Full-Season Agent Evaluation in Soybean Farm Operations under Real-World Agricultural Process Dynamics

Ao Qu [*1 2] **Panagiotis Michelakis** [*† 3] **Yiannis Hadjiyianni** [*3] **Feng Li** [1 2] **Jingchi Jiang** [1 2] **Dimitrios Stamoulis** [1 2] **Jie Liu** [1 2]

Abstract

In this preliminary field study, we evaluate contemporary agent systems on the operational pipeline of an industry-grade soybean research farm at HIT. The environment is not a bespoke benchmark wrapper around an LLM tool interface; it is built from the physics-grounded process models, sensing infrastructure, machinery workflows, and historical operation records used to manage and study the farm. We evaluate nine representative agent-controller methodologies across three levels of farm operations, ranging from atomic tasks and episode chains to full-season scenarios that span the entire harvest cycle. Our results show that human-expert context is the dominant factor in long-horizon performance: without it, full-season agents remain 34% below the human-oracle soybean yield, while expert context reduces this shortfall to 7%. We observe that full-path agentic correctness and crop yield drop under longer horizons, and that agriculture-domain LLM-as-an-Expert guidance can recover yield while still lagging human experts. With these field observations from an operating research farm, we hope to inform AI research in domains where actions interact with physical processes, delayed feedback, and production-level outcomes.

1. Farm Setting and Research Question

We operate a soybean research farm at Harbin Institute of Technology's smart-agriculture research site (Figure 1). The site has production-grade machinery, a sensing layer with fixed soil and canopy sensors, multispectral and thermal drones, satellite vegetation products, a weather station, a calibrated physics-grounded process model, and multi-season historical harvest and yield records over multiple years of on-field studies. Field operations span ridge preparation, planting, irrigation, fertilization, pest and disease treatment, harvest, grain handling, drying, and storage. We are preparing a controlled portion of this surface to be fully operated by agents alongside the existing human workflows, with the goal of comparing agent-managed and human-managed farm operations in upcoming harvest seasons. This makes the evaluation question **practical**: we need to understand how existing agent systems behave on real farm operations before deciding what to deploy in the field.

Figure 1. Agentic world model for the HIT soybean research farm. The system follows an "*everything is an event*" abstraction, connecting farm apps for drones, sensors, weather, robots, tractors, and farm-state updates to full-season soybean operations.

Most agent evaluations measure tool-use success on web, coding, or OS tasks where feedback is immediate and the action space is bounded by application APIs. Farm operation differs in three concrete ways (Yan et al., 2026; Zhang et al., 2026; Xu et al., 2025; Seo & Lee, 2026; Zuzuárregui et al., 2025). **Feedback is delayed:** a poor planting decision surfaces as weak crops emergence days later, a missed irrigation window appears as yield loss months later. **Observations are partial:** sensors are zone-level across ridges, drones are weather-gated, and ground inspection is sparse. **Consequences propagate:** operational errors compound across the season, so an action can be syntactically valid and operationally wrong (e.g., harvesting before grain moisture is suitable). The question is therefore not whether agents can call farm tools, but whether they preserve correct operational judgment under spatiotemporal, partially observed, and delayed-feedback conditions.

[*]Equal contribution [†]Currently with new affiliation [1]Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China [2]National Key Laboratory of Smart Farm Technologies and Systems, Harbin, China [3]School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece. Correspondence to: Dimitrios Stamoulis <dimi@hit.edu.cn>.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Copyright 2026 by the author(s).

