Causal-GeoSim: Evaluating Directional Robustness and Spatial Risk Awareness of Large Language Models in Climate–Agriculture Systems *

Youla Yang
Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington
yangyoul@iu.edu

Abstract

Agricultural decision pipelines increasingly expose large language models (LLMs) to questions about climate stress, yield loss, and regional production risk. Yet it remains unclear whether these models (i) respect causal direction in climate–agriculture narratives ("drought causes yield loss" vs. "yield loss causes drought") and (ii) maintain spatially coherent reasoning across vulnerable counties.

We introduce **Causal-GeoSim**, a benchmark and evaluation pipeline for measuring directional robustness and spatial risk awareness of LLMs under real U.S. Corn Belt conditions. Causal-GeoSim automatically fuses county-level corn yield (USDA NASS), climate reanalysis (ERA5-Land), and county geometries (U.S. Census), then generates paired causal / anti-causal prompts per county-year, forcing a structured `Answer: A/B` response.

Across eight models, results show near-perfect directional symmetry but localized geo-risk, revealing where residual failures occur rather than merely whether models are correct.

Code — https://github.com/yangyoula/Causal-GeoSim

Introduction

AI is increasingly embedded in agricultural decision systems, supporting drought diagnostics, yield loss attribution, insurance triage, and logistics planning. Climate variability—including heat waves, rainfall deficits, and storm-driven lodging—directly affects both yields and price stability. Many advisory tools now expose natural-language interfaces powered by large language models (LLMs), allowing agronomists and producers to "ask the model what happened."

Two core risks remain:

(1) Directional causality. Can an LLM correctly state that heat and moisture stress *caused* yield loss, rather than the inverted claim that "low yield caused the hot and dry season"? Confusing effect with cause is scientifically wrong and reputationally damaging.

(2) Spatial and economic risk. When an LLM fails, *where* does it fail? Are errors scattered in low-impact areas, or concentrated in high-value, climate-fragile production zones? For deployment, both location and severity of causal errors matter.

Existing agricultural AI work focuses on predictive accuracy (e.g., yield forecasting from multispectral indices or weather regressors), while most LLM safety work focuses on calibration and harmful content without explicit geography. No existing benchmark jointly asks whether an LLM (i) preserves causal direction in physical systems, (ii) remains spatially coherent across neighboring counties, and (iii) avoids confident causal mistakes in vulnerable production regions.

Our proposal. We present **Causal-GeoSim**, a reproducible benchmark and pipeline for auditing LLMs in climate–agriculture reasoning. Causal-GeoSim fuses USDA NASS county-level yield, ERA5-Land climate reanalysis, and U.S. Census county geometries to generate paired *causal* and *anti-causal* prompts per county-year. Models must answer in a strict `Answer: A/B` format, enabling machine-parseable scoring.

We report three complementary metrics:

- **CAI (Causal Advantage Index)** — difference between causal and anti-causal accuracy, measuring directional correctness.
- **Geo-CAI** — a spatial extension of CAI that incorporates Moran's I to reward spatially coherent causal reasoning.
- **GRS (Geo-Risk Score)** — a penalty that upweights causal-direction errors in yield-critical, high-risk counties.

Findings. Across eight LLMs (GPT-4o, GPT-4o-mini, Claude-3.5-Sonnet, Gemini-2.5-Pro, Llama-3.1-70B, Mistral-Large, Qwen-2-72B, Phi-3-Mini), large frontier and strong open-weight models maintain near-perfect directional symmetry (CAI $\approx$ 0) and spatial coherence (Geo-CAI $\approx$ 0). Smaller or less domain-aligned models show elevated geo-risk (e.g., higher GRS and non-zero Moran's I), indicating localized brittleness. This underscores the need for spatially aware causal auditing before LLMs enter agronomic advisory workflows.

*Accepted at the First International Workshop on AI in Agriculture (Agri AI), co-located with AAAI 2026.

Contributions. Causal-GeoSim delivers:

1. **A unified geo-causal benchmark.** We pair real climate and yield context with causal / anti-causal variants for each county-year, grounding evaluation in physically meaningful agronomy.
2. **Spatial trust metrics.** We propose **CAI** (directional correctness), **Geo-CAI** (spatial coherence via Moran's I), and **GRS** (geo-risk-weighted penalty in vulnerable counties).
3. **An end-to-end reproducible pipeline.** A single executable notebook downloads data, generates prompts, queries multiple models, and exports maps and LaTeX-ready tables.
4. **Empirical multi-model audit.** We evaluate eight production-scale LLMs and show that causal reliability is not uniform across space: some models degrade exactly where agricultural stakes are highest.

