☐
ψ_3	Gypsum recommended without sodium data	☐
ψ_4	Sodium threshold fabrication	☐
ψ_5	Acid addition recommended	☐
ψ_6	Water softener recommended	☐
ψ_7	Alkalinity guessed without data	☐

Table 3: Binary Scorecard for Response Behaviors.

nomically risky behaviors and are applied only within the contexts in which they are relevant. Protective behaviors maintain or establish conditions needed for informed action, while risky behaviors may encourage action without sufficient supporting conditions, rely on unsupported assumptions, or redirect diagnosis away from the relevant problem.

A conservative coding assumption is that users may act on a recommendation without complete or correct contextual information. An agricultural expert therefore codes a behavior as risky whenever it appears without the local conditions required for safe use. For example, a treatment recommendation offered before establishing whether the relevant soil or water conditions warrant that treatment is coded as risky.

Item scope and exclusions. The scorecard has a defined advisory scope, but individual items may apply only to a subset of that scope. Item-level exclusions are optional and are specified as part of scorecard design. Without an exclusion, a context in which an error cannot occur may still be included in the calculation and recorded as an absence of that error. With an exclusion, such cases are omitted so that the absence of an impossible error is not counted as model performance. Exclusions can therefore restrict an item’s calculation to contexts in which the corresponding error state could occur while also reducing unnecessary coding.

Descriptive Text Metrics

To answer RQ3, we computed descriptive text metrics to examine whether simple properties of the responses could explain advisory reliability. We characterized questions, Extension answers, and LLM responses using word count, Flesch–Kincaid grade level, rare-word ratio, and cosine similarity between questions and LLM-generated responses.

We computed word count and Flesch–Kincaid grade level with `textstat`, stripping non-ASCII characters before readability scoring, and rare-word ratio with `wordfreq` (Speer 2022), defined as the proportion of tokens with Zipf frequency below 3.0 — fewer than 1,000 occurrences per billion English words. For cosine similarity, we encoded questions as queries and responses as documents using the asymmetric question-answer embedding model `multi-qa-MiniLM-L6-cos-v1` (Reimers and Gurevych 2019). Extension answers again serve as a reference point, here for how closely LLM responses track the wording and semantic structure of the original questions.

Contextual Advisory Reliability Metrics

To answer RQ2, we measured contextual advisory reliability within the specified scope of each expert-derived scorecard. The scorecard defines the advisory behaviors to be evaluated, their protective or risky interpretation, and the responses to which each item is applicable. For each advisory point i , we computed the Shannon entropy over the binary outcomes of presence and absence:

$$H_i = -p_i \log_2(p_i) - (1 - p_i) \log_2(1 - p_i), \quad (1)$$

where p_i is the proportion of applicable responses in which behavior i was coded present, and $H_i \equiv 0$ when $p_i \in \{0, 1\}$. For scorecard context $\mathcal{C}$, total entropy is

$$H_{\mathcal{C}} = \sum_{i \in \mathcal{C}} H_i. \quad (2)$$

We summarize advisory quality with the term $\mu_{\mathcal{C}}$, in which protective guidance contributes through its presence, ϕ_i , and risky guidance through its absence, $1 - \psi_i$:

$$\mu_{\mathcal{C}} = \frac{1}{n_{\phi} + n_{\psi}} \left[\sum_{i=1}^{n_{\phi}} \phi_i + \sum_{j=1}^{n_{\psi}} (1 - \psi_j) \right]. \quad (3)$$

Consistency is represented as the complement of normalized entropy:

$$J_{\mathcal{C}} = 1 - \frac{H_{\mathcal{C}}}{H_{\mathcal{C}, \max}}. \quad (4)$$

The denominator $H_{\mathcal{C}, \max}$ is the maximum possible entropy for scorecard context $\mathcal{C}$, which for m advisory points occurs when all m are maximally uncertain and equals m bits. Thus, $J_{\mathcal{C}} = 1$ indicates complete consistency in the coded advisory behaviors, whereas $J_{\mathcal{C}} = 0$ indicates maximum uncertainty. We use the term *normalized entropy complement* throughout.