We use the farm's digital twin as a realistic agentic **world model** to evaluate nine representative agent-controller architectures across three scenario levels: atomic tasks, episode chains, and full-season scenarios. Our preliminary results yield observations that we hope are useful for AI researchers building agents for physical workflows. First, domain expertise is the dominant lever: human-expert context reduces the full-season yield shortfall from 34.0% under zero context to 7.0%. Second, agent and farm performance diverge under longer horizons: on shorter episodic tasks, agents reach 99.4% correctness with no crop yield loss, while in full-season management scenarios correctness drops to 94.2%, 7.0% below the human-oracle soybeans yield. Third, multi-agent orchestration introduces coordination cost, decreasing correctness by up to 17.1%. Fourth, frontier agriculture LLMs are only partial substitutes for human experts: LLM-as-an-Expert context reduces the yield shortfall, but remains 5.7% behind human-expert in-context learning.

2. Methodology

Farm site and growth process dynamics. The study site is the industry-grade university soybean research farm at Harbin Institute of Technology's smart-agriculture research facility. The modeled plot is a $268\,\mathrm{m} \times 71\,\mathrm{m}$ field organized into 64 ridges. The farm process-modeling stack reflects approaches used in on-field studies: a WGEN/Richardson-style daily weather generator (Siler & Singh, 2022) drives a soil-water and soil-temperature model, GDD-based phenology with seed-type-specific maturity targets (Akyuz et al., 2017), a canopy and biomass model with water, nutrient, stand-fraction, and biotic-stress multipliers (Monteith, 1977; Monsi & Saeki, 2005), a biotic-pressure model for weed, insect, and disease dynamics with treatment effects (Steduto et al., 2009), and a yield-recovery model that converts crop state at maturity into harvested grain through grain moisture, field loss, and machine loss (Humburg, 2019).

Farm instrumentation and Tool APIs. The agent acts through tools derived from the APIs and operational interfaces installed in the farm, exposed as LLM function-calling tools. Sensing tools provide fixed, mobile, and remote observations, including weather-station and forecast queries, soil temperature and moisture readings from DF-G3012/DF-HRS sensor units, canopy-index readings from Apogee sensors, DLS light and total-radiation measurements, SPAD chlorophyll checks, multispectral DJI drones for NDVI crop-vigor products, thermal drone inspection for canopy-temperature and water-stress assessment, ground robot inspection for local emergence, pest, disease, and damage checks, and satellite vegetation-index products.

Stateful ARE agentic engine. We build the farm execution engine on top of the ARE framework (Froger et al., 2025), which represents agent environments through stateful apps, an event queue, notifications, scenarios, and logged execution traces. We adopt this abstraction because farm operation is naturally event-based: operation tools execute ridge preparation, planting, irrigation, fertigation, pesticide application, harvest, and storage as timestamped farm events. Each event has arguments such as operation duration, preconditions such as weather and equipment readiness. This keeps the action space tied to farm operations without forcing the agent to carry the entire seasonal state inside a monolithic reasoning trace: the environment maintains the farm state, event queue, notifications, and operation history, while the controller observes and acts through the tool interface. Without loss of generality, tool invocation is emulated inside the digital twin, so operations retain their farm-state effects and timing dependencies without real-clock execution.

Scenarios. We curate scenarios from representative on-field procedures based on historical processes, crop records, seasonal observations, machinery operations, and expert workflows. We consider three levels of increasing complexity: **L1** atomic tasks require one or two tool calls, such as checking weather before drone flight; **L2** episodes chain observation, diagnosis, and intervention, such as detecting an anomaly from a drone survey and applying targeted treatment; and **L3** full-season scenarios span planting through post-harvest storage, with multiple interventions, delayed consequences, and accumulated effects on crop state and recovered yield. For each scenario, we follow the ARE scenario annotation (Froger et al., 2025): three domain experts independently specify oracle solutions using the same farm-operation primitives exposed to the agents. If the first two workflows diverge, we inspect the scenario to resolve task ambiguity and revise the action sequences. The third expert confirms consistency and addresses any omissions. We use a compact suite of 25 scenarios: 8 L1, 8 L2, and 9 L3 tasks (examples shown in Appendix A). We replay the resulting workflows in the ARE engine, producing human-oracle farm-state operation trajectories and target crop yields against which agent runs are compared.