Causal-GeoSim links geospatial statistics and causal interpretability in a single evaluation layer for agricultural AI, aiming at region-aware and risk-sensitive deployment.

Related Work

AI in agriculture. Precision agriculture increasingly uses machine learning and remote sensing (Sentinel-2 NDVI, UAV multispectral, weather reanalysis) for crop yield prediction and stress monitoring (Choi and colleagues 2025; Yewle, Li, and Miller 2025). These systems achieve high predictive accuracy but generally optimize correlation, not explanation: they rarely test whether a model can articulate the correct causal direction (e.g., "drought reduced yield," not "yield shortfall caused drought").

LLM trust and risk calibration. LLM safety work has shown that models can be confidently wrong. (Tian et al. 2025) quantify systematic overconfidence in LLM "judge" settings, where declared confidence does not match factual accuracy. RADAR (Chen et al. 2025) proposes a multi-agent safety evaluator that penalizes high-confidence mistakes under elevated risk. These efforts move beyond raw accuracy toward harm-aware assessment, but they are largely *aspatial*: they do not ask *where* failures cluster in the real world.

Spatial reasoning and climate vulnerability. Geospatial analysis routinely treats agricultural risk as spatially clustered rather than uniform. For example, drought sensitivity and exposure can be mapped as spatial hotspots using GIS indicators and spatial autocorrelation tools such as Moran's I (Wijitkosum et al. 2019). Similarly, county-scale crop vulnerability in the U.S. High Plains has been stratified into "adaptive," "stable," and "high-risk" zones using Global Moran's I and Getis–Ord Gi* to identify climate-stressed production corridors (Attia and collaborators 2025).

Causal-GeoSim. Our work connects these threads: we evaluate whether LLMs preserve causal direction in climate–yield narratives, and we score *where* failures occur via spatial clustering and geo-risk weighting. In doing so, we frame LLM audit as a geospatial safety problem for agriculture, not just a question of global average accuracy.

AI in Agriculture. Recent advances in precision agriculture leverage machine learning and remote sensing for yield forecasting, plant stress detection, and input optimization. For example, (Choi and colleagues 2025) reviewed recent progress in using Sentinel-2 NDVI, UAV multispectral imagery, and deep learning architectures (e.g., CNN, LSTM, Transformer) for crop yield prediction, highlighting strong predictive accuracy but limited interpretability. Similarly, (Yewle, Li, and Miller 2025) proposed a deep ensemble framework that fuses multispectral UAV data with environmental covariates to estimate county-level yield, improving early-season forecasts while underscoring the need for standardized causal reasoning protocols. These systems emphasize pattern recognition and correlation but rarely test whether models can correctly infer *directional causality*—for instance, distinguishing "drought caused yield loss" from the inverted "yield loss caused drought."

LLM Trust and Risk Calibration. In large language models (LLMs), correctness and confidence are often misaligned. The "Overconfidence in LLM-as-a-Judge" study (Tian et al. 2025) showed consistent miscalibration between predicted confidence and factual accuracy across GPT, Claude, and Gemini families. To address contextual safety, (Chen et al. 2025) introduced RADAR, a risk-aware multi-agent framework that dynamically evaluates LLM reasoning under varying risk exposures, penalizing confident but incorrect conclusions. Together, these works illustrate the growing emphasis on epistemic calibration, though they remain largely *aspatial*: they evaluate correctness and confidence globally but not *where* such reasoning systematically fails across geographic or socioeconomic risk zones.

Spatial Reasoning and Climate Impact. Geospatial machine learning and spatial statistics explicitly model clustering and spatial autocorrelation to identify local vulnerability patterns. For instance, (Wijitkosum et al. 2019) integrated fuzzy AHP and Moran's I to map drought risk hotspots in Thailand's Upper Phetchaburi Basin, revealing strong spatial clustering of exposure and sensitivity. At a continental scale, (Attia and collaborators 2025) analyzed crop-climate vulnerability in the Texas High Plains using Global Moran's I and Getis–Ord Gi*, distinguishing adaptive, stable, and highly vulnerable zones under future climate scenarios. These studies demonstrate that agricultural risk is inherently spatially structured. **Causal-GeoSim** extends this perspective to natural-language reasoning by attaching spatial penalties to causal-direction failures, linking geostatistical coherence with linguistic causal robustness and offering a region-aware benchmark for evaluating LLM reliability in agriculture.