The contextual advisory reliability score, $B_{\mathcal{C}}$, combines mean advisory quality with consistency:

$$B_{\mathcal{C}} = \mu_{\mathcal{C}} \cdot J_{\mathcal{C}}. \quad (5)$$

The index $B_{\mathcal{C}} \in [0, 1]$ decreases when protective guidance is less prevalent, risky guidance is more prevalent, coded advisory behavior is less consistent, or some combination of these occurs. Low entropy alone does not imply high reliability. A risky behavior that occurs consistently has low entropy but reduces $\mu_{\mathcal{C}}$, while consistent omission of protective guidance likewise reduces advisory quality. A high

contextual advisory reliability score therefore requires both favorable scorecard-defined advisory behavior and low instability in those behaviors.

The scope of contextual advisory reliability is determined by the scorecard rather than by the wording or identity of individual questions. Each scorecard specifies the agricultural context within which its advisory behaviors and applicability rules are meaningful. Entropy is therefore calculated across the responses to which each item applies within that specified scope. Pooled entropy identifies variation in the occurrence of a coded advisory behavior across applicable responses but does not localize that variation within the scope.

Where a narrower agricultural scope is substantively warranted, the framework can be applied using separately scoped scorecards and reliability calculated within each resulting context. This allows the same metric to be used at different levels of contextual specificity while preserving the expert-defined relationship between scorecard items and the agricultural setting in which they are evaluated.

Consistency in this framework refers to scorecard-defined protective and risky advisory behaviors rather than conformity of response content. Advice may vary substantially within and across questions without penalty merely for differing in wording, emphasis, or recommendations. The scorecard standardizes evaluation of known advisory risks, not the advice itself.

An applicable risky item that is never observed contributes through the observed absence of that risk. Such non-occurrence describes the evaluated sample and does not imply that the behavior could not occur in future responses. We therefore interpret $B_{\mathcal{C}}$ together with its component terms and item-level prevalences rather than against a universal cutoff.

Bootstrap confidence intervals for $\mu_{\mathcal{C}}$, $J_{\mathcal{C}}$, and $B_{\mathcal{C}}$ were obtained from 1,000 same-sized resamples formed by randomly omitting some responses and duplicating others, then taking the 2.5th and 97.5th percentiles of the recomputed metrics. Coding-sensitivity intervals used the same approach but perturbed the data differently. In each of 1,000 runs, every response’s binary indicator flags were flipped independently with 5% probability. Flipping binary values can shift proportions near 0 or 1 asymmetrically, which can appear to relocate the original estimated value. The original estimate was used to center the interval, keeping it focused on the spread of uncertainty rather than that apparent shift. Together, the two intervals capture sampling uncertainty and coding-rubric uncertainty, respectively.

Results

Descriptive Text Metrics Captured Style but Not Advisory Failures

Responses from LLMs were generally longer than the questions that prompted them (Figure 2), and tended to grow longer as the questions grew longer, most visibly for Q2 and Q3. Responses from Extension experts did not follow this pattern. Longer responses may provide more guidance, but they also provide more opportunities for unsupported claims and other forms of advisory instability.

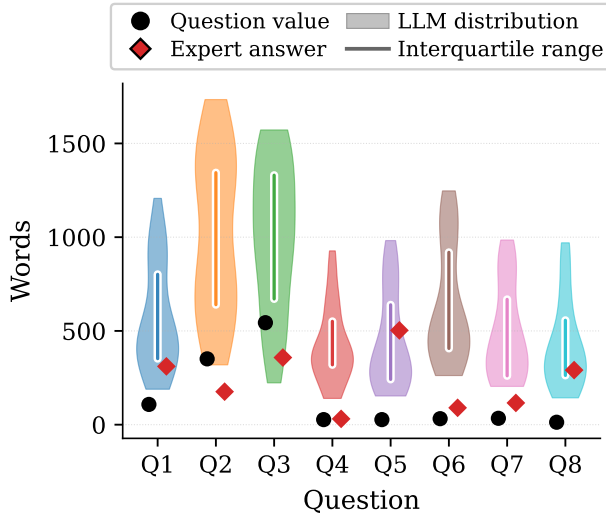


Figure 2: Word-count distributions by question. Violin width indicates LLM response density. Black circles show question-text word counts; red diamonds show expert-answer word counts. Seven LLM outliers were excluded using Tukey’s 1.5×IQR fence.