Agent families. We integrate nine controller architectures into ARE and evaluate them over the scenario suite: ReAct (Yao et al., 2023), planner-executor Plan-and-Act (Erdogan et al., 2025), reflective memory Reflexion (Shinn et al., 2023), multi-specialist AutoGen (Wu et al., 2024), adaptive verifier MMRL (Tan et al., 2026), ReWOO (Xu et al., 2023), tree search LATS (Zhou et al., 2024), critic-refiner CRITIC (Gou et al., 2024), and graph memory GoT (Besta et al., 2024). Each controller supports two modes: *direct* tool access and *agent-to-agent (A2A)* (A2A Project, 2025), where queries are routed through specialist app agents (weather, sensor, machinery, operations experts).

Metrics. We report two metrics per run: agentic correctness, measured as Kendall-style temporal correctness

Table 1. Agent performance on farm operations for atomic tasks (**L1**), episode chains (**L2**), and full-season scenarios (**L3**) across agent families and prompting schemes with Qwen3.6-35B-A3B (Yang et al., 2025). Longer task horizons surface the yield gap between zero-context and expert-instructed in-context learning. **Yield Loss** reports the percentage drop from the human oracle yield; **KTC Score** is temporally grounded full-path agent task correctness following ARE evaluation (Froger et al., 2025; Michelakis et al., 2025).

Controller	L1: Tasks		L2: Episodes						L3: Full-Season Scenarios			
	Zero Cont.	Expert Inst.	Zero Context		LLM-as-an-Expert		Expert Instruct		Zero Context		Expert Instruct	
	KTC Score	KTC Score	Yield Loss	KTC Score	Yield Loss	KTC Score	Yield Loss	KTC Score	Yield Loss	KTC Score	Yield Loss	KTC Score
ReAct (Yao et al., 2023)	62.3%	100%	0.8%	82.8%	12.6%	52.0%	0%	99.9%	36.1%	78.1%	6.4%	94.1%
ReWOO (Xu et al., 2023)	77.3%	99.7%	0.1%	72.9%	0.1%	90.7%	0%	99.9%	34.3%	71.8%	16.2%	94.6%
Reflexion (Shinn et al., 2023)	75.7%	100%	12.5%	60.0%	12.5%	78.6%	0%	99.9%	32.4%	81.1%	5.6%	93.4%
CRITIC (Gou et al., 2024)	89.3%	100%	0.1%	74.3%	0.1%	76.4%	0%	99.9%	45.0%	81.7%	5.1%	91.3%
GoT (Besta et al., 2024)	66.2%	100%	0.1%	81.6%	0.1%	67.7%	0%	99.9%	14.2%	79.1%	7.0%	93.7%
AutoGen (Wu et al., 2024)	80.5%	100%	0%	94.6%	1.3%	78.8%	0%	99.9%	11.4%	84.0%	**4.7%**	94.3%
LATS (Zhou et al., 2024)	78.2%	97.6%	12.5%	70.5%	12.5%	80.6%	0%	99.9%	43.7%	75.9%	5.8%	94.7%
Plan&Act (Erdogan et al., 2025)	71.2%	100%	0.8%	74.8%	12.6%	68.4%	0%	99.9%	34.8%	80.0%	6.1%	95.5%
MMRL (Tan et al., 2026)	64.8%	100%	0%	72.0%	0%	84.1%	0%	99.9%	54.3%	72.3%	6.3%	**96.1%**
Average	73.9%	99.7%	3.0%	75.9%	5.7%	75.3%	0%	99.9%	34.0%	78.2%	7.0%	94.2%

Table 2. Agent performance on farm operations for atomic tasks (**L1**) and episode chains (**L2**) across agent families and prompting schemes with DeepSeek-V4-Flash (DeepSeek-AI, 2026) and GPT-4o-mini (OpenAI et al., 2024).