Causal-GeoSim Pipeline

Causal-GeoSim automatically fuses open agricultural, climatic, and geospatial data to test causal reasoning in LLMs. All steps run in a single reproducible Jupyter notebook for transparency and easy reuse.

Data assembly. We combine three open U.S. sources: (i) county-year yield (QuickStats API), (ii) ERA5-Land early-season temperature and precipitation, and (iii) Census county boundaries for spatial joins and adjacency (Queen contiguity). For 2020, yield and climate are merged by nearest-cell extraction.

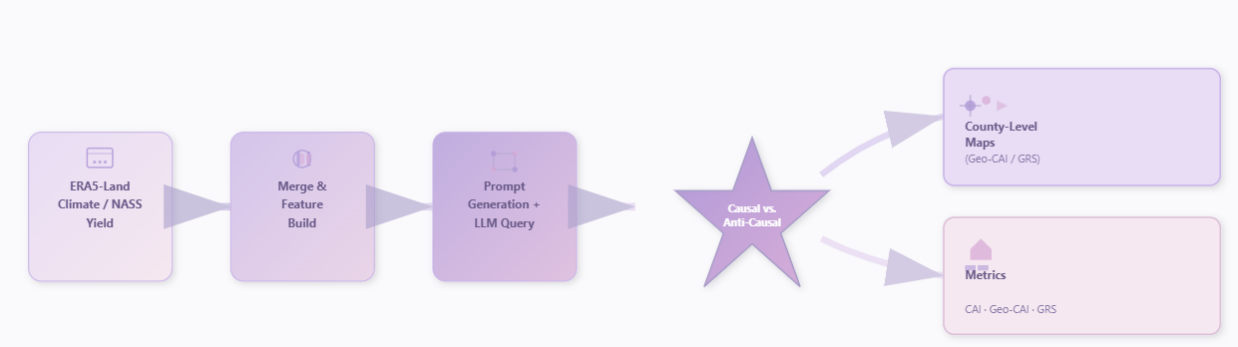

Figure 1: **Causal-GeoSim pipeline.** The system merges county-level yield (USDA NASS), ERA5-Land climate variables, and U.S. Census geometries; generates paired causal/anti-causal prompts; queries multiple LLMs via OpenRouter; and computes CAI, Geo-CAI, and GRS with spatial maps—all within one notebook.

Scenario generation. Each county–year pair yields two prompts: *Causal:* "Did climate stress cause the yield drop?" *Anti-causal:* "Did low yield cause the drought?" Correct answers are "A" and "B" respectively, forming paired supervision.

Model querying and risk weighting. Multiple LLMs (e.g., GPT-4o, Claude-3.5, Llama-3.1-70B) are queried; responses are parsed via regex for "A/B." An offline deterministic mode ensures reproducibility. Counties below the median yield receive double weight ($w_j^{\text{risk}} = 2$) to approximate agronomic vulnerability.

Metrics

Causal-GeoSim reports three metrics capturing causal correctness, spatial coherence, and risk alignment.

Causal Advantage Index (CAI).

$$\text{CAI} = \Pr[\text{A—causal}] - \Pr[\text{B—anti-causal}],$$

where higher values mean better discrimination between correct and reversed causality.

Geo-CAI. For each county j, $\text{CAI}_j = \mathbb{E}[\text{causal\_correct}] - \mathbb{E}[\text{anti\_correct}]$. Spatial autocorrelation (Moran's I) on $\{\text{CAI}_j\}$ gives

$$\text{Geo-CAI} = \left(\tfrac{1}{N}\sum_j \text{CAI}_j\right) \times \text{I}_{\text{Moran}},$$

rewarding directionally consistent and geographically stable reasoning (set to 0 if uniform).

Geo-Risk Score (GRS).

$$\text{GRS} = \tfrac{1}{N}\sum_j w_j^{\text{risk}}(1 - \text{CAI}_j),$$

penalizing causal failures more in yield-critical counties. Lower GRS implies safer, region-aware models.