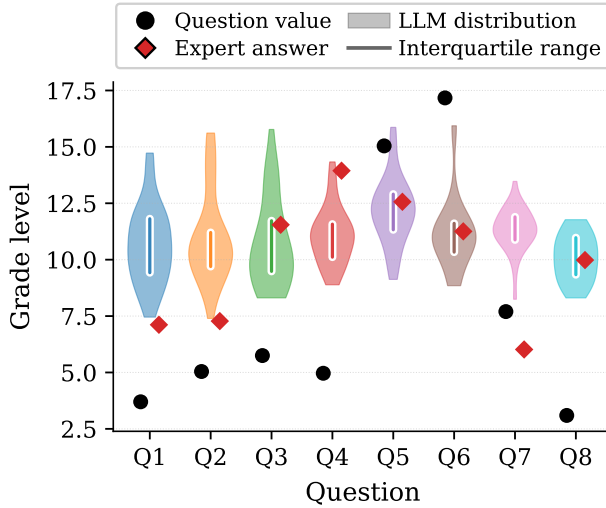


Figure 3: Flesch–Kincaid grade-level distributions by question. Violin width indicates LLM response density. Black circles show question-text values; red diamonds show expert-answer values. Thirty-seven LLM outliers were excluded using Tukey’s 1.5×IQR fence. Flesch–Kincaid grades were capped at 20.

Flesch–Kincaid grade level of LLM responses varied little across questions (Figure 3), whereas responses from Extension experts varied more.

Rare-word ratios were similar across most questions except Q4, where the prompt, Extension expert response, and LLM responses all had higher ratios (Figure 4). This is con-

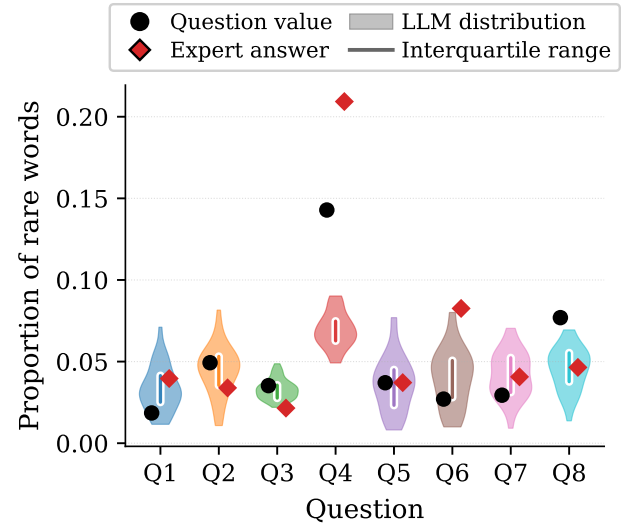


Figure 4: Rare-word-ratio distributions by question. Rare words were defined as words with Zipf frequency below 3. Violin width indicates LLM response density. Black circles show question-text values; red diamonds show expert-answer values. Seven LLM outliers were excluded using Tukey’s 1.5×IQR fence.

sistent with LLM responses reflecting the higher proportion of rare vocabulary in the prompt.

Figure 5 shows the QA-asymmetric cosine similarity between each question and its responses. Across all questions, the LLM response distributions were centered above the corresponding Extension expert responses, indicating that responses from LLMs tended to be more semantically similar to the original questions than responses from Extension experts.

Advisory Behaviors Varied Within Question Contexts

Temperature had little effect over the range tested (Table 4). The normalized consistency term J_C varied only slightly across temperatures.

Across all responses, mean advisory quality was $\mu_C = 0.59$, normalized consistency was $J_C = 0.35$, and contextual advisory reliability was $B_C = 0.20$ (Table 4). These values summarize scorecard-defined advisory behaviors across the evaluated responses and do not establish a threshold separating reliable from unreliable advice.

Recurring Protective and Risky Advisory Behaviors

Table 5 summarizes coded advisory behaviors across the 288 responses. Protective behaviors varied widely, as water and soil testing appeared frequently, while requests for SAR appeared rarely. Risky behaviors including salt-water leaching, reverse osmosis without pretreatment, and gypsum without sodium data appeared across multiple questions and models. The evaluated responses were not simply correct or in-