Controller	DeepSeek-V4-Flash (DeepSeek-AI, 2026)								GPT-4o-mini (OpenAI et al., 2024)			
	L1: Tasks		L2: Episodes						L2: Episodes			
	Zero Cont.	Expert Inst.	Zero Context		LLM-as-an-Expert		Expert Instruct		Zero Context		Expert Instruct	
	KTC Score	KTC Score	Yield Loss	KTC Score	Yield Loss	KTC Score	Yield Loss	KTC Score	Yield Loss	KTC Score	Yield Loss	KTC Score
ReAct (Yao et al., 2023)	90.1%	100%	12.5%	93.4%	0.1%	77.9%	12.5%	99.2%	25.0%	49.9%	12.5%	100%
ReWOO (Xu et al., 2023)	85.7%	100%	11.7%	82.7%	0%	91.2%	0%	100%	12.6%	25.0%	12.5%	99.8%
Reflexion (Shinn et al., 2023)	70.6%	99.1%	5.2%	90.5%	0.1%	80.3%	0%	95.5%	12.5%	35.4%	12.5%	99.7%
CRITIC (Gou et al., 2024)	87.7%	99.4%	11.7%	97.6%	0.1%	82.3%	0%	98.8%	12.5%	58.3%	12.5%	98.6%
GoT (Besta et al., 2024)	90.0%	97.0%	12.5%	81.6%	0%	82.4%	0%	86.3%	0.1%	37.5%	12.5%	100%
AutoGen (Wu et al., 2024)	93.9%	100%	0.9%	93.9%	6.2%	76.7%	2.3%	97.7%	0.1%	64.5%	12.5%	98.9%
LATS (Zhou et al., 2024)	84.2%	99.7%	0%	72.3%	0.1%	87.1%	0%	87.0%	12.7%	45.0%	12.5%	99.7%
Plan&Act (Erdogan et al., 2025)	89.0%	97.8%	0%	94.4%	0.1%	84.6%	0%	100%	12.8%	37.4%	12.5%	99.7%
MMRL (Tan et al., 2026)	64.5%	99.4%	12.5%	81.7%	0%	78.8%	0%	98.6%	12.7%	58.1%	12.5%	99.7%
Average	84.0%	99.1%	7.5%	87.6%	0.7%	82.4%	1.7%	95.9%	11.2%	45.7%	12.5%	99.6%

(**KTC**) (Michelakis et al., 2025), which captures the alignment between the agent trajectory and the oracle workflow at the level of farm events, rewarding required operations executed in causally valid order and penalizing missing, invalid, or reordered steps; and **Yield Loss**, which reports how far the agent's final crop yield is below the human oracle.

3. Results

Expert context dominates the controller spread. Table 1 reports Qwen3.6-35B-A3B (Yang et al., 2025) results under direct tool access. Across all L1/L2/L3 scenario levels, zero-context agents average 73.9%, 75.9%, and 78.2% KTC on L1, L2, and L3, respectively. With expert context, these scores increase to 99.7%, 99.4%, and 94.2%. The yield gap appears mainly at the full-season horizon: zero-context agents have 3.0% average Yield Loss on L2, but 34.0% on L3. Expert-driven in-context learning reduces the L3 Yield Loss to 7.0%, a 27.0% reduction from a single context change. In terms of agent families, we observe that controller choice matters most when domain context is missing. Similar expert-context gains are observed for the DeepSeek and GPT variants on L2 scenarios, as reported in Table 2.

Table 3. Controller results on episodes (**L2**) via Agent2Agent (A2A Project, 2025) routing with Qwen3.6-35B-A3B (Yang et al., 2025).