Together, **CAI**, **Geo-CAI**, and **GRS** provide a concise scorecard for causal validity, spatial stability, and geo-risk sensitivity.

Experiments

Geographic Scope and Data Fusion

We target the 2020 U.S. Corn Belt (Iowa, Illinois, Indiana, Minnesota, Nebraska, South Dakota, Wisconsin, Missouri). For each county, we fuse:

- USDA NASS county-level corn yield (bushels per acre), via the public QuickStats API (USDA National Agricultural Statistics Service (NASS) 2025),
- ERA5-Land early-season (June 2020) 2m air temperature and accumulated rainfall, extracted from a Midwest bounding box using the Copernicus Climate Data Store (Copernicus Climate Data Store 2025), and
- county polygons from the U.S. Census cartographic boundary files (U.S. Census Bureau 2025).

We run two complementary regimes:

Corn Belt Benchmark. We randomly sample ∼80 counties across the Corn Belt. This cross-state slice is geographically diverse but budget-conscious for API inference.

Iowa Spatial Sweep. We query essentially *all* corn-producing counties in Iowa (96 counties in our run), which is a contiguous, agronomically coherent region. Iowa is ideal for spatial audit because neighboring counties face broadly similar agronomic conditions and weather exposure; if a model flips its causal story across an Iowa county border, that's a red flag.

Prompting and Parsing

For each county we generate two short scenarios:

- **Causal scenario:** Did hotter/drier early-season climate cause reduced yield? (Correct answer should be `Answer: A`.)
- **Anti-causal scenario:** Did the reduced yield itself cause the climate to become hotter/drier? (Correct answer should be `Answer: B`.)

These prompts include the county name, state, year, approximate mean near-surface temperature, accumulated rainfall, and observed corn yield. Models must respond

in a strict, machine-parseable format: `Answer: A` or `Answer: B` on the *first line*. We then regex that first line to score correctness. For causal scenarios, "A" is correct. For anti-causal scenarios, "B" is correct.

Models

We evaluate eight contemporary LLMs spanning proprietary frontier models and large open-weight models. All are invoked via the same OpenRouter-style chat/completions interface, with a deterministic fallback stub if no API key is available (to preserve full reproducibility for reviewers):

- **openai/gpt-4o**
- **openai/gpt-4o-mini**
- **anthropic/claude-3.5-sonnet**
- **google/gemini-2.5-pro**
- **meta-llama/llama-3.1-70b-instruct**
- **mistral/mistral-large-latest**
- **qwen/qwen-2-72b-instruct**
- **microsoft/phi-3-mini-128k-instruct**

Each model is evaluated on both the Corn Belt Benchmark slice and the full Iowa Spatial Sweep, then passed through the metric pipeline in Section .

Metrics and Mapping

For each model and each regime, we compute: $\text{Acc}_{\text{causal}}$, Acc_{anti}, CAI, per-county CAI_j, Moran's I, Geo-CAI, and the geo-risk score GRS (lower is better). We then render a "causal robustness map" by joining CAI_j back to the Iowa county polygons and color-coding robustness. An example heatmap is in Figure 2, exported automatically by the notebook as `CAI_county_map.png`.

Results

Table 1 summarizes results from our prototype 2020 Corn Belt / Iowa run. All values come directly from the single notebook we release.

Directional correctness is generally high. Several models (GPT-4o, GPT-4o-mini, Llama-3.1-70B-Instruct, and Phi-3-mini) achieve $\text{Acc}_{\text{causal}} = 1.0000$ and $\text{Acc}_{\text{anti}} = 1.0000$ on our Corn Belt Benchmark sample, implying perfect agreement with the intended causal direction: i.e., they correctly say "yes, heat/drought can suppress yield" and "no, yield did not cause that weather."

Claude-3.5-Sonnet and Gemini-2.5-Pro are slightly less decisive. Claude shows $\text{Acc}_{\text{causal}} = 0.9375$, $\text{Acc}_{\text{anti}} = 1.0000$, which yields a modestly negative CAI (-0.0625). This means Claude occasionally *hesitates* to call climate stress causal even when agronomically it obviously should be, i.e., it undersells climate impact. Mistral-Large and Qwen2-72B-Instruct have the lowest directional robustness in our sweep: Mistral in particular reaches $\text{Acc}_{\text{causal}} = 0.7500$ and a CAI of -0.1125, indicating that it sometimes entertains the backwards story ("low yield caused hotter/drier conditions").