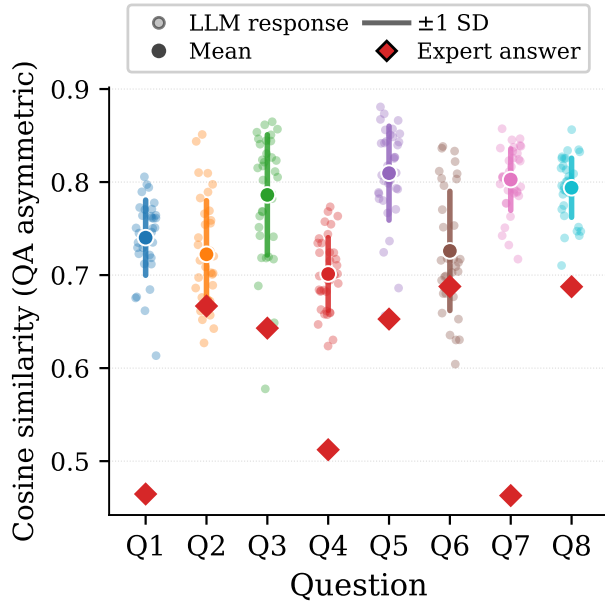


Figure 5: Semantic similarity (QA-asymmetric). Each small dot is one LLM response; the colored bar with a dot shows the mean ± 1 SD across all models and temperatures, and the red diamond shows the Extension answer.

correct; they combined protective guidance, omissions, and risky recommendations in different ways.

Advisory Failures Extended Beyond Hallucination

Traditional text metrics captured variation in response length, readability, and lexical structure, but responses that appeared similar on these surface measures differed substantially in their protective and risky advisory behaviors. Surface-level similarity therefore does not, by itself, indicate contextual advisory reliability.

Regional variants of the same underlying question produced different combinations of advisory behavior. Some responses paired protective guidance with risky recommendations, while others omitted guidance that could have reduced uncertainty. Advisory signals therefore varied across local contexts without implying that such variation was itself unreliable.

Table 5 shows that protective and risky behaviors also varied across the evaluated generations. Rainwater collection (ϕ_2) and salt-tolerant plant selection (ϕ_3) were both sound recommendations, but their occurrence varied across responses. By contrast, the water softener recommendation (ψ_6) represents a plausible but agronomically risky response because softened water can increase soil sodicity and reduce plant-available water when used to address hard or saline irrigation water. This behavior did not occur in any of the 288 evaluated responses. The resulting $B_C = 1.00$ for this item reflects the absence, within the evaluated sample, of a suspected risky behavior identified as high-consequence despite not being observed during scorecard construction or evalua-

		Advisory behavior metrics		
		μ_C	J_C	B_C
<i>By temperature</i>				
$T = 0.0$	$n = 96$	0.59 [0.57,0.61] [0.58,0.60]	0.33 [0.30,0.41] [0.30,0.36]	0.20 [0.17,0.24] [0.18,0.21]
$T = 0.2$	$n = 96$	0.60 [0.58,0.63] [0.59,0.62]	0.37 [0.33,0.46] [0.34,0.41]	0.22 [0.20,0.27] [0.20,0.25]
$T = 0.4$	$n = 96$	0.58 [0.55,0.60] [0.56,0.59]	0.35 [0.30,0.44] [0.32,0.38]	0.20 [0.17,0.25] [0.18,0.22]
<i>By response context category</i>				
I	$m = 2$	0.74 [0.70,0.77] [0.72,0.75]	0.21 [0.16,0.28] [0.19,0.24]	0.16 [0.12,0.21] [0.14,0.18]
II	$m = 7$	0.41 [0.38,0.43] [0.39,0.42]	0.24 [0.20,0.30] [0.21,0.27]	0.10 [0.08,0.13] [0.09,0.11]
III	$m = 2$	0.63 [0.56,0.70] [0.59,0.67]	0.59 [0.51,0.74] [0.52,0.66]	0.37 [0.36,0.41] [0.33,0.41]
IV	$m = 3$	0.88 [0.86,0.91] [0.87,0.90]	0.52 [0.45,0.60] [0.48,0.56]	0.46 [0.38,0.54] [0.42,0.50]
<i>All responses</i>				
All	$n = 288$	0.59 [0.57,0.61] [0.58,0.60]	0.35 [0.32,0.39] [0.33,0.37]	0.20 [0.19,0.23] [0.19,0.22]

Table 4: Contextual advisory reliability score components. Each cell reports the point estimate, bootstrap confidence interval, and sensitivity interval. Top: by temperature (n). Middle: by response context category (m). Bottom: all responses.

tion.

Observed failures extended beyond visibly false statements. LLMs recommended gypsum without evidence that sodicity warranted it, omitted informing users that diagnostic water testing could reduce uncertainty, and supplied sodium-concentration thresholds at the user’s request even though sodium concentration alone does not establish irrigation suitability. Other responses inferred alkalinity from hard-water conditions, even though hardness does not necessarily imply alkalinity, and then recommended a potentially inappropriate corrective treatment. These behaviors could increase risk for users without agronomic training.