Controller	Zero Context		Expert Instruct	
	Yield Loss	KTC Score	Yield Loss	KTC Score
ReAct (Yao et al., 2023)	8.6%	65.4%	11.8%	99.2%
ReWOO (Xu et al., 2023)	12.6%	54.2%	0.1%	99.7%
Reflexion (Shinn et al., 2023)	10.3%	58.3%	0.1%	96.4%
CRITIC (Gou et al., 2024)	12.6%	44.6%	0.1%	98.5%
GoT (Besta et al., 2024)	0.1%	67.2%	0.1%	99.4%
AutoGen (Wu et al., 2024)	12.6%	49.2%	0.1%	99.0%
LATS (Zhou et al., 2024)	0.1%	85.0%	0.1%	97.6%
Plan&Act (Erdogan et al., 2025)	12.6%	47.5%	0.1%	100.0%
MMRL (Tan et al., 2026)	12.6%	58.3%	12.6%	98.8%
Average	9.1%	58.9%	2.8%	98.7%

LLM-as-an-Expert. We ablate domain-LLM against human-expert context on L2 scenarios with an LLM-as-an-Expert: we use Agri-CM^3 (Wang et al., 2025) to generate

scenario policies from the task description, available tools, and world state, using the same response template as the human-expert instructions. For Qwen3.6-based agent families, domain-LLM context reduces average Yield Loss from 7.5% under zero context to 0.0%, but average KTC remains 75.3%, compared with 99.9% under human-expert context. This suggests that agriculture-specialized LLMs can recover high-level agronomic decisions, but might not consistently encode the procedural ordering required for farm operations.

Agent2Agent orchestration. We evaluate A2A routing (A2A Project, 2025) where, following ARE (Froger et al., 2025), the main controller no longer calls all tools directly. Instead, weather, sensing, machinery, and farm-operation interfaces are exposed through specialist app agents, and the main controller coordinates with them to complete the task. Table 3 reports the L2 results for Qwen3.6. Under zero context, A2A reduces average KTC from 75.9% to 58.9% and degrades yield, increasing crop loss from 3.0% to 9.1%. With expert context, the gap narrows but does not disappear: KTC remains high at 98.7%, but Yield Loss increases from 0.0% to 2.8%. The effect is also controller-dependent: tree search improves under zero-context A2A, while other controllers lose correctness or yield. We postulate that this degradation comes from the coordination overhead, consistent with observations in other expert-level agents (Lee et al., 2025; Badmus et al., 2025).

4. Discussion

From expert instructions to reusable skills. Our field study corroborates the role of tacit expert knowledge as a core component of farm-agent deployment. Farm management relies on practices, timing rules, and local operating discipline that are often implicit in human expertise. Encoding this knowledge as scenario-specific in-context instructions is useful for our evaluation, but it is not a scalable deployment strategy (Stamoulis & Marculescu, 2025): each new task, crop condition, equipment configuration, or seasonal edge case would require another written policy. A more scalable direction for future work is to organize this knowledge into reusable **skill libraries** that can be retrieved, composed, and adapted across adjacent farm tasks. This is consistent with the growing use of automated skill-based reasoning in other expert domains, such as power-grid (Badmus et al., 2025) and Earth-observation analysis (Bhattaram et al., 2025; Shabbir et al., 2026; Chen et al., 2026b; Feng et al., 2026), where recent work has shown how structured knowledge pools can improve agent generalizability.

Full-path evaluation paired with downstream objectives. Atomic tasks and short episodes mostly test whether the scaffold can select tools, parse observations, and satisfy local preconditions. Full-season scenarios test whether decisions remain coherent after their effects propagate through soil state, crop growth, stress accumulation, treatment residuals, harvest timing, and storage outcomes. A planting, irrigation, or spraying decision can look locally plausible while still creating downstream yield loss, so the downstream objective must be evaluated alongside the action trace. In our setting, KTC agent correctness and Yield Loss diagnose different failure modes. KTC captures whether the agent follows the expected farm-event trajectory with the right ordering and dependencies. Yield Loss captures whether the final crop outcome survives the agent's decisions. A high-KTC, high-yield-loss run indicates that small operational differences can still have physical consequences, while a lower-KTC, low-yield-loss run indicates that the agent deviated from the reference trace but preserved the outcome. This motivates evaluation protocols that combine full-path correctness with domain-level objectives and reasoning uncertainty (Chen et al., 2026a; Froger et al., 2025; Michelakis et al., 2025).