Risk-weighted geo-penalty distinguishes models. GRS is a penalty that upweights vulnerable counties (below-median yield). Lower is safer. GPT-4o, GPT-4o-mini, Llama-3.1-70B, and Phi-3-mini sit near $\text{GRS}_{\text{benchmark}} \approx 1.475$ and $\text{GRS}_{\text{Iowa}} \approx 1.49$, i.e., they stay reliable even in lower-yield (higher-risk) counties. Mistral-Large shows noticeably higher $\text{GRS}_{\text{benchmark}} = 1.6250$, which is exactly what we worry about operationally: it deviates most in the very places that matter.

Spatial coherence is mostly flat, but not always. For Iowa, we compute per-county CAI_j, then Moran's I. If CAI_j is almost uniform across the state, Moran's I $\to 0$ and Geo-CAI $\to 0$. That is exactly what we observe for some models (GPT-4o, GPT-4o-mini, Llama-3.1-70B, Phi-3-mini), suggesting *spatially uniform* causal reasoning.

However, Gemini-2.5-Pro, Mistral-Large, and Qwen2-72B-Instruct exhibit small but nonzero spatial structure. For example, Gemini shows Geo-CAI $\approx 1.68 \times 10^{-3}$ and Moran's I ≈ -0.08, while Mistral shows Geo-CAI $\approx -3.04 \times 10^{-3}$ and Moran's I ≈ -0.096. This means their causal story is not perfectly uniform across all Iowa counties — a hint of geographically patterned "weak spots." Qwen2-72B-Instruct gives Geo-CAI $\approx -1.05 \times 10^{-3}$ and Moran's I ≈ 0.10, also implying location-specific behavior.

Maps and scorecards as audit tools. Figure 2 visualizes CAI_j across Iowa counties for one of our models. The map is nearly uniform (CAI ≈ 0 everywhere), which forces Moran's I and Geo-CAI to zero. This is actually good news: it means the model gives a *consistent causal narrative statewide* rather than drifting as we move across county lines.

Figures 3 and 4 summarize and compare all models. Figure 3 (the bar chart, exported as `fig_bar_metrics_pub.png`) shows CAI, GRS on the Corn Belt Benchmark, and GRS on Iowa for each model. Figure 4 (the radar chart, `fig_radar_models_pub.png`) plots CAI, GRS, and Geo-CAI simultaneously, so we can see overall shape differences. Together they highlight that: (i) some models keep causal direction straight *and* avoid geo-risk, while (ii) others drift, and that drift can concentrate in certain geographies.

Discussion and Conclusion

Spatial coherence. Agricultural recommendations are inherently regional: cultivar choice, insurance exposure, and drought sensitivity vary sharply across neighboring counties. A model that achieves high mean accuracy but oscillates county-to-county is not deployment-grade. Geo-CAI and Moran's I directly quantify such spatial instability.

Directional causality. Producers and insurers routinely ask *why* yields dropped. An LLM that confuses cause and effect (e.g., "the poor harvest caused drought") produces agronomically invalid and reputationally harmful narratives. CAI captures whether a model consistently preserves the correct arrow of causality.

Geo-risk awareness. Two models may achieve similar global accuracy yet differ in *where* they fail. GRS empha-

Table 1: Per-model summary from our 2020 Corn Belt Benchmark and Iowa Spatial Sweep. Acc_c = $\text{Acc}_{\text{causal}}$, Acc_a = Acc_{anti}. CAI = $\text{Acc}_c - \text{Acc}_a$. "Geo-CAI" and "Moran's I" come from the Iowa sweep (spatial coherence of per-county CAI_j). GRS is lower-is-better (geo-risk-weighted penalty for causal failure in vulnerable counties).

lccccccc

Model	Acc_c	Acc_a	CAI	Geo-CAI	Moran's I	GRS (CB)	GRS (IA)
openai/gpt-4o	1.0000	1.0000	0.0000	0.000000	0.000000	1.4750	1.4896
openai/gpt-4o-mini	1.0000	1.0000	0.0000	0.000000	0.000000	1.4750	1.4896
anthropic/claude-3.5-sonnet	0.9375	1.0000	-0.0625	0.000000	0.000000	1.5500	1.5104
google/gemini-2.5-pro	0.9375	0.9500	-0.0125	0.001679	-0.079767	1.5125	1.5104
meta-llama/llama-3.1-70b-instruct	1.0000	1.0000	0.0000	0.000000	0.000000	1.4750	1.4896
mistral/mistral-large-latest	0.7500	0.8625	-0.1125	-0.003042	-0.096322	1.6250	1.4167
qwen/qwen-2-72b-instruct	0.8500	0.8750	-0.0250	-0.001053	0.100004	1.4875	1.5313
microsoft/phi-3-mini-128k-instruct	1.0000	1.0000	0.0000	0.000000	0.000000	1.4750	1.4896