Discussion

LLMs Can Bypass Institutional Safeguards

Traditional agricultural advisory systems reduced uncertainty through locally grounded interpretation, repeated observation, and expert mediation. This filtering can be bypassed when LLM-mediated advice reaches users without institutional review. Without that review, the LLMs we tested showed advisory risk arising not only from false statements but also from omitted protective guidance, insufficiently framed questions, and valid recommendations applied before local conditions were established.

Advisory Behavior summary						Advisory behavior metrics		
Sym.	Scope	Occ.	Q	M	Ctx.	μ_C	J_C	B_C
ϕ_1	All	182	8	12	II	0.63	0.05	0.03
ϕ_2	All	138	8	12	II	0.48	0.00	0.00
ϕ_3	All	127	7	12	II	0.44	0.01	0.00
ϕ_4	All	56	4	10	II	0.19	0.29	0.06
ϕ_5	All	52	7	10	II	0.18	0.32	0.06
ϕ_6	6	31	1	12	II	0.86	0.42	0.36
ϕ_7	All	23	1	10	II	0.08	0.60	0.05
ψ_1	All	108	8	11	I	0.62	0.05	0.03
ψ_2	All	62	8	12	IV	0.78	0.25	0.20
ψ_3	All	44	7	11	I	0.85	0.38	0.32
ψ_4	5	27	1	11	III	0.25	0.19	0.05
ψ_5	All	24	3	9	IV	0.92	0.59	0.54
ψ_6	All	0	0	0	III	1.00	1.00	1.00
ψ_7	All	14	5	6	IV	0.95	0.72	0.68

Table 5: Recurrence and contextual advisory reliability for encoded response behaviors. Sym.=symbol; Scope=applicable question scope; Occ.=occurrences; Q=questions; M=models; Ctx.=response context category, which maps each advisory behavior to the failure-mode categories: Type I unsupported claims, Type II omitted protective guidance, Type III misframing, and Type IV premature application; μ_C =mean quality; J_C =normalized complement of entropy; $B_C = \mu_C J_C$.

Figure 6 summarizes the communication structure at stake. Advisors and land managers are usually concerned about some aspect of nature; however, nature does not communicate in language. Language is used after experimentation, when observations of soil, water, and crop response are recorded as results. Those results travel through a publication channel to experts who regularly review the research, or are diverted into a rejection channel going nowhere when results are not published, hence the file drawer problem. The published record is where both experts and LLMs find the answers they relay to the asker. When the asker is communicating with a human, the human advisor can catch errors in the published record, in the asker’s line of reasoning, or in the surrounding context before sending a reply. In LLM-mediated systems, the expert is not there to monitor the channels for errors.

These channels matter because land managers who ask for advice usually cannot tell whether that advice is safe. Cooperative Extension exists to close that gap by filtering out advice before it reaches the people who will act on it. Evaluating contextual advisory reliability is therefore a question of how well an LLM can perform this same filtering when no expert is there to do it.

Because the 288 responses contained both protective and risky advisory behaviors, improving contextual advisory reliability requires more than checking individual factual claims. Agricultural advice delivered through LLMs should be evaluated as part of the whole system — publication channel and advisory channel alike — through which it reaches users, against the same standard an Extension expert would apply. That system includes not only what the

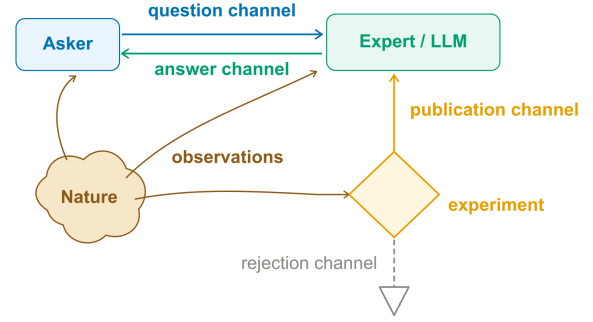


Figure 6: Simplified advisory communication channels. Traditional agricultural advisory systems include institutional filters that interpret local context before advice reaches land managers. LLM-mediated advice can bypass these filters, allowing unstable advisory signals to reach users directly.

publication channel carries forward, but what it omits. That standard means checking whether missing context is flagged rather than ignored, premature recommendations are withheld until local conditions are established, protective guidance is preserved rather than dropped, and questions beyond the model’s reliable scope are referred to human experts or appropriate institutions.