Agent orchestration should preserve operational state. Controller design affects farm-agent behavior, especially when domain context is limited. Planning, memory, retrieval, verification, specialist decomposition, and graph-style state tracking can improve the execution trace, but they do not remove the need to maintain agronomic and operational assumptions across time. The same applies to multi-agent layouts. Splitting the farm interface into weather, sensing, machinery, and operation specialists matches the structure of real farm systems, but it also introduces communication and state-sharing requirements. If the main controller loses track of what each specialist observed or executed, orchestration might be error-prone. The main design requirement is therefore not only better decomposition, but preserving shared farm state, timing constraints, and task context across controller steps and agent boundaries.

LLM-as-an-Expert context. Using LLMs as experts, judges, or instruction generators is common in agent evaluation pipelines, especially in domains such as OS, web, and code where task structure, documentation, and feedback signals are more standardized. The LLM-as-an-Expert result is therefore a useful stress test for agriculture: even when an agriculture-specialized LLM is given the scenario, tools, world state, and a response template aligned with the human-expert instructions, it does not fully match human experts on procedural ordering. This does not make the domain LLM unhelpful; it provides meaningful agronomic context and recovers high-level crop-management choices, further motivating structured knowledge libraries that complement both LLM-distilled and human-written in-context guidance.

Conclusion. Our results inform the next stage of our research: operating a plot under agent management alongside the human-operated farm. We hope our preliminary observations provide the agent community with a useful reference point for evaluation in other physical-process settings.

References

A2A Project. Agent2Agent protocol (A2A). https://github.com/a2aproject/A2A, 2025. GitHub repository.

Akyuz, F. A., Kandel, H., and Morlock, D. Developing a growing degree day model for north dakota and northern minnesota soybean. *Agricultural and Forest Meteorology*, 239:134–140, 2017.

Badmus, E. O., Sang, P., Stamoulis, D., and Pandey, A. Powerchain: A verifiable agentic ai system for automating distribution grid analyses, 2025. URL https://arxiv.org/abs/2508.17094.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., and Hoefler, T. Graph of thoughts: solving elaborate problems with large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i16.29720. URL https://doi.org/10.1609/aaai.v38i16.29720.

Bhattaram, A., Chung, J., Chung, S., Gupta, R., Ramamoorthy, J., Gullapalli, K., Marculescu, D., and Stamoulis, D. Geoflow: Agentic workflow automation for geospatial tasks. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '25, pp. 1150–1153, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720864. doi: 10.1145/3748636.3763217. URL https://doi.org/10.1145/3748636.3763217.

Chen, T., Yao, H., Chen, J., Papalexakis, E. E., and Wei, H. Every response counts: Quantifying uncertainty of llm-based multi-agent systems through tensor decomposition, 2026a. URL https://arxiv.org/abs/2604.08708.

Chen, Z., Wang, H., Yao, J., Ghamisi, P., Zhou, J., Atkinson, P. M., and Zhang, B. Cangling-knowflow: A unified knowledge-and-flow-fused agent for comprehensive remote sensing applications, 2026b. URL https://arxiv.org/abs/2512.15231.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Erdogan, L. E., Furuta, H., Kim, S., Lee, N., Moon, S., Anumanchipalli, G., Keutzer, K., and Gholami, A. Plan-and-act: Improving planning of agents for long-horizon tasks. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ybA4EcMmUZ.

Feng, P., Lv, Z., Ye, J., Wang, X., Huo, X., Yu, J., Xu, W., Zhang, W., Bai, L., He, C., and Li, W. Earth-agent: Unlocking the full landscape of earth observation with agents, 2026. URL https://arxiv.org/abs/2509.23141.