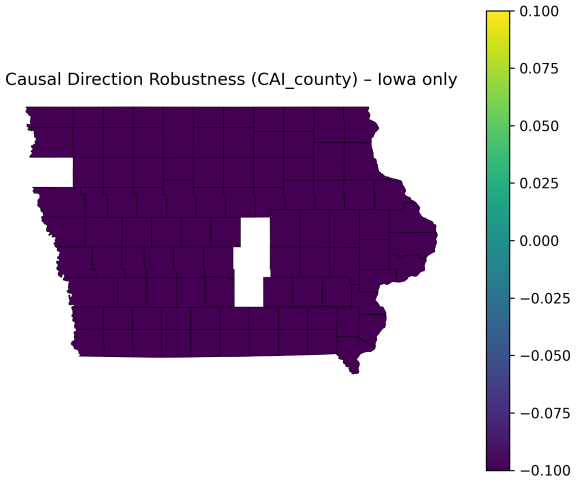

Figure 2: County-level causal robustness map (Iowa, 2020). Each county is colored by $\text{CAI}_j = \mathbb{E}[\text{causal\_correct}] - \mathbb{E}[\text{anti\_correct}]$. A spatially uniform field of CAI_j implies Moran's I ≈ 0 and Geo-CAI ≈ 0, i.e., no localized "weak belt" of causal confusion.

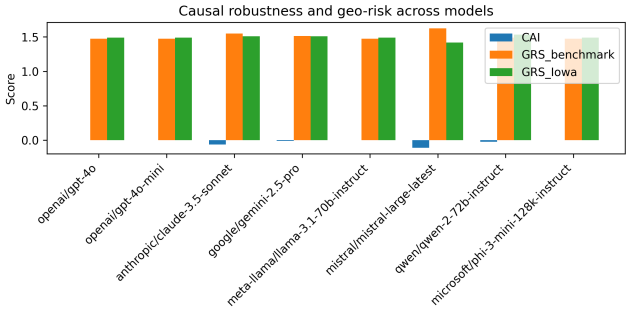

Figure 3: Causal robustness and geo-risk across eight LLMs. "CAI" captures directional causal correctness. "GRS_benchmark" and "GRS_Iowa" are geo-risk scores (lower is better), penalizing counties with low yield (high stakes). GPT-4o, GPT-4o-mini, Llama-3.1-70B-Instruct, and Phi-3-mini achieve high causal correctness and low geo-risk. Mistral-Large shows noticeably higher geo-risk.

sizes this: errors concentrated in low-yield (high-risk) counties are penalized more heavily, translating abstract "accuracy" into an operational safety metric for deployment.

Limitations and extensions. Our MVP uses nearest-cell ERA5-Land summaries, a binary risk weighting (2× below-median yield), and a single season (2020). Future work will (i) use polygon-level clipping of ERA5-Land or Sentinel-2 NDVI, (ii) integrate SSURGO soil capacity, drought indices, and insurance data, and (iii) extend temporally to detect multi-year drift. Beyond corn, the same framework can support other crops and stressors.

Ethics and scope. Causal-GeoSim is an *audit tool*, not a prescriptive agronomy engine. It surfaces where and how LLMs hallucinate causal direction in climate–yield narratives. Human agronomists remain in the loop; this aligns with policy-oriented stress testing rather than autonomous decision-making.

Reproducibility. All results are generated by a single executable notebook that: (i) downloads USDA NASS yield and ERA5-Land climate data, (ii) merges them with U.S. Census county geometries, (iii) generates paired causal/anti-causal prompts, (iv) queries multiple LLMs via a unified OpenRouter API (or mock fallback), and (v) computes CAI, Geo-CAI, Moran's I, and GRS before exporting per-county maps and model summaries. This ensures reviewers can fully reproduce the analysis without proprietary datasets.