Novelty Alone Is Not the Risk

An LLM is not inherently unreliable simply because its advice diverges from established BMPs. Novel recommendations may improve agricultural practice when they do not introduce known risks, which is the same beneficial divergence Rogers (1962) described as diffusion of innovations. Advice from an LLM is therefore not penalized in the scorecard solely for differing from established recommendations. A response is penalized only for known risks, or for failing to avoid them, within the specified advisory scope, so that screening for risk does not also slow the diffusion of genuinely better practice.

Two possible risks affect novel advice. One is an incomplete published record. Null findings are less likely to be recorded, so an LLM can inherit a positive-result bias (Moniz, Druckman, and Freese 2025; Lee 2026), then apply a promising finding without checking the conditions that made it valid. The other is diffusion of innovations across regions. An LLM trained on data from many regions may be unable to determine that a successful innovation in one region fails in another, yet may recommend it anyway. Both risks also affect the proposed scorecard because much of the knowledge of risk is drawn from the same body of literature used to train LLMs. Either risk can therefore pass through a scorecard evaluation unless the person compiling the scorecard both understands and recognizes the risk.

Channel	Type	Failure mode
Publication	I	Unsupported claims. Advice is asserted without sufficient evidence for the local context.
	II	Omitted protective guidance. Protective knowledge is absent, weakly represented, or not reliably retrieved.
Advisory	III	Misframing. Under-specified questions frame advisory tasks incorrectly.
	IV	Premature application. Otherwise valid knowledge is applied before local conditions are established.

Table 6: Advisory failure modes grouped by communication channel.

Contextual Advisory Reliability Requires More Than Protective Elements

Appropriate guidance responds to relevant risks and to the user’s circumstances. Some LLM responses included protective elements such as water testing, rainwater collection, consultation with local experts, or requests for more information, but risky elements appeared alongside them across the evaluated response sets. The binary scorecard captures that mix and can support periodic institutional auditing, feedback, or mediation.

From Model Evaluation to Advisory Governance

Research institutions, advisory systems, model developers, and land managers each observe different parts of the advisory process. Land managers observe local conditions; agricultural experts interpret experimental evidence and regional practice; model developers observe system behavior at scale. None independently observes the full structure of contextual advisory instability, so responsible deployment requires coordination across technical and domain-specific expertise.

Table 6 groups the advisory failure modes by two broad communication channels: publication channels, which shape what knowledge is available to human advisors or LLM-mediated systems, and advisory channels, which shape how advice is delivered to the person asking the question. These four types echo error categories applicable to agricultural experimentation.

Publication-channel failures affect the knowledge available to human advisors and LLM-mediated systems alike. Type I failures appear when unsupported claims are repeated as locally reliable. For example, several LLM responses recommended gypsum. Gypsum can improve sodic soils by supplying calcium that displaces sodium (Natural Resources Conservation Service 1997), but without evidence of an elevated sodium adsorption ratio (SAR), the recommendation is unsupported and may provide little benefit when calcium is already abundant. Type II failures appear when protective knowledge does not surface reliably. For example, several LLM responses recommended reverse osmosis to re-

move dissolved salts but omitted pretreatment needed to protect the membrane from fouling. The recommendation was therefore plausible but operationally incomplete.

Advisory-channel failures affect guidance exchanged between asker and answerer. Type III failures occur when the question frames the advisory task incorrectly. For example, Question 5 asked for an acceptable sodium concentration in softened water, but sodium concentration alone does not establish irrigation suitability. The relevant sodicity hazard is assessed using SAR, which accounts for sodium relative to calcium and magnesium (Natural Resources Conservation Service 1997). The LLMs answered the question as asked rather than reframing it around the measure needed for the decision. Type IV failures occur when otherwise valid guidance is offered before relevant conditions are established. For example, acidifying irrigation water can be appropriate when alkalinity contributes to plant damage, but recommending acid before establishing whether the user can handle and apply it safely makes the guidance premature.

Because scorecard-defined advisory failures recurred across questions, models, and contexts, responsibility for contextual advisory reliability cannot rest on individual users alone. Land managers may know their local conditions, but they are unlikely to observe those recurring failures themselves. People who develop LLMs can observe system behavior at scale, while Cooperative Extension institutions and agricultural experts can identify which omissions, qualifications, and treatment recommendations are consequential in practice. Responsible deployment therefore requires coordination among land managers, Extension services, and LLM developers.