Froger, R., Andrews, P., Bettini, M., Budhiraja, A., Cabral, R. S., Do, V., Garreau, E., Gaya, J.-B., Laurençon, H., Lecanu, M., Malkan, K., Mekala, D., Ménard, P., Bertran, G. M.-T., Piterbarg, U., Plekhanov, M., Rita, M., Rusakov, A., Vorotilov, V., Wang, M., Yu, I., Benhalloum, A., Mialon, G., and Scialom, T. Are: Scaling up agent environments and evaluations, 2025. URL https://arxiv.org/abs/2509.17158.

Gou, Z., Shao, Z., Gong, Y., yelong shen, Yang, Y., Duan, N., and Chen, W. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Sx038qxjek.

Humburg, D. Chapter 38: Determining harvest losses in soybeans. In *South Dakota State University iGrow Soybean Best Management Practices*. South Dakota State University Extension, 2019.

Lee, C., Paramanayakam, V., Karatzas, A., Jian, Y., Fore, M., Liao, H., Yu, F., Li, R., Anagnostopoulos, I., and Stamoulis, D. Multi-agent geospatial copilots for remote sensing workflows. In *IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1084–1089, 2025. doi: 10.1109/IGARSS55030.2025.11243915.

Michelakis, P., Hadjiyianni, Y., and Stamoulis, D. Core: Full-path evaluation of llm agents beyond final state, 2025. URL https://arxiv.org/abs/2509.20998.

Monsi, M. and Saeki, T. On the factor light in plant communities and its importance for matter production. *Annals of botany*, 95(3):549–567, 2005.

Monteith, J. L. Climate and the efficiency of crop production in britain. *Philosophical transactions of the royal society of London. B, Biological Sciences*, 281(980):277–294, 1977.

OpenAI, :, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Seo, S. and Lee, K. Density-driven multidrone coordination for efficient farm coverage and management in smart agriculture. *IEEE Transactions on Control Systems Technology*, 34(2):711–724, March 2026. ISSN 2374-0159. doi: 10.1109/tcst.2025.3631091. URL http://dx.doi.org/10.1109/TCST.2025.3631091.

Shabbir, A., Munir, M. A., Dudhane, A., Sheikh, M. U., Khan, M. H., Fraccaro, P., Moreno, J. B., Khan, F. S., and Khan, S. Thinkgeo: Evaluating tool-augmented agents for remote sensing tasks, 2026. URL https://arxiv.org/abs/2505.23752.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. R., and Yao, S. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vAElhFcKW6.

Siler, T. B. and Singh, M. P. Optimal soybean maturity group selection is influenced by planting date in northern production systems. *Crop Science*, 62(6):2462–2475, 2022.

Stamoulis, D. and Marculescu, D. Geo-olm: Enabling sustainable earth observation studies with cost-efficient open language models & state-driven workflows. In *Proceedings of the 2025 ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, COMPASS '25, pp. 608–619, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714849. doi: 10.1145/3715335.3735494. URL https://doi.org/10.1145/3715335.3735494.

Steduto, P., Hsiao, T. C., Raes, D., and Fereres, E. Aquacrop—the fao crop model to simulate yield response to water: I. concepts and underlying principles. *Agronomy journal*, 101(3):426–437, 2009.

Tan, R., Peng, B., Yang, Z., Cheng, H., Mees, O., Zhao, T., Tupini, A., Meijier, I., Wu, Q., Yang, Y., Liden, L., Gu, Y., Zhang, S., Liu, X., Wang, L., Pollefeys, M., Lee, Y. J., and Gao, J. Multimodal reinforcement learning with adaptive verifier for ai agents, 2026. URL https://arxiv.org/abs/2512.03438.

Wang, H., Guan, Y., Meng, F., Zhao, C., Yan, L., Yang, Y., and Jiang, J. Agri-CM^3: A Chinese massive multimodal, multi-level benchmark for agricultural understanding and reasoning. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11729–11754, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.576. URL https://aclanthology.org/2025.acl-long.576/.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=BAakY1hNKS.