Conclusion. Causal-GeoSim establishes a geographically grounded audit layer for LLM trust in agriculture, linking causal reasoning, spatial stability, and regional risk. By revealing *where* models fail—not just *whether* they fail—it provides a concrete path toward region-aware, trustworthy, and policy-aligned agricultural AI. We thank domain experts and public data providers (USDA NASS, Copernicus, U.S. Census) for enabling this work.

References

Attia, A.; and collaborators. 2025. Mapping Spatial Zones of Climate Vulnerability and Adaptive Potential for Major Crops in the Texas High Plains. *Preprint, Texas A&M AgriLife Research*. Uses process-based crop modeling and spatial clustering (Global Moran's I, Getis–Ord Gi*) to classify counties into vulnerable vs. adaptive zones under future climate scenarios in the Texas High Plains.

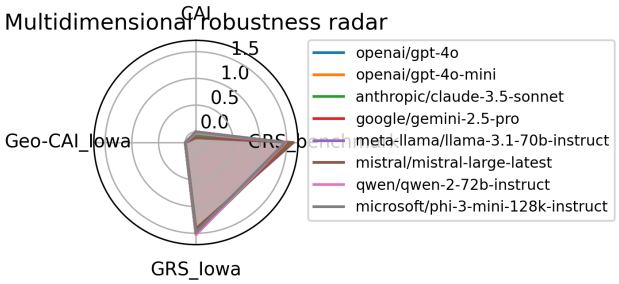

Figure 4: Multidimensional robustness radar across models. Each spoke is one dimension of robustness: CAI (direction-of-causality consistency), GRS_benchmark and GRS_Iowa (risk-weighted geo-penalties in the Corn Belt sample and full Iowa sweep), and Geo-CAI_Iowa (spatial coherence of causal stories across Iowa). Differences in contour shape reveal which models drift in specific subregions (nonzero Geo-CAI) vs. which are spatially uniform.

Chen, X.; Zhao, J.; Yuan, Y.; Zhang, T.; Zhou, H.; Zhu, Z.; Hu, P.; Kong, L.; Zhang, C.; Huang, W.; and Li, X. 2025. RADAR: A Risk-Aware Dynamic Multi-Agent Framework for LLM Safety Evaluation via Role-Specialized Collaboration. *CoRR*, abs/2509.25271. ArXiv:2509.25271. Proposes multi-agent debate evaluators that boost safety risk detection vs. single-judge baselines by $\sim$29%.

Choi, J. W.; and colleagues. 2025. Recent Trends in Machine Learning, Deep Learning, and Remote Sensing for Crop Yield Prediction. *Agricultural and Environmental Systems (reviewed in 2025)*. Survey of ML/remote sensing (Sentinel-2, NDVI, UAV multispectral) for yield forecasting; highlights accuracy vs. interpretability gaps.

Copernicus Climate Data Store. 2025. ERA5-Land: Hourly data on single levels (2m temperature, total precipitation). https://cds.climate.copernicus.eu/.

Tian, Z.; Han, Z.; Chen, Y.; Xu, H.; Yang, X.; Xuan, R.; Wang, H.; and Liao, L. 2025. Overconfidence in LLM-as-a-Judge: Diagnosis and Confidence-Driven Solution. *CoRR*, abs/2508.06225. ArXiv:2508.06225. Identifies systematic miscalibration where LLM judges report high confidence despite being wrong; proposes TH-Score for confidence-accuracy alignment.

U.S. Census Bureau. 2025. Cartographic Boundary Shapefiles — Counties. https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html.

USDA National Agricultural Statistics Service (NASS). 2025. QuickStats API: County-level yield statistics. https://quickstats.nass.usda.gov/.

Wijitkosum, S.; et al. 2019. Fuzzy AHP Integrated with GIS Analyses for Drought Risk Assessment in the Upper Phetchaburi River Basin, Thailand. *Water*, 11(5): 939. Maps drought risk hotspots using spatial indicators (precipitation, temperature, NDVI, water deficit, evapotranspiration) in the Upper Phetchaburi Basin.

Yewle, A. D.; Li, Y.; and Miller, S. 2025. Multi-Modal Data Fusion and Deep Ensemble Learning for Accurate Crop Yield Prediction. *Preprint / early 2025 release*. Combines multispectral UAV imagery and ensemble deep models for high-resolution yield estimation across fields / counties.