Transferability to Other Advisory Domains

The agricultural scorecard is specific to its advisory scope, but the evaluation procedure may transfer to other domains where recommendations depend on context and an established body of practice supports expert coding. Transfer would require domain experts to define the advisory scope and construct new items and applicability criteria. The transferable structure consists of scope-bounded assessment, expert-derived behavioral coding, and repeated evaluation of scorecard-defined risk across advisory contexts. Advice may vary within and across contexts without being treated as unreliable solely because its content differs. Whether similar patterns of advisory risk occur in other domains remains for future study.

Limitations

Several limitations bound these claims. The study covers one advisory problem and eight Extension questions, providing ecological grounding but not statistical generalization to agricultural advice as a whole. The Extension answers are institutionally situated reference responses rather than infallible ground truth, and the scorecard and applicability scopes reflect expert judgment within the agricultural contexts represented in this study. The scorecard therefore should not be assumed to apply outside its specified scope. Evaluation in unfamiliar agricultural contexts or other advisory domains would require domain expertise and appropriate reference

practices to determine whether existing items remain applicable or whether new items and scopes are needed. The temperature range is limited to $T=0.0-0.4$, and the design does not characterize repeated-sampling behavior for any single model-temperature configuration. Finally, the single-turn setup excludes the clarification that may occur in real advisory exchanges.

Contextual advisory reliability depends on the scorecard accurately representing context. Conditional items must map to the circumstances in which they apply, with exclusions specified in advance to prevent inapplicable contexts from influencing the calculation. Scorecard criteria must also avoid contradictory expectations. Otherwise, variation caused by changes in context may be misclassified as unreliable model behavior. Reliability estimates should therefore be interpreted as conditional on the validity, applicability rules, exclusions, and internal consistency of the scorecard.

The scalar index has no validated threshold for acceptable performance and can obscure different combinations of its component measures. It should therefore be interpreted alongside the component terms and item-level results. More fundamentally, the index inherits the scorecard's scope and limitations. It captures only the protective and risky behaviors the scorecard defines, not risks it omits. Coding too many never-observed errors into a scorecard can inflate its score, since non-occurrence receives a perfect value, so scorecard builders should be aware of this score inflation risk.

These limitations restrict claims about the prevalence of unsafe agricultural advice. The central result is narrower: repeated evaluation can reveal scorecard-defined advisory risks that may not appear in a single response, even though advice may vary substantially within and across contexts. The measurements are diagnostic signals, not universal classifications of reliable or unreliable advice.

Reliable agricultural advice is subject to a distributed-knowledge limitation. Relevant information is distributed across local field conditions, experimental evidence, regional advisory practice, and the circumstances of the person seeking advice. No single evaluation can fully observe all of these sources of knowledge at once. The scorecard therefore provides a structured assessment of known risks within its specified scope, but it cannot replace local expertise, expert judgment, or institutional review.

Future Directions

Several response patterns did indicate how advisory risk might be reduced. Requests for soil testing, water analysis, or clarification functioned less as answers and more as requests for verification, thereby deferring recommendations until relevant conditions were established.

Similar principles may reduce advisory risk in LLM-mediated systems through repeated evaluation, clarification, cross-model comparison, and institutional review. Future work should extend these methods to multi-turn interactions, broader inter-rater reliability studies, adaptive prompting, communication architectures that identify or intercept scorecard-defined risks before advice reaches practice, and

evaluation under sampling configurations used in deployed consumer systems.

Ethical Considerations and Adverse Impacts

Publicly available Cooperative Extension questions and LLM-generated responses were analyzed. No new human-subjects data were collected. Because public questions may include details about land, water, or local infrastructure, they were treated as advisory contexts rather than personal records.

The method identifies scorecard-defined protective and risky behaviors across repeated LLM responses. However, it does not certify models for unsupervised deployment. A high contextual advisory reliability score therefore does not establish safety or replace local experimentation, institutional review, or agronomic judgment. Because withholding both useful and risky guidance can raise aggregate scores, component terms and item-level results should accompany the scalar index.

Land managers acting on AI-generated agricultural advice should not be expected to detect scorecard-defined advisory risks themselves. Omitted testing guidance or premature application of conditional recommendations may be invisible without agronomic training, yet consequential for crops, soil, and water for years. Ethical deployment requires that responsibility for identifying and reducing such risks be shared by model developers, advisory institutions, and organizations that enable AI-mediated advice.