Xu, B., Peng, Z., Lei, B., Mukherjee, S., Liu, Y., and Xu, D. Rewoo: Decoupling reasoning from observations for efficient augmented language models, 2023. URL https://arxiv.org/abs/2305.18323.

Xu, Z., Xu, J., Zhang, M., Wang, P., Deng, C., and Liu, C.-L. Multimodal agricultural agent architecture (ma3): A new paradigm for intelligent agricultural decision-making, 2025. URL https://arxiv.org/abs/2504.04789.

Yan, L., Wang, H., Tang, C., Liu, H., Sun, T., Liu, L., Guan, Y., and Jiang, J. Agrieval: A comprehensive chinese agricultural benchmark for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34205–34213, 2026.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.03629.

Zhang, Z., Zhang, J., Liu, H., Lv, Q., Yang, J., Cai, K., and Wang, K. Agriworld:a world tools protocol framework for verifiable agricultural reasoning with code-executing llm agents, 2026. URL https://arxiv.org/abs/2602.15325.

Zhou, A., Yan, K., Shlapentokh-Rothman, M., Wang, H., and Wang, Y.-X. Language agent tree search unifies reasoning, acting, and planning in language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Zuzuárregui, M. A., Toslak, M. M., and Carpin, S. One for all: Llm-based heterogeneous mission planning in precision agriculture, 2025. URL https://arxiv.org/abs/2506.10106.

A. Example Scenarios by Task Level

We show representative samples from the 25-scenario suite, covering real-world soybean farm operations used in our preliminary field study. Each task is phrased as a natural-language user request, injected as an ARE `user event`, which the agent system must resolve through executable farm operations.

L1: Atomic tasks. *Aphid verification and treatment.* Yesterday, the drone detected signs of aphids on ridges 15–25. Please verify the issue and apply an appropriate pesticide treatment. Report back after the operation is complete.

Irrigation after dry weather. It has not rained recently. Check which areas are water-stressed, irrigate them as needed, and confirm the irrigation outcome. Tell me when the operation is complete.

L2: Episode chains. *Low-NDVI diagnosis at V4.* Remote-sensing analysis shows a localized low-NDVI area during the V4 vegetative stage. Diagnose the cause without directly assuming drought or pest damage. Distinguish among water stress, pest or disease pressure, and nutrient deficiency. If the issue is confirmed to be nutritional, treat it through the ridge-level fertigation system.

Threshold-based pest response. The sensors show signs of mild pest pressure. Determine whether the pressure reaches the treatment threshold by monitoring it *daily* over one week. If pest pressure continues to increase and is confirmed by ground inspection, spray according to the affected area.

Harvest-window selection at R8. The soybean crop has reached maturity R8 stage, but grain moisture is high today. Select an appropriate harvest window using the available weather, crop-state, and grain-moisture information.

L3: Full-season scenarios. *Full-season management from planting to storage.* Pre-planting operations have already been completed. Complete planting, monitor crop status weekly, address nutrient, water, pest, and disease issues using sensor and drone observations, select an appropriate harvest window after maturity, and complete drying and storage; before each operation, reassess the current weather, soil, and crop state and revise actions as needed.

Full-season drought management during grain filling. Pre-planting operations have already been completed. Early-season conditions are normal, but assess where and when sustained drought risk may emerge before the August grain-filling period. Manage the full season from planting onward, with daily monitoring of soil sensors and thermal imagery during R5/R6 to confirm water stress. Use rainfall, soil-moisture, and crop-state evidence to schedule irrigation in time to protect yield potential.

Full-season high-density management with low-NDVI diagnosis. Pre-planting operations have already been completed for high-density planting, so take over farm management from planting onward. Monitor crop growth using soil moisture, thermal imagery, and SPAD measurements to detect conditions that may lead to low-NDVI areas. Confirm suspected issues with drone inspection before selecting an intervention. Use the ridge-level fertigation system when the evidence indicates nutrient stress.