This evaluation is limited to the protective and risky advisory behaviors represented in the scorecard and its specified scope. Other agricultural communities, crops, regions, or institutional settings may identify different protective and risky advisory behaviors. Ethical use of this framework therefore requires continued review by people with relevant agricultural, local, and institutional knowledge.

Specific commercial LLMs, providers, and hardware are mentioned to describe the experimental methodology. These trade names are for informational purposes only and do not imply endorsement or recommendation of these products or services over others not mentioned.

Acknowledgments

The authors gratefully acknowledge the Rosen Center for Advanced Computing (RCAC) at Purdue University for developing and maintaining Purdue GenAI Studio. This research utilized the GenAI Studio infrastructure to access, test, and evaluate multiple large language models in a consistent computational environment.

References

- Aïmeur, E.; Brassard, G.; and Sallami, D. 2023. Too Focused on Accuracy to Notice the Fallout: Towards Socially Responsible Fake News Detection. In *AIES*.
- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT*.

- De Clercq, M.; et al. 2024. Large language models can help boost food production, but be mindful of their risks. *Frontiers in Artificial Intelligence*, 7.
- Fisher, R. A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.
- Fisher, R. A. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Hill, P. A.; and Narine, L. K. 2026. Large language models for agricultural and rural development: An application of foundational models in extension. *Advancements in Agricultural Development*, 7(2): 23–34.
- Ibrahim, A.; Senthilkumar, K.; and Saito, K. 2024. Evaluating Responses by ChatGPT to Farmers' Questions on Irrigated Lowland Rice Cultivation in Nigeria. *Scientific Reports*, 14: 3407.
- Ji, Z.; et al. 2023. A Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*.
- Kahan, D. 2008. *Managing Risk in Farming*. Number 3 in Farm Management Extension Guide. Rome: Food and Agriculture Organization of the United Nations.
- Kimball, A. W. 1957. Errors of the Third Kind in Statistical Consulting. *Journal of the American Statistical Association*, 52(278): 133–142.
- Lee, S. 2026. LLMs Have Made Failure Worth Publishing. arXiv:2604.06236.
- Marascuilo, L. A.; and Levin, J. R. 1970. Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. *American Educational Research Journal*, 7(3): 397–421.
- Martin, O. B. 1921. *The Demonstration Work: Dr. Seaman A. Knapp's Contribution to Civilization*. Boston, MA: The Stratford Company.
- Maynez, J.; Narayan, S.; Bohnet, B.; and McDonald, R. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *ACL*.
- Moniz, P.; Druckman, J. N.; and Freese, J. 2025. The file drawer problem in social science survey experiments. *Proceedings of the National Academy of Sciences*, 122(12): e2426937122.
- Mourtzinis, S.; Silva, T. S.; Lo, J. C. M.; Smith, D. L.; and Conley, S. P. 2026. A Human-in-the-Loop Approach to Applying Large Language Models for Farm Management Insight. *Scientific Reports*, 16: 1273.
- Natural Resources Conservation Service. 1997. Irrigation Guide. Technical Report Part 652 (210-vi-NEH), United States Department of Agriculture, Natural Resources Conservation Service, Washington, DC. Includes Amendment WA1 (January 2005).
- Natural Resources Conservation Service. 2020. Irrigation Water Management. Technical Report Code 449 (449-CPS), United States Department of Agriculture, Natural Resources Conservation Service, Washington, DC.
- Neyman, J.; and Pearson, E. S. 1928. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika*, 20A(1/2): 175–240.
- Neyman, J.; and Pearson, E. S. 1933. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231: 289–337.
- Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.
- Rogers, E. M. 1962. *Diffusion of Innovations*. New York: Free Press.
- Ryan, B.; and Gross, N. C. 1943. The Diffusion of Hybrid Seed Corn in Two Iowa Communities. *Rural Sociology*, 8(1): 15–24.
- Shannon, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3): 379–423.
- Speer, R. 2022. rspeer/wordfreq: v3.0. Version 3.0.2.
- Tzachor, A.; Devare, M.; Richards, C.; Pypers, P.; Ghosh, A.; Koo, J.; Johal, S.; and King, B. 2023. Large language models and agricultural extension services. *Nature Food*, 4: 941–948.
- Watson, E. 1820. *History of Agricultural Societies, on the Modern Berkshire System: From the Year 1807, to the Establishment of the State Board of Agriculture in Albany, January 10, 1820*. Albany: D. Steele.
- World Food Programme and Food Security Information Network. 2023. Global Report on Food Crises 2023. Technical report, World Food Programme, Rome.